B9145: Reliable Statistical Learning
Hongseok Namkoong

Causality

https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB

https://scholar.harvard.edu/imbens/files/efficient_estimation_of_average_treatment_effects_using_the_estimated_propensity_score.pdf

https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097

https://www.pnas.org/content/116/10/4156

https://arxiv.org/pdf/1712.04912.pdf

Prediction and causality

- A central goal of ML is to predict an outcome given variables describing a situation
  - Given patient characteristics, will their outcome improve?

- Most decision-making problems revolve around a decision / intervention / treatment
  - What would happen if we changed the system?
  - Given patient characteristics, will their outcome improve if **they follow a new diet**?

- We want to develop a scientific understanding of a decision

Prediction and causality

- Causal inference is a multi-disciplinary field built across economics, epidemiology, and statistics

- Focus is on questions about **counterfactuals**
  - What structure of data do we need to answer this question?
  - How do we interpret the key estimands?

- ML models can predict outcomes; when can it predict counterfactuals?
  - How can we leverage flexible ML models to infer causality?

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 de facto standard
  - Control is status quo, treatment is a new elaborate 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

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.

WIRED Technology | Science | Culture | Video | Reviews | Magazine

Liking curly fries on Facebook reveals your high IQ

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

By PHILIPPA WARR

Causality

- You came up with a new diet regimen that you believe will alleviate symptoms of rheumatism (e.g. chronic joint pain)
- To test it, you recruit people to try the diet
- You find that
  - Small fraction on the diet experience chronic pain
  - Large fraction not on the diet (aka all rheumatism patients outside your volunteer pool) experience chronic pain
  - Awesome! Everyone should try this diet
- But after years of adoption, you realize the diet does not affect chronic pain

Figure credit: data science central

Causality

- What could have gone wrong?
  - Volunteers to the diet may have been people with healthy predispositions, and affluent socioeconomic backgrounds

- **Fundamental problem:** we don't observe counterfactuals

- How do we model this?

Figure credit: data science central

Potential outcomes

- Framework for explicitly modeling counterfactuals
- A: binary treatment assignment (1: treated, 0: control)
- Y(1) and Y(0) are potential outcomes
- X is observed covariates

First goal: Estimate average treatment effect

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

Problem: We only observe $Y := Y(A)$

ATE

First goal: Estimate average treatment effect

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

- We only observe Y := Y(A)
- What could go wrong?
  - Volunteers to the diet (A = 1) may have been people with healthy predispositions, and affluent socioeconomic backgrounds

Person	5	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 control trials

also called A/B testing, (randomized) experiments

- First try: let's **randomize** treatment assignments

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

- By virtue of randomized assignments, we have

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

$$= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$$ ← **observable**

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

Randomized control trials

Randomized control trials

RCT with covariates

- If you have access to covariates X, and can estimate $\mathbb{E}[Y \mid X, A]$ accurately, then we can improve this

- If by randomness more treatments get assigned to young patients with a better prognosis, then we will exaggerate the treatment effect
  - Problem goes away in large samples, but matters for small samples

- Using any regression model, we can estimate $\mathbb{E}[Y \mid X, A = 1], \mathbb{E}[Y \mid X, A = 0]$ ← **observable**
  - Random forests, boosted decision trees, kernels, NNs etc

Estimator

Fitting outcome models

CLT for covariate adjustments

Beyond RCTs

- What if clean randomization is not possible?
- Randomization sometimes affected by the site
  - Oxford / AstraZeneca trial made a dosage mistake at a location
  - Turned out to be more effective
- Ignoring variables that affect treatment assignment leads to biases

Beyond RCTs

- Run large-scale experiment, randomized for each sex

	Men		Women	
	No disease ($Y=1$)	Disease ($Y=0$)	No disease ($Y=1$)	Disease ($Y=0$)
Treatment ($A=1$)	0.1500	0.2250	0.1000	0.0250
Control ($A=0$)	0.0375	0.0875	0.2625	0.1125

(Here the numbers are the fractions of individuals in each category.)

- $\mathbb{P}(Y=1 \mid A=1)=0.5$ vs $\mathbb{P}(Y=1 \mid A=0)=0.6$
  - So maybe treatment is not effective?

Slide by Ramesh Johari

Simpson's paradox

- But if you compute treatment effect for each sexes,

$$\mathbb{E}[Y(1) - Y(0) \mid X = m] = \mathbb{E}[Y(1) - Y(0) \mid X = w] = 0.1$$

- So ATE = 0.1. What happened?

- Women are more likely to be in control than treatment; men are more likely to be in treatment than control. And women have higher potential outcomes on average than men.

Simpson's paradox

- Issue here is that

$$\mathbb{E}[Y(1) - Y(0)] \neq \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0]$$

- If you ignore sex as a confounding variable, you create a **omitted variable bias** in estimating the ATE

Slide by Ramesh Johari

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

Observational studies

- Randomization is sometimes infeasible or prohibitively expensive
  - e.g. post-market drug surveillance, effect of air pollution on long-term health outcomes

- Experimentation can be risky in high-stakes scenarios
  - operational scenarios: new inventory system for Amazon, new pricing algorithm for Uber

- May want to use existing large-scale data collected under some data-generating policy (e.g. legacy system)

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

No unobserved confounding

Overlap

- We need enough samples for both control and treatment throughout the covariate space
  - This governs the 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

- This 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
- Governs difficulty of estimation. Often violated in practice.

