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
  - If you predict housing demand based on price, then a prediction model will say high price means high demand

Prediction and causality

- Causal inference is a multi-disciplinary field spanning 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?

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 $Y := Y(A)$

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 policy 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)

Observational studies

- Historically, many important findings from observational data
  - "citrus fruit curing scurvy described in the 1700s or insulin as a treatment for diabetes in the 1920s long preceded the advent of the modern randomized clinical trial."
  - "these methods had in common a reliable method of diagnosis, a predictable clinical course, and a large and obvious effect of the treatment." [Corrigan-Curay et al. 2018]

- These results need to be contextualized and viewed with more skepticism than RCTs

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

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

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

- Under **no unobserved confounding,**

$$\mathbb{E}_P[Y(1)] = \mathbb{E}_P[\mathbb{E}[Y(1) \mid X]] = \mathbb{E}[\mathbb{E}_P[Y(1) \mid X, Z = 1]]$$

- Directly regress Y on X for treated units (Z=1) to get $\mathbb{E}_{\hat{P}}[Y \mid X, Z = 1]$

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.

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

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

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

Assessing overlap

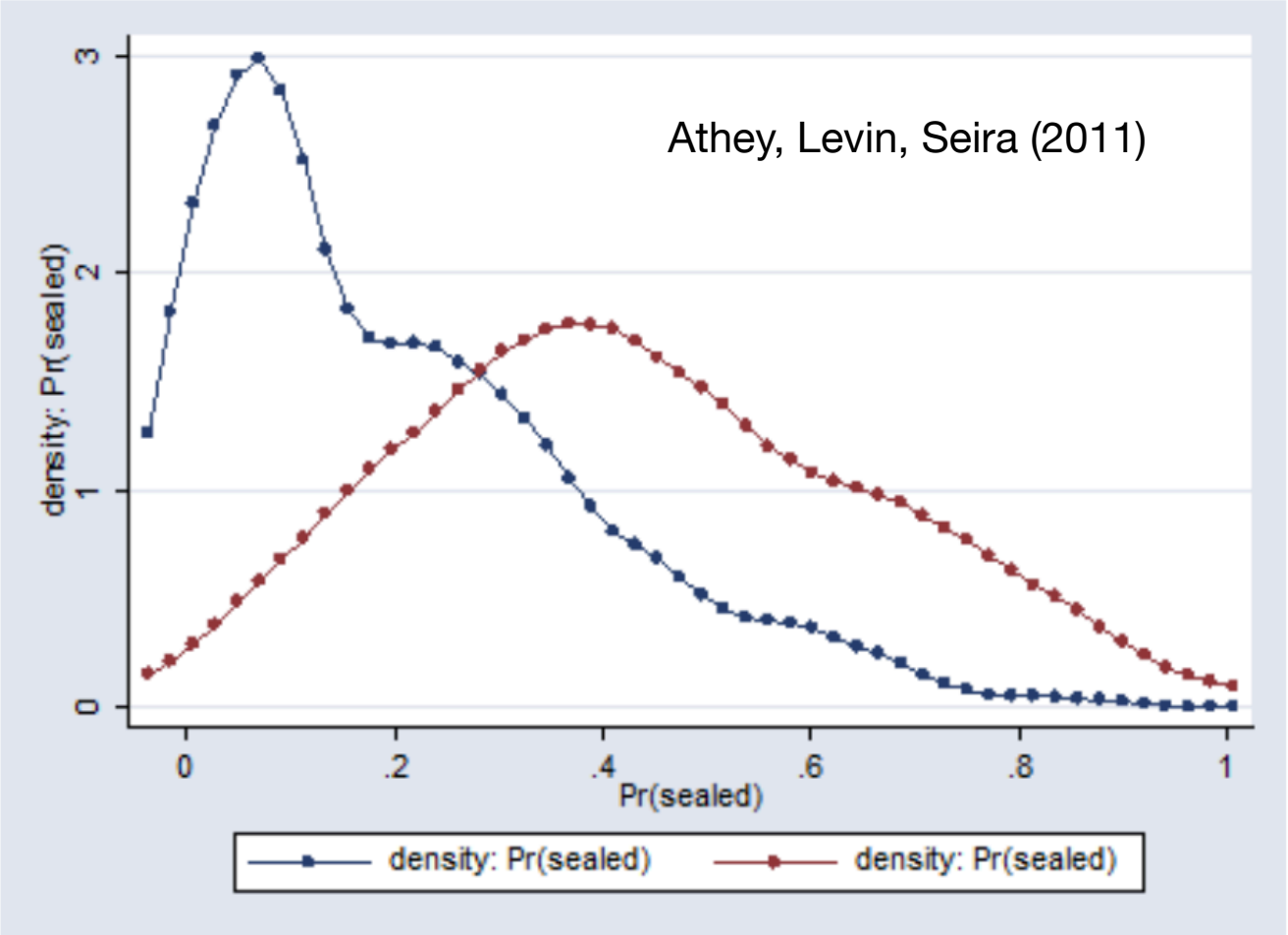
- A common visualization is to look at the pdf of the propensity score across treatment groups
- Plot approximates pdfs of the distribution $\mathbb{P}(e^\star(X) \in \cdot \mid A = a)$
- For each $q \in (0,1)$, plot fraction of observations in the treatment group with $e^\star(x) = q$ (and similarly 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

