Causality

B9145: Reliable Statistical Learning

“What if” questions

You’ll often want to know the causal effect of some action.

You’ll want to answer “**what if**” questions:

- Will patients get better **if** they take this drug?
- Will fewer people smoke **if** we add a cigarette tax?
- Will more people buy this product **if** they see this new advertisement?
- Will I have a better outcome **if** I went to surgeon A vs surgeon B?

Drug example

- There’s a disease. Some people with the disease get better on their own.

- You develop a drug. You recruit people to try out the drug.
    - Anyone who wants the drug gets the drug.

- You find that a larger fraction of the recruited people, who take the drug, get better.

- However, years later, you find out the drug does not work. What could have gone wrong?

Drug example

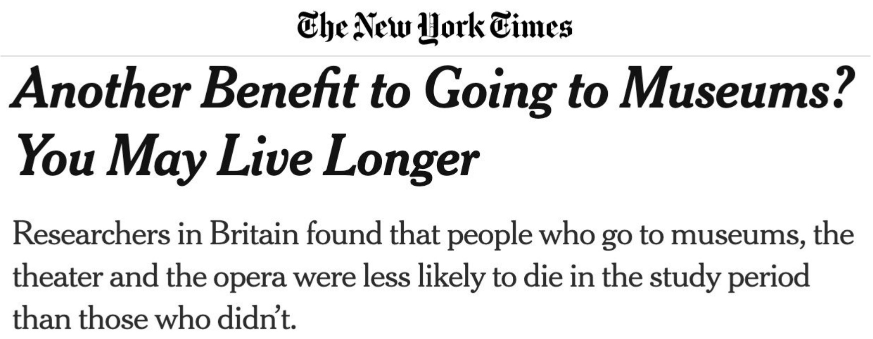
Perhaps the disease is debilitating, so that only people with milder symptoms were able to try out the drug, and people with milder symptoms are more likely to get better.

Or perhaps people who are more affluent are more likely to get better on their own, and also more likely to try the drug.

Problem: we did not observe counterfactuals

- For each person, what would have happened if they took the drug vs didn’t?

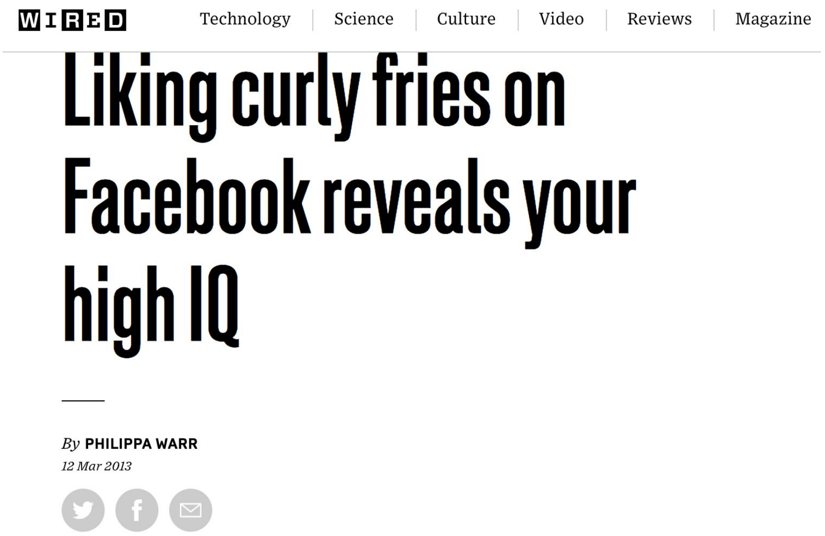
Secret to life



The New York Times

Another Benefit to Going to Museums? You May Live Longer

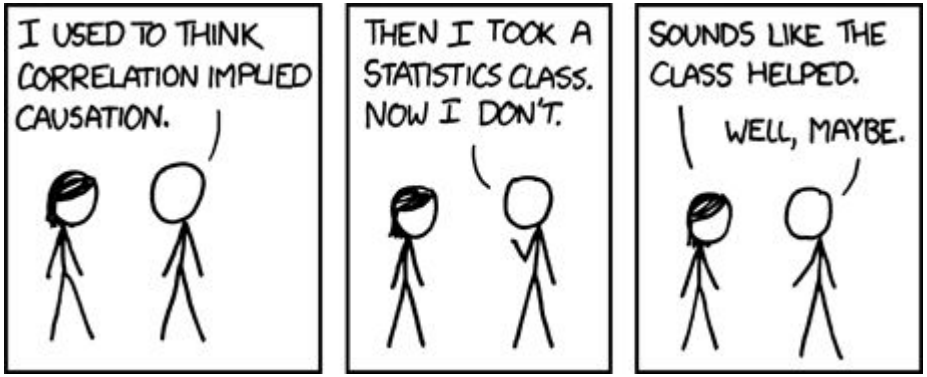
Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.