Direct method

Direct method

Inverse probability weighting

- What if the outcome models are very complex and difficult to estimate?

- A natural approach is to reweight samples, to change the distribution $\mathbb{E}[\,\cdot \mid A=1,X]$ to $\mathbb{E}[\,\cdot \mid X]$
  - Essentially importance sampling

Unbiasedness

CLT for IPW

Estimating propensity score

Inverse probability weighting

- 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

Augmented IPW

- Can we combine the best of both worlds?
  - Direct method + IPW

- Propensity weight residuals to debias the direct method

Unbiasedness

CLT for AIPW

Control variate

Control variate

Efficiency

- In fact, this is the best asymptotic variance we can get
- AIPW has optimal asymptotic variance, regardless of whether the propensity score is known or not
- Formalizing this requires a lot of work

Nuisance parameters

- If a good parametric model exists, then can estimate at the usual $1/\sqrt{n}$ rates

- In general, these are infinite dimensional objects. Can be difficult to estimate.

Semiparametrics

- We only care about estimating the ATE
  - One-dimensional estimand, infinite dimensional nuisance parameters

- Estimation accuracy of nuisance parameters is good only insofar as it helps with estimating the ATE

- Due to its high-dimensional nature, often difficult to estimate nuisances at parametric rates

- Goal: semiparametric estimators that are insensitive to errors in nuisance estimates

Doubly robust

- One main advantage of AIPW is that even if one of the nuisance parameter models are **misspecified,** you can still get correct asymptotic behavior

Doubly robust

Doubly robust

Orthogonality

- When is a semiparametric estimator insensitive to errors in nuisance estimates?

- Directional derivative of functional wrt nuisance parameters at true value is near-zero

- Ensures that a little perturbation in nuisance parameters near the truth values does not affect functional

Orthogonality

Orthogonality of AIPW

Orthogonality of AIPW

Why orthogonality?

- Allows getting central limit rates on ATE estimation even when we can only estimate nuisance parameters at slower rates

- In addition to no unobserved confounding, $e^\star(X), \hat{e}(X) \in [\epsilon, 1-\epsilon]$, we assume the following rate condition

$$\|\hat{e} - e^\star\|_{P,2}(\|\hat{\mu}_1 - \mu_1^\star\|_{P,2} + \|\hat{\mu}_0 - \mu_0^\star\|_{P,2}) = o_p(n^{-1/2})$$

- This allows us to trade-off errors between nuisance parameters. Only their product needs to go down at this rate!

Central limit result

- CLT for the semiparametric AIPW, even when nuisance estimates converge at slower-than-parametric rates

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{\mathrm{AIPW}}(X_i, Y_i, A_i; \hat{\mu}_0, \hat{\mu}_1, \hat{e}) - \tau\right) \Rightarrow N(0, \sigma^2_{\mathrm{AIPW}})$$

where $\sigma^2_{\mathrm{AIPW}} := \mathrm{Var}\left(\psi_{\mathrm{AIPW}}\left(X, Y, A; \mu_0^\star, \mu_1^\star, e^\star\right)\right)$

- This is the oracle asymptotic variance; when the true nuisance parameters are known

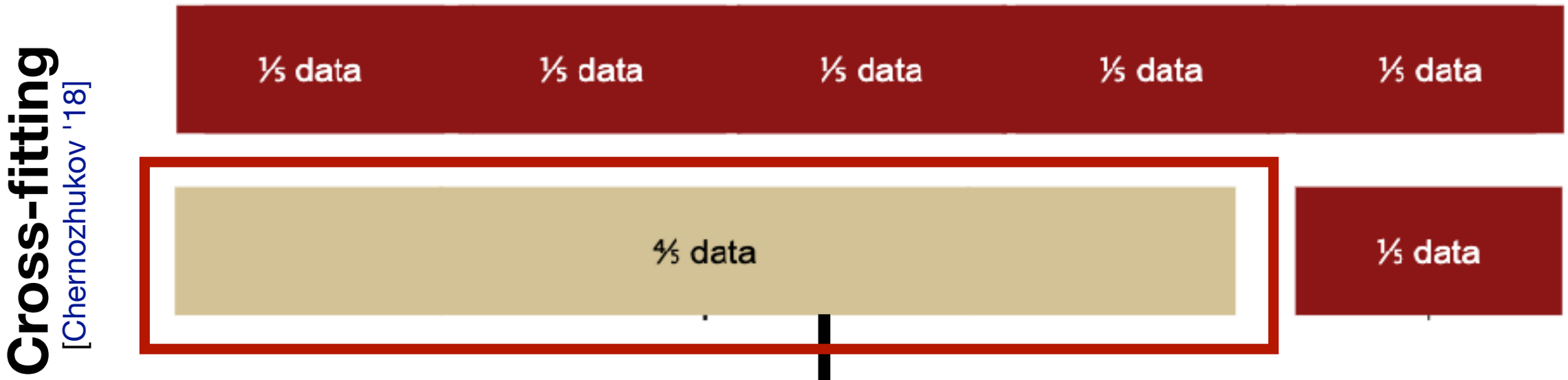
- AIPW achieves optimal asymptotic efficiency