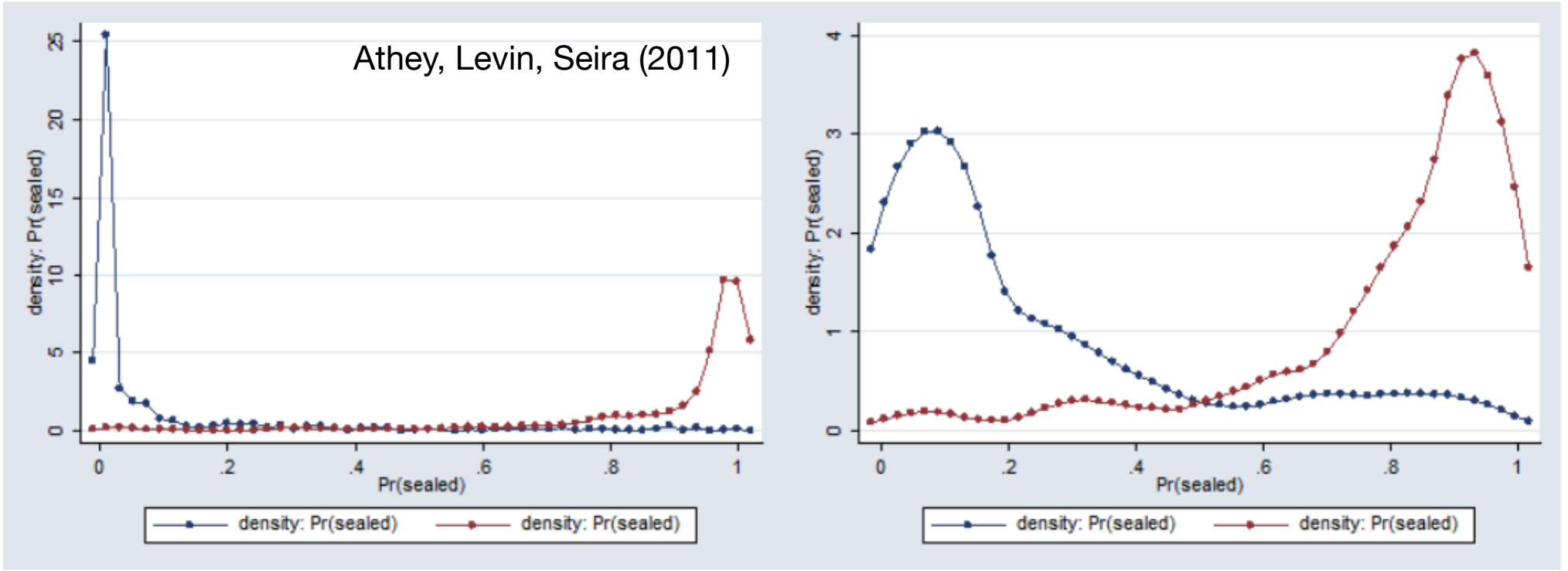
Very few observations with extreme propensity scores