Liking curly fries on Facebook reveals your high IQ

By PHILIPPA WARR

12 Mar 2013


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://xkcd.com/552/

Prediction and causation

Lots of ML is about prediction.

- How and when can we use prediction to estimate causal effects?
- What structure of the data do we need?

Causal inference is a multidisciplinary field built across economics, epidemiology, and statistics.

Binary actions

- Today we will focus on the setting with two actions
  - One action represents treatment (1), the other is control (0)
- This is still foundational
  - Key difficulties still persist here despite the simplicity
  - Core technical insights will translate to more general settings
- In complex problems, this is often the standard
  - Control is status quo, treatment is the new program
  - Throughout economics, medicine, and tech, it requires a tremendous amount of domain knowledge and effort to come up with an alternative to the current system

Potential outcomes

Also known as **bandit feedback**

Framework for explicitly modeling counterfactuals

- A: binary treatment assignment (1: treated, 0: control)
- Y(1) and Y(0) are potential outcomes under treatment and control, respectively.
  - Assume we observe Y=Y(A).
- X is observed covariates

First goal: Estimate average treatment effect

$$\tau := \mathbb{E}[Y(1) - Y(0)]$$

Problem: We only observe Y=Y(A)

It's a missing data problem!

Potential outcomes

Also known as **bandit feedback**

Framework for explicitly modeling counterfactuals

- A: binary treatment assignment (1: treated, 0: control)
- Y(1) and Y(0) are potential outcomes under treatment and control, respectively.
  - Assume we observe **Y=Y(A)**.
- X is observed covariates

SUTVA assumption

First goal: Estimate average treatment effect

$$\tau := \mathbb{E}[Y(1) - Y(0)]$$

Problem: We only observe Y=Y(A)

SUTVA assumption: **Y=Y(A)** deceptively simple!

SUTVA = Stable Unit Treatment Value Assumption

1. No interference between units

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

2. No hidden variation of treatment

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

When might these assumptions not hold?

- Network effects: effect of vaccines on disease prevalence
- Two-sided platforms: ridesharing, ad auctions
- Equilibria: if everyone gets a job training, it won't increase everyone's income
- Different ways of administering a drug, expired vs un-expired medication

Average treatment effect (ATE)

First goal: Estimate average treatment effect

$$\tau := \mathbb{E}[Y(1) - Y(0)]$$

- We only observe Y := Y(A)
- What could go wrong?

Person	A	Y(0)	Y(1)	Y(1) - Y(0)
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	0	1	1	0
6	0	1	1	0
7	0	1	1	0
8	0	1	1	0

Randomized controlled trials (RCT)

also called A/B testing, (randomized) experiments

- First try: let's randomize treatment assignments

$$Y(1), Y(0) \perp A$$

- By randomized assignments and then SUTVA, we have

$$\boxed{\tau} = \mathbb{E}[Y(1) - Y(0)] \overset{\text{independence}}{\boxed{=}} \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0]$$

$$y = y(A) \longrightarrow \; = \underline{\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]}$$ ← **observable**

- We can estimate final line from i.i.d. data (Y_i, A_i)

  → **difference in means estimator**

Difference in means estimator

$$\tau = E[y \mid A=1] - E[y \mid A=0]$$

estimand

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{i:A_i=1} y_i - \frac{1}{n_0} \sum_{i:A_i=0} y_i$$

estimator

$$n_a = \#\{A_i = a\}$$

$$\mathrm{var}(\hat{\tau}_{DM} \mid n_0, n_1) = \frac{1}{n_0} \mathrm{var}[y(0)] + \frac{1}{n_1} \mathrm{var}[y(1)]$$

CLT $\rightsquigarrow$ $\sqrt{n}(\hat{\tau}_{DM} - \tau) \to N(0, V_{DM})$

$$V_{DM} = \frac{\mathrm{var}(y(1))}{p(A=1)} + \frac{\mathrm{var}(y(0))}{p(A=0)}$$

Linear regression adjustments

Randomization might not be perfect in small samples. Can we do better?