Sketch of asymptotics

Sketch of asymptotics

Cross-fitting

- Instead of sample-splitting, we can alternate the role of main and auxiliary samples over multiple splits



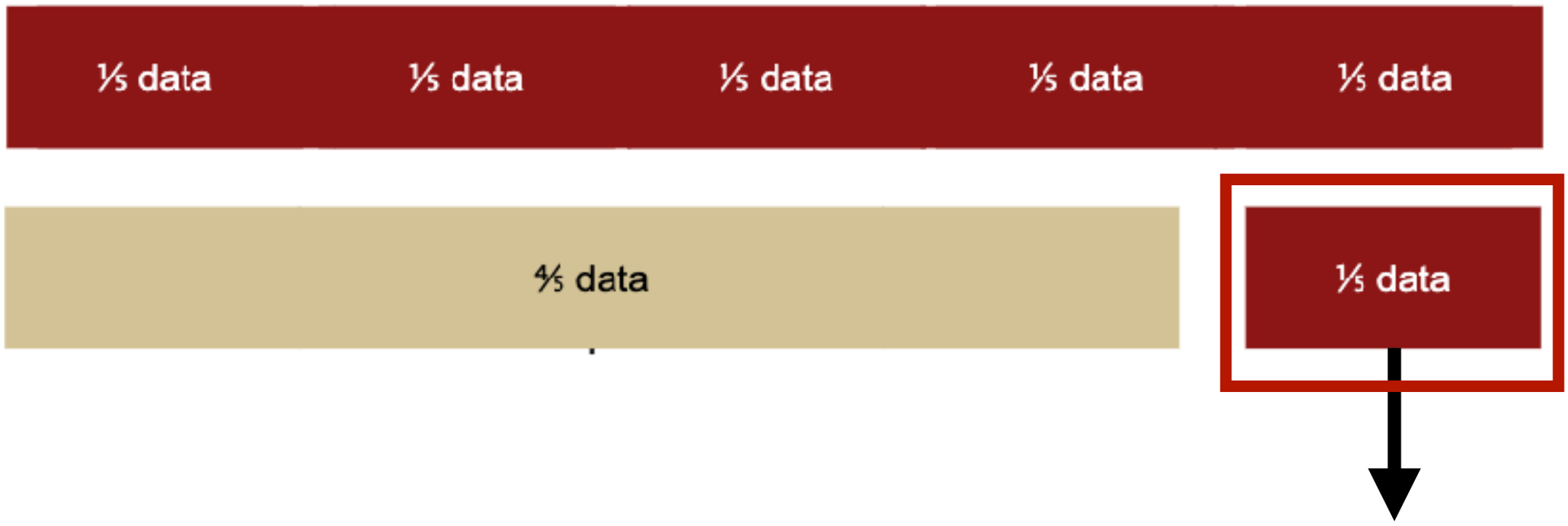
$$\widehat{\mu}_a(X) \approx \mathbb{E}[Y(a) \mid X = x],\ a \in \{0, 1\}$$

$$\widehat{e}(X) \approx \mathbb{P}(A = 1 \mid X)$$

- Estimate nuisance parameters on the auxiliary sample

Cross-fitting

Cross-fitting
[Chernozhukov '18]

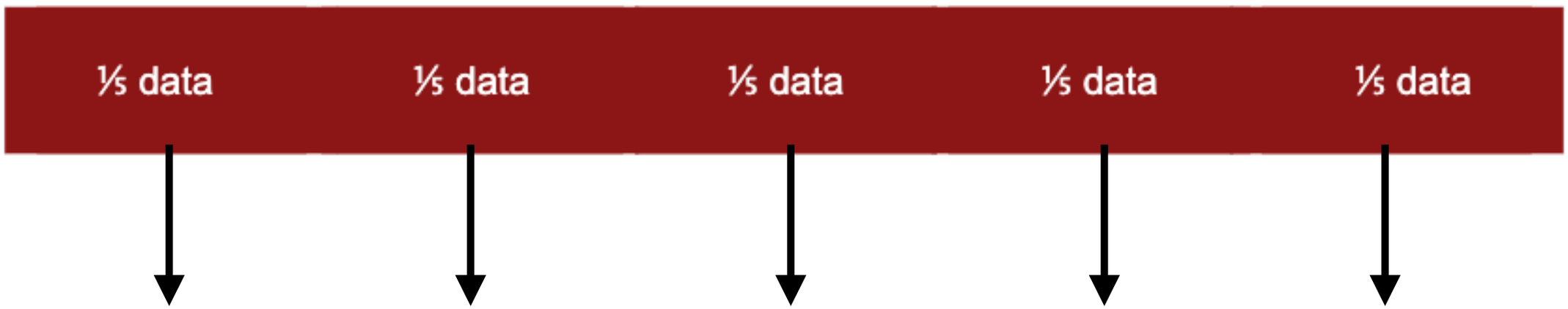


$$\hat{\tau}_1 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{A_i}{\hat{e}(X_i)}(Y - \mu_1(X_i)) - \frac{1 - A_i}{1 - \hat{e}(X_i)}(Y - \mu_0(X_i))$$