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

Estimators

Direct method

- By no unobserved confounding,

$$\mu_a^\star(X) := \mathbb{E}[Y(a) \mid X] = \mathbb{E}[Y(a) \mid X, A = a]$$
$$= \mathbb{E}[Y \mid X, A = a]$$

observable

- Fit $\mu_a^\star(X)$ via the loss minimization problem

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

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

- Good if the outcome models are easy to learn

Inverse propensity weighting

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

- A natural approach is to reweight samples to correct for confounding bias
  - Essentially importance sampling
  - Reweight treated units to look like “everyone”

- First, estimate the propensity score $e^\star(X) := \mathbb{P}(A = 1 \mid X)$
  - e.g. run logistic regression to predict A given X

Inverse propensity weighting

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

- 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

Unbiasedness

Propensity score $e^*(x) := \mathbb{E}[A|X] = \mathbb{P}(A=1|X)$

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

$$\mathbb{E}\left[\frac{AY}{e^*(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{AY}{e^*(X)}\,|X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{AY}{e^*(X)}\,|X, A=1\right]e^*(X)\right]$$

$$= \mathbb{E}\left[\mathbb{E}[AY\,|X, A=1]\right] = \mathbb{E}\left[\mathbb{E}[Y(1)\,|X, A=1]\right]$$

$$= \mathbb{E}\left[\mathbb{E}[Y(1)|X]\right] \quad \text{by no unobserved confounding}$$

Similarly $\mathbb{E}\left[\frac{(1-A)}{1-e^*(X)}Y\right] = \mathbb{E}[\mathbb{E}[Y(0)|X]] = \mathbb{E}[Y(0)]$.

Estimating propensity score

But e^* is often unknown. How do we estimate e^*?

Logistic regression

$$\log \frac{e(x)}{1-e(x)} = f(x;\theta) \qquad \text{e.g. } f(x;\theta) = \theta^T x$$

Then solve

$$\min_{\theta \in \Theta} \mathbb{E}\left[\log\left(1+\exp\left((2A-1)\cdot f(x;\theta)\right)\right)\right]$$

$A \in \{-1,1\}$

MLE

If model is well-specified, then $\hat{e} \approx e^*$.

Inverse propensity 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

Debiasing

- **Idea: Correct plug-in estimator using the first-order error**

$$\psi(\hat{P}) - \psi(P) = \nabla\psi(\hat{P})^\top(\hat{P} - P) + \mathsf{Rem}_2$$

- Debiased estimator

$$\psi(\hat{P}) - \nabla\psi(\hat{P})^\top(\hat{P} - P)$$

- Automatically achieves second-order error Rem_2

Debiasing

- Outcome model $\mu_1^\star(X) := \mathbb{E}[Y \,|\, X, A = 1]$,
  propensity score $e^\star(X) := \mathbb{P}(A = 1 \,|\, X)$
- Debiasing gives **doubly robust** estimator

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\quad \mu_1^\star(X) \quad + \frac{A}{\mathbb{P}(A = 1 \mid X)}(Y - \mu_1^\star(X)) \right]$$

- Propensity weight residuals to debias the direct method
- Accurate if you can do either well; **insensitive** to errors in nuisance estimates

Debiasing

- Outcome model $\mu_1^\star(X) := \mathbb{E}[Y \mid X, A = 1]$,
  propensity score $e^\star(X) := \mathbb{P}(A = 1 \mid X)$

$$\hat{\tau}_{\text{AIPW}} := \frac{1}{n}\sum_{i=1}^{n} (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$$
$$+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{1 - A_i}{1 - \hat{e}(X_i)}(Y_i - \hat{\mu}_0(X_i))\right)$$