1. Assume linear model
2. Construct alternative estimator $\hat{\tau}_{OLS}$
3. compare with $\hat{\tau}_{DM}$!

Assume linear model: (X_i, A_i, y_i) as ground truth:

$$y(a) = c_a + X\beta_a + \varepsilon(a)$$

$$E[\varepsilon(a) \mid X] = 0, \quad \operatorname{var}[\varepsilon(a) \mid X] = \sigma^2$$

$$P(A=0) = 1/2, \quad E[X] = 0, \quad \operatorname{var} X = \Sigma$$

Asymp. variance of $\hat{\tau}_{DM}$:

$$V_{DM} = 4\sigma^2 + \|\beta_0 + \beta_1\|_\Sigma^2 + \|\beta_0 - \beta_1\|_\Sigma^2$$

where $\|v\|_\Sigma^2 := v^T \Sigma v$

Linear regression adjustments

under linear assumptions,

$$\tau = E[y(1) - y(0)] = c_1 - c_0 + EX(\beta_1 - \beta_0)$$

Then we can use OLS to learn $\hat{\beta}_a, \hat{c}_a$:

$$\hat{\beta}_a, \hat{c}_a = \underset{\beta, c}{\operatorname{argmin}} \frac{1}{n} \sum_{i : A_i = a} \left(y_i - X_i \hat{\beta}_a - \hat{c}_a \right)^2$$

and estimate

$$\hat{\tau}_{OLS} = \hat{c}_1 - \hat{c}_0 + \bar{X}(\hat{\beta}_1 - \hat{\beta}_0), \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i .$$

Then, calculations show

$$\sqrt{n}\,(\hat{\tau}_{OLS} - \tau) \to N(0, V_{OLS})$$

$$V_{OLS} = 4\sigma^2 + \|\beta_0 - \beta_1\|_\Sigma^2 = V_{DM} - \|\beta_0 + \beta_1\|_\Sigma^2$$

Linear regression adjustments

Perhaps surprisingly, let $y(a) = \mu_a(x) + \varepsilon(a)$

instead of $y(a) = c_a + X\beta_a$.

Then with $\hat{\tau}_{DM}$, $\hat{\tau}_{OLS}$ defined as before,

we also have

$$V_{OLS} = V_{DM} - \|\beta_0^* + \beta_1^*\|_\Sigma^2$$

where $(c_a^*, \beta_a^*) = \arg\min_{c,b} \frac{1}{n} \sum_{i:A_i=a} (y_i - X_i \beta_a^* - c_a^*)^2$.

You don't need to estimate w/ OLS. See Direct Method slides later

see Stefan Wager's Stats 361 notes

Beyond RCTs: observational studies

- Randomization is sometimes infeasible or prohibitively expensive
  - post-market drug surveillance
  - effect of air pollution on long-term health outcomes
  - effect of a government policy on some economic outcome

- May want to use existing data collected under a previous data generating policy

Simpson's paradox

# successful procedures / # total procedures

	actual surgery	band-aid	total
	0 / 1	99 / 100	99 / 101
	40 / 50	1 / 1	41 / 51

Who is the better doctor?

or, what is the effect of doctor choice on procedure success?

Y: procedure success, A: doctor, X: type of procedure

Berkeley admissions

- Berkeley was sued for gender bias in admissions based on 1973 numbers: 44% of men were admitted but only 35% of women

- But individual department’s admissions record showed no evidence of such gender-based discrimination

- Turns out women systematically applied to more competitive majors

Y, A, X?