- Estimate ATE by plugging in nuisance estimates

Cross-fitting

Cross-fitting [Chernozhukov '18]



$$\hat{\tau} = \frac{1}{5}\Big(\quad \hat{\tau}_1 \; + \; \hat{\tau}_2 \; + \; \hat{\tau}_3 \; + \; \hat{\tau}_4 \; + \; \hat{\tau}_5 \quad \Big)$$

- Same procedure for direct method, IPW
- Similar central limit result follows as before

SUTVA

- Throughout we implicitly assumed there is only a single version of the treatment that gets applied to all treated units
  - This may not be true if drugs go stale in storage, or dosages differ

- We also assumed there is *no interference between units*
  - Whether or not individual i is treated has no impact on the treatment effect of another individual j
  - This can also fail in many real-world scenarios

- Together these assumptions are called stable unit treatment value assumption (SUTVA)

Interference

- Any two-sided platform faces interference between units
- Consider the following scenario:
  - Lyft A/B tests a new promotion strategy for drivers
  - Each driver is randomized into treatment or control
  - It is observed that drivers finish a lot more rides with the promotion
  - So they decide this promotion is worth spending resources on
- But the estimate turned out to be an **overestimate**, not worth the cost of the promotion. Why?

Interference

- Both treated and control drivers see the same set of demand

- If promotion incentivizes treated drivers to work more for less nominal fares, this cannibalizes demand that would usually go to control drivers

- Interference occurs in a number of different settings
  - Two-sided platforms: Airbnb, ridesharing, ad auctions
  - Network effects: e.g. adoption of new education technology

- When this happens, the potential outcomes now depend on all possible 2^n treatment assignments
  - Very active area of research

Assessing overlap

- “If the covariate distributions are similar, as they would be, in expectation, in the setting of a completely randomized experiment, there is less reason to be concerned about the sensitivity of estimates to the specific method chosen than if these distributions are substantially different.”

- “On the other hand, even if unconfoundedness holds, it may be that there are regions of the covariate space with relatively few treated units or relatively few control units, and, as a result, inferences for such regions rely largely on extrapolation and are therefore less credible than inferences for regions with substantial overlap in covariate distributions.”

- Imbens and Rubin

Assessing overlap

- Overlap governs effective sample size
  - Even approaches that don't require propensity weighting is affected under this fundamental restriction

- Causal inference literature has developed various "supplementary analysis" tools for assessing credibility of empirical claims

- One of the most common conventions is to plot the propensity scores of treated and control groups

Assessing overlap

- Difference in covariate distributions between treatment and control group is summarized by the propensity score

- Let $f_1(X)$ be the density of X in the treatment group (similarly $f_0(X)$)

- Let $p := \mathbb{P}(A = 1)$

$$\mathsf{Var}(e^\star(X)) = p(1-p)(\mathbb{E}\left[e^\star(X) \mid A = 1\right] - \mathbb{E}\left[e^\star(X) \mid A = 0\right])$$

$$= p^2(1-p)^2 \cdot \mathbb{E}\left[\left(\frac{f_1(X) - f_0(X)}{pf_1(X) + (1-p)f_0(X)}\right)^2\right]$$

Assessing overlap

- A common visualization is to look at the pdf of the propensity score across treatment groups

- Plot approximates pdfs of the distribution $\mathbb{P}(e^\star(X) \in \cdot \mid A = a)$

- For each $q \in (0,1)$, plot fraction of observations in the treatment group with $e^\star(x) = q$ (and similarly for control)