Control variate

- Control variate is a centered RV, such that if you want to estimate E[Y], you estimate E[Y + V] instead
- You get a variance reduction whenever $\mathrm{Var}(V) - 2\mathrm{Cov}(Y, V) < 0$

$$\mathrm{Var}\left(\frac{AT}{e^a(x)} + \left(1 - \frac{A}{e^a(x)}\right) M_i^a(x) \right)$$

$$= \mathrm{Var}\left(\frac{AT}{e^a(x)}\right) + \mathrm{Var}\left(\left(1 - \frac{A}{e^a(x)}\right) M_i^a(x)\right) + 2\mathrm{Cov}\left(\frac{AT}{e^a(x)}, \left(1 - \frac{A}{e^a(x)}\right) M_i^a(x)\right)$$

Control variate

$$\mathrm{Var}\left(\left(1-\frac{A}{e^*(x)}\right)M_1^*(x)\right) = \mathbb{E}\left[M_1^*(x)^2\cdot\left(1-\frac{A}{e^*(x)}\right)^2\right]$$

$$= \mathbb{E}\Big[M_1^*(x)^2\cdot\Big\{e^*(x)\cdot\mathbb{E}\Big[\left(1-\frac{A}{e^*(x)}\right)^2\mid X, A=1\Big]$$

$$+\left(1-e^*(x)\right)\mathbb{E}\Big[\left(1-\frac{A}{e^*(x)}\right)^2\mid X, A=0\Big]\Big\}\Big]$$

$$= \mathbb{E}\left[M_1^*(x)^2\cdot\left\{\frac{(e^*(x)-1)^2}{e^*(x)}+1-e^*(x)\right\}\right]$$

$$= \mathbb{E}\left[M_1^*(x)^2\cdot\frac{1}{e^*(x)}\cdot\left\{e^*(x)^2-2e^*(x)+1+e^*(x)-e^*(x)^2\right\}\right]$$

$$= \mathbb{E}\left[M_1^*(x)^2\left(-1+\frac{1}{e^*(x)}\right)\right].$$

Control variate

$$\mathrm{Cov}\left(\frac{AY}{e^*(x)}, \left(1-\frac{A}{e^*(x)}\right)\mu_1^*(x)\right) = \mathbb{E}\left[\frac{AY}{e^*(x)}\left(1-\frac{A}{e^*(x)}\right)\mu_1^*(x)\right]$$

$$= \mathbb{E}\left[e^*(x)\cdot \mathbb{E}\left[\frac{AY}{e^*(x)}\left(1-\frac{A}{e^*(x)}\right)\mu_1^*(x)\,\middle|\, X, A=1\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{1}{e^*(x)}\right)\cdot Y(1)\cdot \mu_1^*(x)\,\middle|\, X, A=1\right]\right]$$

$$= \mathbb{E}\left[\left(1-\frac{1}{e^*(x)}\right)\cdot \mu_1^*(x)\cdot \mathbb{E}[Y(1)\,|\,X]\right]$$ by no unobs. conf.

$$= \mathbb{E}\left[\left(1-\frac{1}{e^*(x)}\right)\mu_1^*(x)^2\right]$$

Control variate

$$\mathrm{Cov}\left(\frac{AY}{e(x)}, \left(1-\frac{A}{e(x)}\right)\mu_1(x)\right) = \mathbb{E}\left[\frac{AY}{e(x)}\left(1-\frac{A}{e(x)}\right)\mu_1(x)\right]$$

$$= \mathbb{E}\left[e(x)\cdot \mathbb{E}\left[\frac{AY}{e(x)}\left(1-\frac{A}{e(x)}\right)\mu_1(x)\,\middle|\, X, A=1\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{1}{e(x)}\right)\cdot Y(1)\cdot \mu_1(x)\,\middle|\, X, A=1\right]\right]$$

$$= \mathbb{E}\left[\left(1-\frac{1}{e(x)}\right)\cdot \mu_1(x)\cdot \mathbb{E}[Y(1)\,|\,X]\right]$$ by no unobs. conf.