No unobserved confounding

- Previous regression-based direct method still works if there are no unobserved confounders (also called **ignorability**)

Assumption: $Y(1), Y(0) \perp A \mid X$

- Observed treatment assignments are based on covariate information alone (+ random noise)

- Treatment assignment does not use information about counterfactuals

- Strong assumption, often violated in practice.
    - e.g. doctors often use unrecorded info to prescribe treatments

Direct method

- By no unobserved confounding (and then SUTVA),

$$\begin{aligned} \underline{\mu_a^\star(X)} &:= \mathbb{E}[Y(a) \mid X] \\ &= \mathbb{E}[Y(a) \mid X, A = a] \\ &= \mathbb{E}[Y \mid X, A = a] \longleftarrow \textbf{observable} \end{aligned}$$

- Fit **outcome models** by loss minimization

$$\text{minimize}_{\mu_a \in \mathfrak{M}_a} \quad \mathbb{E}[(Y(a) - \underline{\hat{\mu}_a}(X))^2 \mid A = a]$$

- ATE estimator

$$\hat{\tau}_{\mathrm{DM}} := \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Good if outcome models are easy to learn
- Similar to RCT adjustment

Overlap

- We need enough samples for both control and treatment throughout the covariate space (i.e. for each X)
  - You want to compare treatment and control over the same X's
  - If treatment contains X's that are rare or nonexistent in the control, it's hard to compare
  - Overlap governs effective sample size

- **Propensity score** $e^\star(X) := \mathbb{P}(A = 1 \mid X)$

- Assume that there exists $\epsilon > 0$ such that $\epsilon \leq e^\star(X) \leq 1 - \epsilon$ almost surely
  - This means I have at least ϵn number of samples for fitting the two outcome models

Overlap

$$P(A=1 \mid X>50) = 1 \notin [\varepsilon, 1-\varepsilon]$$

- Overlap breaks if data is generated by a deterministic policy
  - e.g. always assign the drug (treatment) when age > 50

- We need sufficient amount of randomness in treatment assignment in all covariate regions

- Often violated in practice

Inverse propensity weighting

- What if the outcome models are very complex and difficult to estimate?
    - Direct method less good
- A natural approach is to reweight samples to correct for confounding bias
    - Essentially importance sampling
- First, estimate the propensity score
    - e.g. run logistic regression to predict A given X $e^{\star}(X) := \mathbb{P}(A = 1 \mid X)$

$$e^\star(X) := \mathbb{P}(A = 1 \mid X)$$

Inverse propensity weighting

$$\varepsilon < e^\star(x) < 1 - \varepsilon$$

- Estimator

$$\hat{\tau}_{\mathrm{IPW}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{e}(X_i)}Y_i - \frac{1-A_i}{1-\hat{e}(X_i)}Y_i\right)$$

- Can work well if propensity score is simple to estimate

- But estimating this well over the entire covariate space can be difficult
  - Calibration is hard, especially in high-dimensions

- When overlap doesn’t hold, importance weights blow up

$$e^{\star}(X) := \mathbb{P}(A = 1 \mid X)$$

Inverse propensity weighting

$$\hat{\tau}_{\mathrm{IPW}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{e}(X_i)}Y_i - \frac{1-A_i}{1-\hat{e}(X_i)}Y_i\right)$$

want to check $E[\hat{\tau}_{IPW}] = \tau$.

only check this term

(same works for)

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{A_i y_i}{e(X_i)}\right]$$

$$= E\left[\frac{A y}{e(X)}\right] = E\left[\frac{A y(1)}{e(X)}\right]$$ SUTVA

$$= E\left[E\left[\frac{A y(1)}{e(X)} \,\middle|\, X\right]\right]$$ tower

$$= E\left[\frac{1}{e(X)}\underbrace{E[A \mid X]}_{e(X)}E[y(1) \mid X]\right]$$ no unobserved confounding

$$= E\left[E[y(1) \mid X]\right] = E[y(1)]$$

Recap of assumptions

- **SUTVA**: Y=Y(A)
- **Ignorability** / no unobserved confounding: $Y(1), Y(0) \perp A \mid X$
- **Overlap**: $\epsilon > 0$ such that $\epsilon \leq e^{\star}(X) \leq 1 - \epsilon$

Conditional Average Treatment Effect (CATE)

Second goal: Estimate **conditional average treatment effect**

$$\tau(X) := \mathbb{E}[Y(1) - Y(0) \mid X]$$

- Estimate **personalized** treatment effects
    - E.g. a drug is more effective in some age groups than in others

- As before, missing data: we may not observe both Y(0) and Y(1)

Estimating heterogeneous treatment effect

There are a few ways to do it, including

- T-Learner
- S-Learner
- R-Learner

T-learner

Separate models for treatment and control

S-Learner

Shared feature representation, similar model class for both treatment and control

R-learner