Assessing overlap

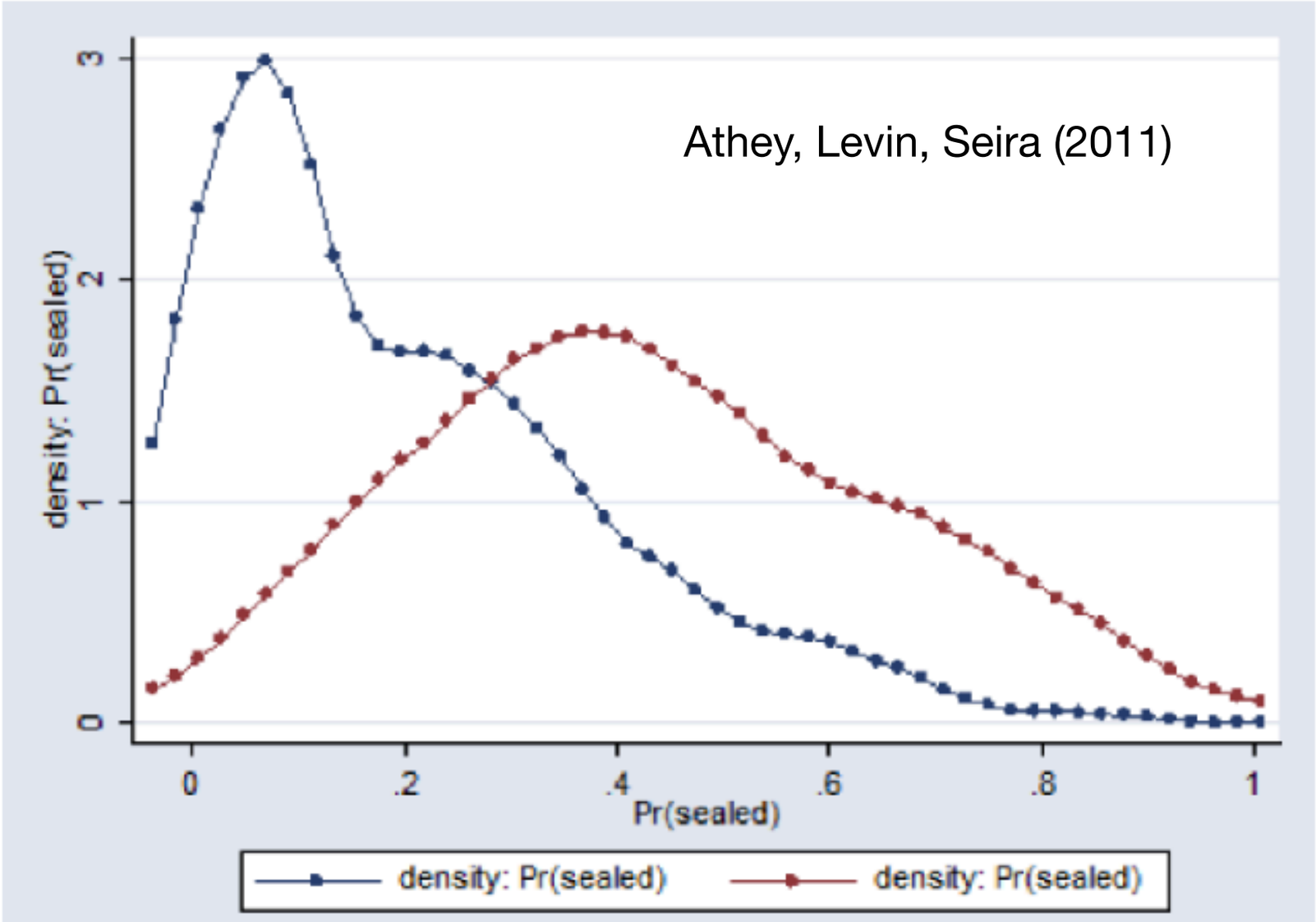
- Athey, Levin, Seira (2011) studied timber auctions
  - Award timber harvest contracts via first price sealed auction or open ascending auction

- Idaho: randomized with different probabilities across different regions

- California: determined by small vs. large sales volume; cutoff varies by region

Idaho

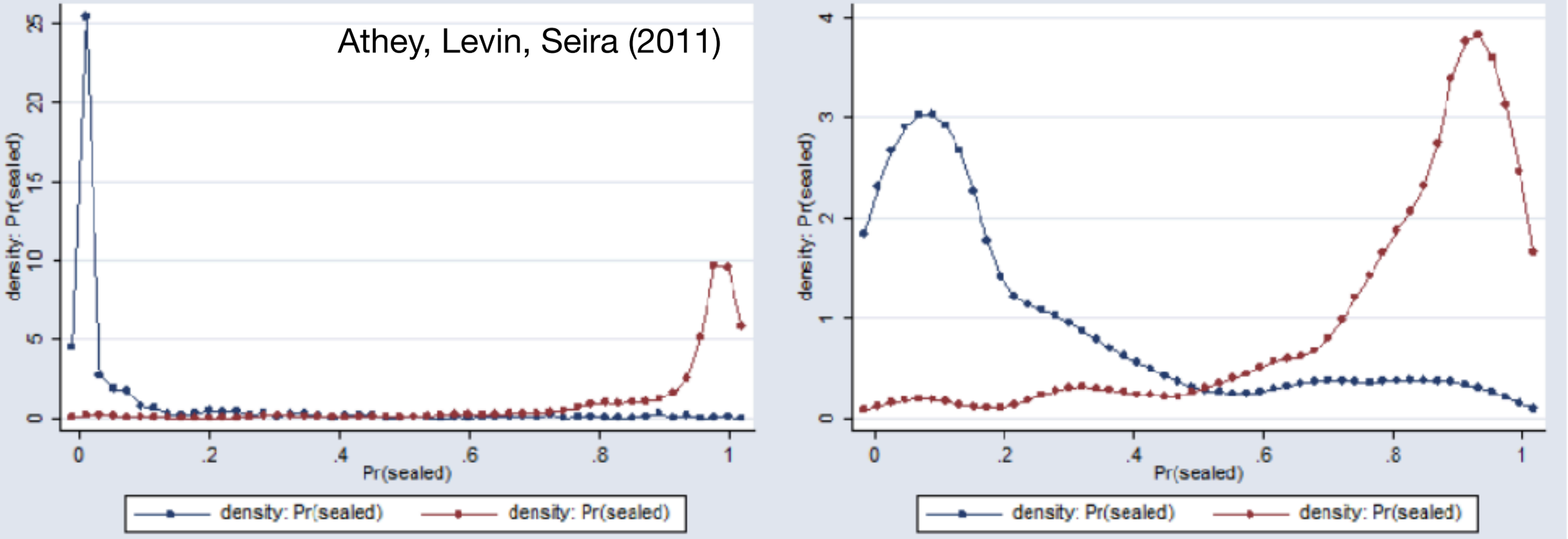
Very few observations with extreme propensity scores