$$= \mathbb{E}\left[\left(1-\frac{1}{e(x)}\right)\mu_1(x)^2\right]$$

So $$\mathrm{Var}\left(\frac{AY}{e(x)} + \left(1-\frac{A}{e(x)}\right)\mu_1(x)\right)$$

$$= \mathrm{Var}\left(\frac{AY}{e(x)}\right) - \mathbb{E}\left[\left(\frac{1}{e(x)}-1\right)\mu_1(x)^2\right]$$ ✓ Variance reduction

Nuisance parameters

- Outcome model $\mu_a^\star(X) := \mathbb{E}[Y \,|\, X, A = a]$,
  propensity score $e^\star(X) := \mathbb{P}(A = 1 \,|\, X)$

- 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

- Consistent estimator of the ATE so long as ***either*** outcome models or propensity score can be estimated consistently

Heuristic derivation

$$\mathbb{E}\left[Y(1) - \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i) + \frac{A_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_1(X_i))\right]$$

$$= \mathbb{E}\left[\left(1 - \frac{e^\star(X)}{\hat{e}(X)}\right)(\mu_1^\star(X) - \hat{\mu}(X))\right]$$

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

- Directional (Gateaux) derivative of functional w.r.t. nuisance parameters around the true values $\gamma^\star = (\mu_1^\star, \mu_0^\star, e^\star, h^\star)$ is zero

Let $\eta = (\mu_0, \mu_1, e)$ be the tuple of nuisance params.

A statistical functional $\mathbb{E}_P \psi(D; \eta)$ is Neyman orthogonal on Λ if

$$\frac{d}{dr} \mathbb{E}_P \psi(D; \eta^* + r(\eta - \eta^*))\Big|_{r=0} = 0 \qquad \forall \eta \in \Lambda$$

↑ Directional derivative at true parameter η^*

$D = (X, Y, A)$. $\eta = (\mu_0, \mu_1, e)$. Let

$$\psi_{AIPW}(D; \eta) := \mu_1(x) - \mu_0(x) + \frac{A}{e(x)}(Y - \mu_1(x)) - \frac{1-A}{1-e(x)}(Y - \mu_0(x))\Big].$$

Orthogonality

Assume $\frac{d}{dr}$ & $\mathbb{E}_p$ are interchangeable throughout.

$$\frac{d}{dr}\mathbb{E}_p\left[\left(\mu_1^* + r(\mu_1 - \mu_1^*)\right)(x)\right]\Big|_{r=0} = \mathbb{E}\left[(\mu_1 - \mu_1^*)(x)\right]$$

$$\frac{d}{dr}\mathbb{E}_p\left[\frac{A}{e^* + r(e - e^*)}\left(Y - \left(\mu_1^* + r(\mu_1 - \mu_1^*)\right)(x)\right)\right]\Big|_{r=0}$$

$$= \mathbb{E}\left[-\frac{A(e - e^*)}{(e^* + r(e - e^*))^2}\left(Y - \mu_1^*(x)\right)\right]\Big|_{r=0} - \mathbb{E}\left[\frac{A}{e^*(x)}(\mu_1 - \mu_1^*)(x)\right]$$

$$= \mathbb{E}\left[-\frac{A(e - e^*)}{e^*(x)^2}\left(Y - \mu_1^*(x)\right)\right] - \mathbb{E}\left[\frac{A}{e^*(x)}(\mu_1 - \mu_1^*)(x)\right]$$

$$= \mathbb{E}\left[e^*(x)\,\mathbb{E}\left[-\frac{A(e - e^*)}{e^*(x)^2}\left(Y - \mu_1^*(x)\right) \middle| X, A = 1\right]\right] - \mathbb{E}\left[(\mu_1 - \mu_1^*)(x)\right]$$

Orthogonality

$$= -\mathbb{E}\left[\frac{(e-e^*)(x)}{e^*(x)}\,\mathbb{E}[Y(1)-\mu_1^*(x)\mid X, A=1]\right] - \mathbb{E}[(\mu_1-\mu_1^*)(x)]$$

$$= -\mathbb{E}[(\mu_1-\mu_1^*)(x)] \qquad \text{by no unobserved confounding}$$

$$\Rightarrow \frac{d}{dr}\mathbb{E}_P[(\mu_1^*+r(\mu_1-\mu_1^*))(x)]\Big|_{r=0} + \frac{d}{dr}\mathbb{E}_P\left[\frac{A}{e^*+r(e-e^*)}\left(Y-(\mu_1^*+r(\mu_1-\mu_1^*))(x)\right)\right]\Big|_{r=0} = 0$$

Similarly for $(1-A)$-related terms.

We conclude $\mathbb{E}_P\psi_{AIPW}$ is Neyman orthogonal at η^*, for all η s.t. $\frac{d}{dr}$ & $\mathbb{E}$ are interchangeable.

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

We use the following decomposition.

$$\frac{1}{n}\sum \psi(D_i;\hat{\eta}) - \mathbb{E}\psi(D;\eta^*) = \underbrace{\frac{1}{n}\sum \psi(D_i;\hat{\eta}) - \mathbb{E}_{D_{np}}\psi(D;\hat{\eta})}_{①} + \underbrace{\mathbb{E}_{D_{np}}\psi(D;\hat{\eta}) - \mathbb{E}\psi(D;\eta^*)}_{②}$$

By the triangular CLT, $\sqrt{n}\cdot ① \Rightarrow N(0, \sigma^2_{NPW})$ if $\hat{\eta} \to \eta^*$

So it suffices to show that $\sqrt{n}\cdot ② \xrightarrow{p} 0$.

Sketch of asymptotics

Define $R(r) := \mathbb{E}_{P_{n,p}}[\psi(\nabla_1 \eta^* + r(\hat{\eta} - \eta^*))]$.

Then ② $= R(1) - R(0)$. If $r \mapsto R(r)$ is cont. diff.

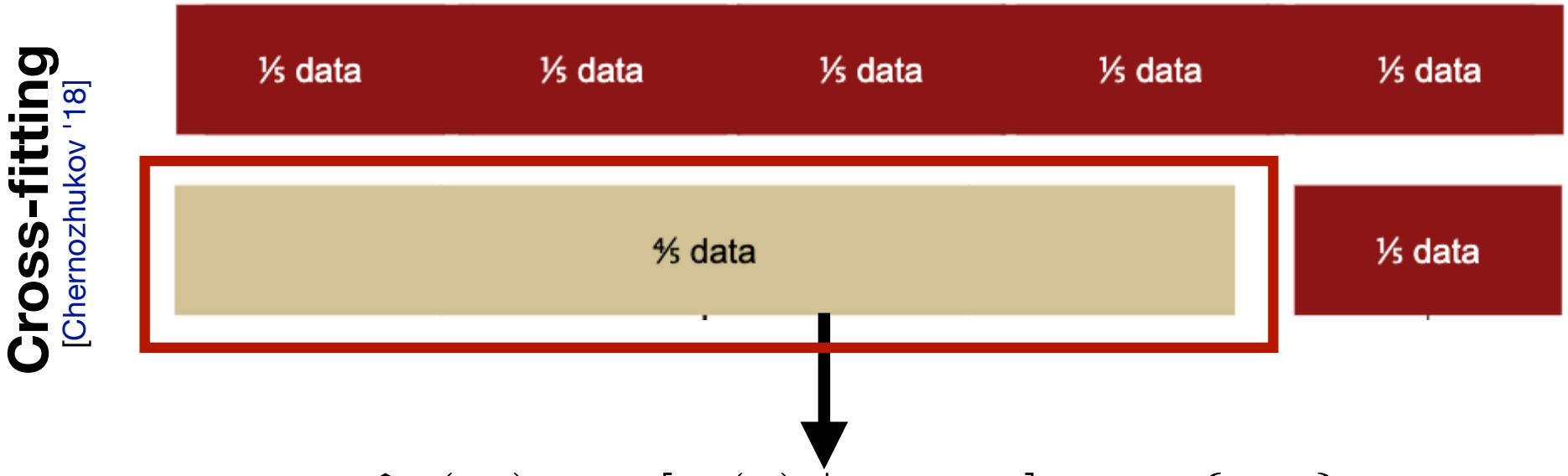
then $= R'(r)$ for some $r \in [0,1]$.