Slide by Susan Athey and Stefan Wager

California

Untrimmed v. trimmed so that $e(x) \in [.025, .975]$



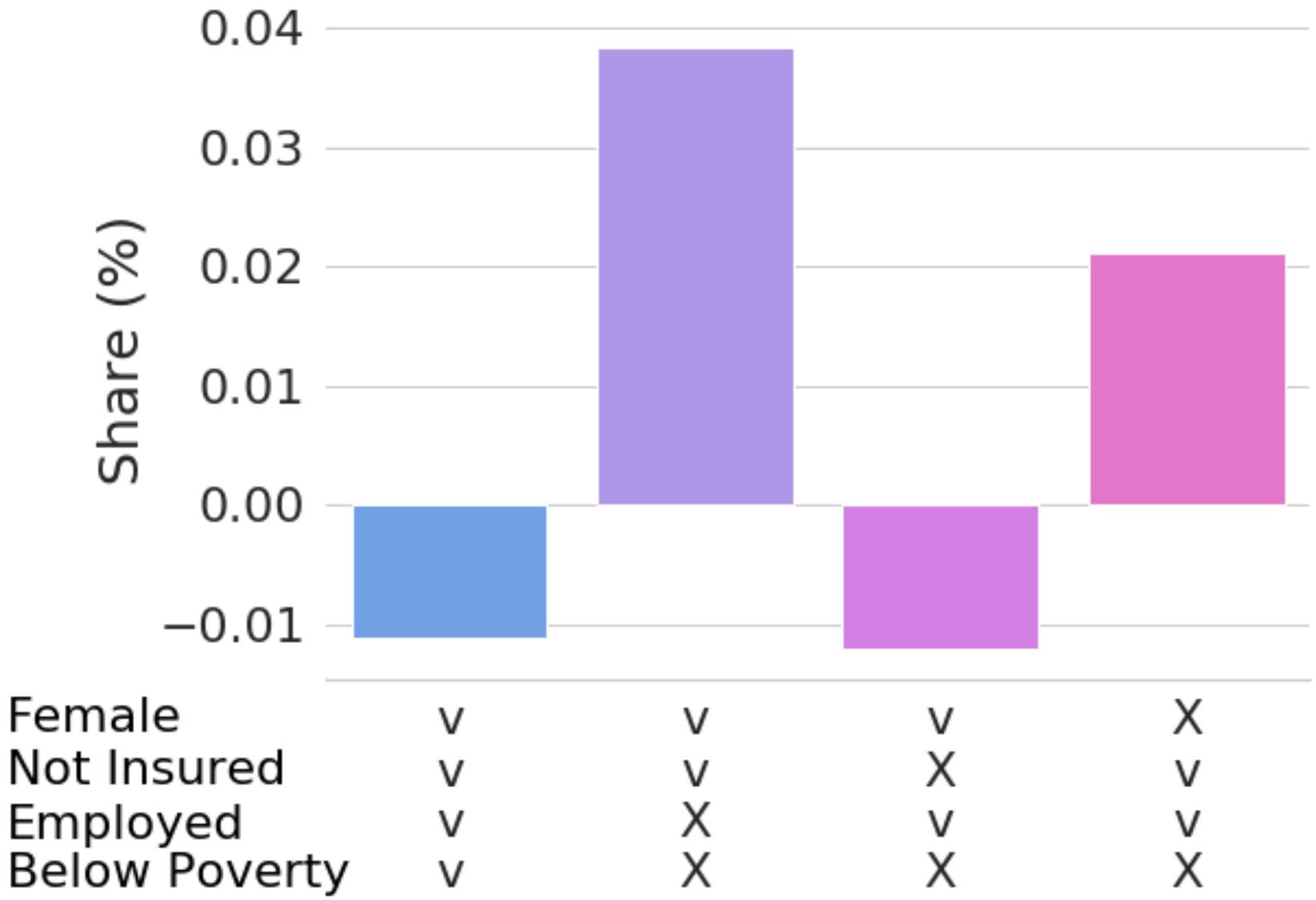
Slide by Susan Athey and Stefan Wager

Heterogeneous treatment effects

- Treatment effect often varies with user / patient / agent characteristics (covariates)

- Example: Oregon Health Insurance Experiment
  - Evaluate effect of Medicaid on low-income adults on emergency department (ED) visits in 2008
  - Precursory study to federal Medicaid expansion in 2014, which cost $553 billion/year
  - Insurance allows visits ED, but access to preventive care may also reduce need of ED visits

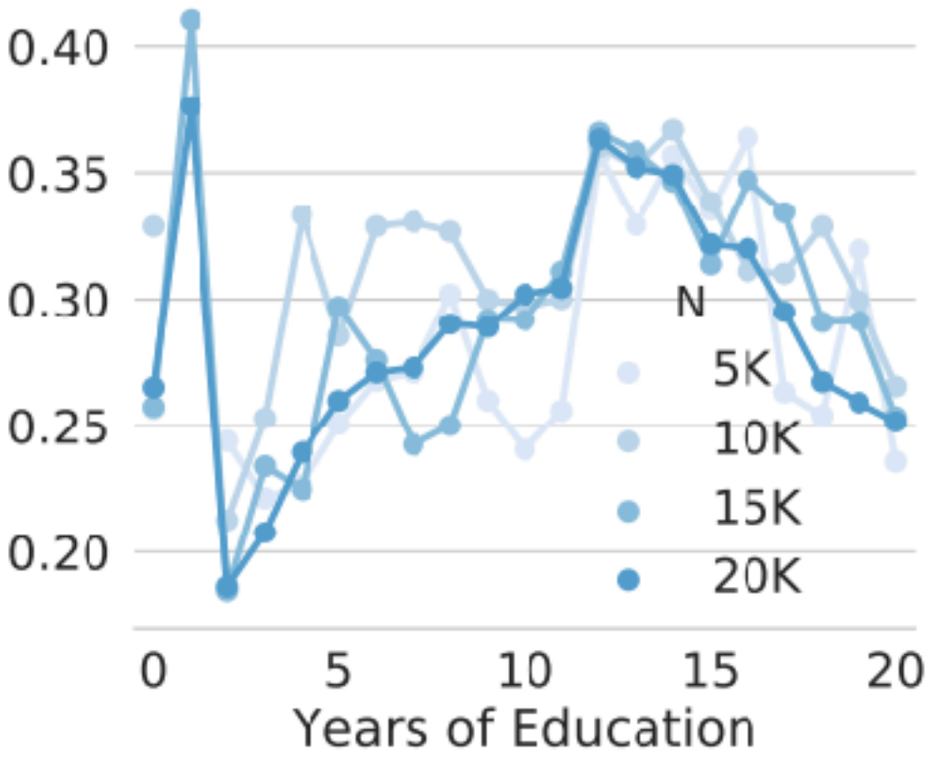
Oregon Health Insurance Experiment

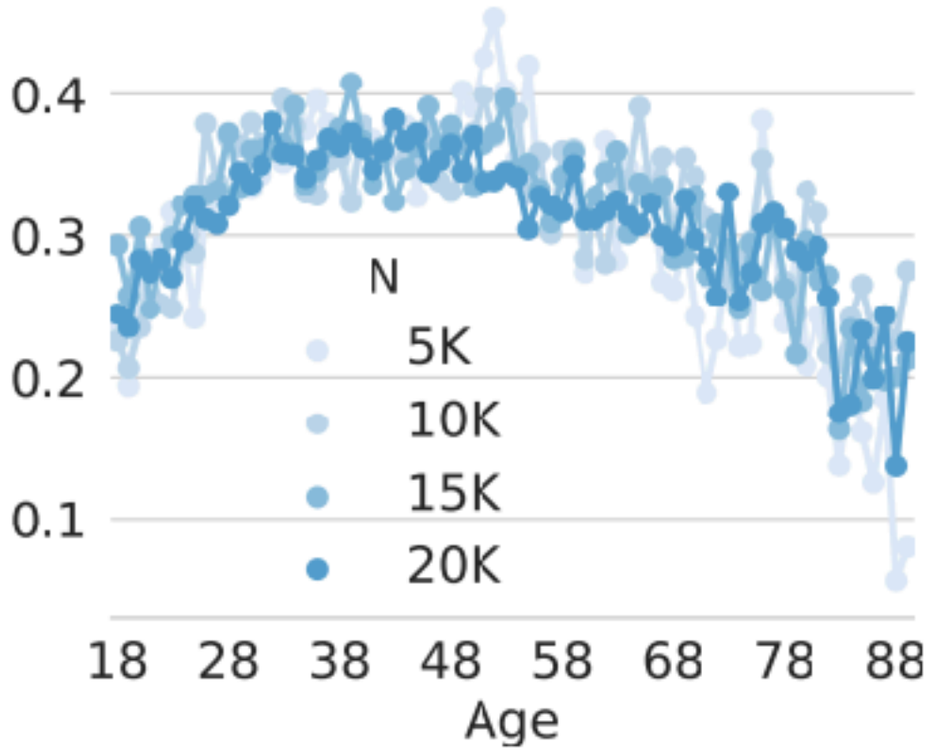


Welfare attitudes experiment

- Evaluate effect of wording on survey results (“welfare” vs “assistance to the poor”)

- Resoundingly positive treatment effects, but significant heterogeneity across covariates





CATE

- To estimate personalized treatment effects, we want to estimate the **conditional average treatment effect (CATE)**

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

- Few different ways to estimate this using black-box ML models
- Again, key challenging is missing data
  - We never observed counterfactuals

S-Learner

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

T-Learner

- Can fit different models over treatment options

X-Learner

Kunzel et al. (2018)

- Regress on the imputed treatment effect Y(1) - Y(0)

- Fit T-learner models and compute imputed treatment effects

$$Y_i - \hat{\mu}_{\theta,0}(X_i) \text{ if } A_i = 1, \hat{\mu}_{\theta,1}(X_i) - Y_i \text{ if } A_i = 0$$

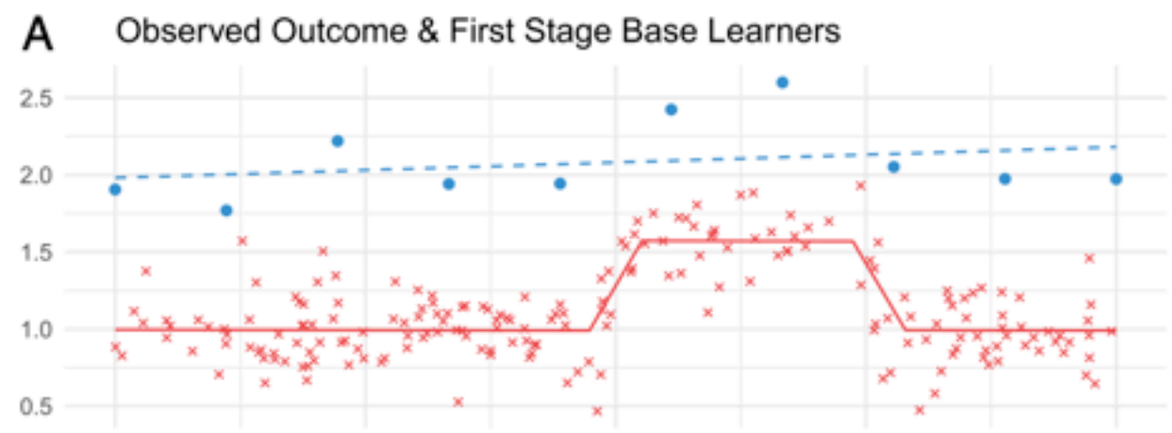
- Fit another set of models $\hat{\tau}_1, \hat{\tau}_0$ on the two category of imputed values, take

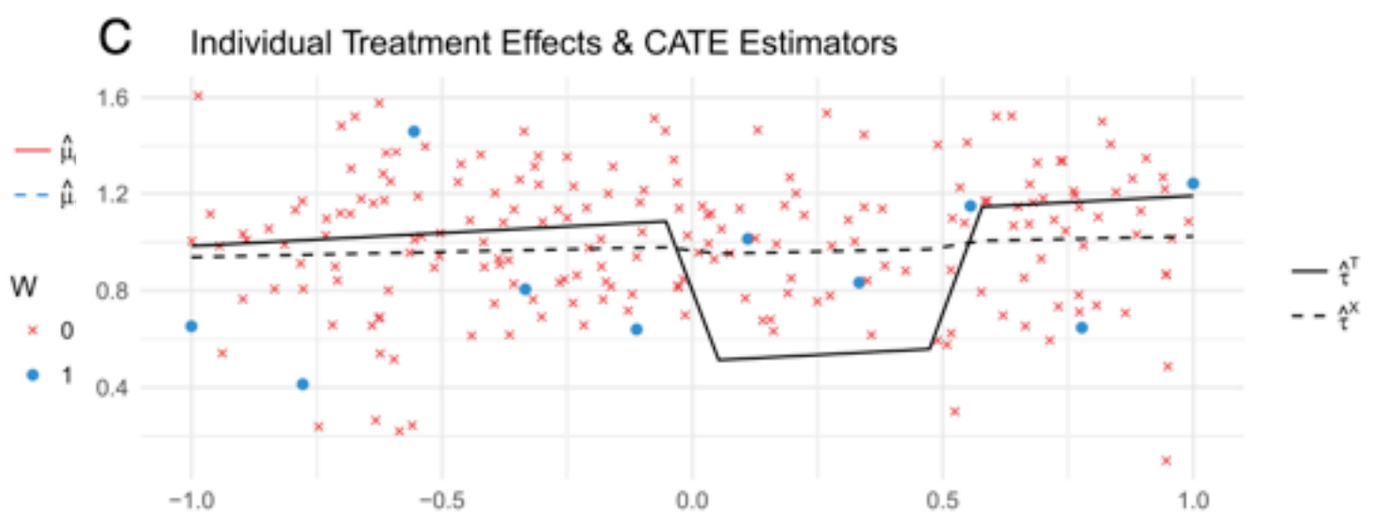
$$\hat{\tau}(X) := \hat{e}(X)\hat{\tau}_0(X) + (1 - \hat{e}(X))\hat{\tau}_1(X)$$

X-Learner

Kunzel et al. (2018)

- Usually, number of samples in treatment >> those in control
- Advantageous if CATE is much smoother than individual outcome functions





R-Learner

Nie and Wager (2020)

R-Learner

Nie and Wager (2020)