So if we can show $\sup_{r \in [0,1]} R'(r) = o_p(n^{-\frac{1}{2}})$ ... (*)

By orthogonality, $R'(0) = 0$. Since $R'(\cdot)$ is sufficiently smooth, then we can argue that (*) holds under the rate condition

Cross-fitting

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



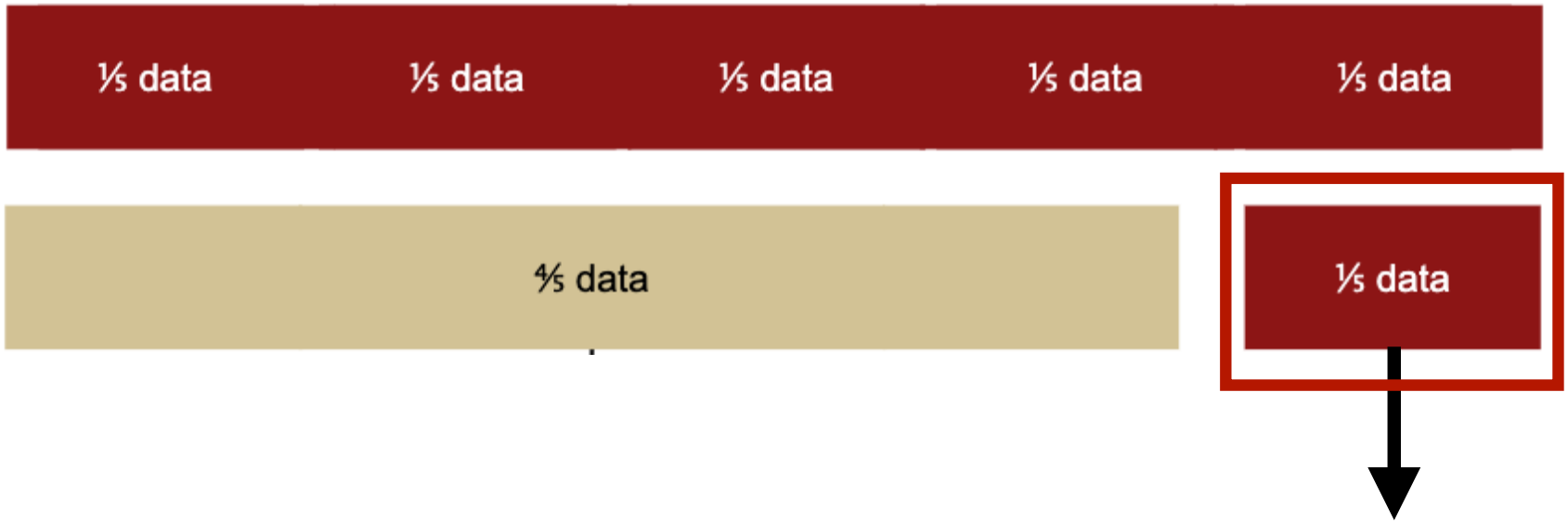
$$\widehat{\mu}_a(X) \approx \mathbb{E}[Y(a) \mid X = x],\ a \in \{0, 1\}$$

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

- 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]

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

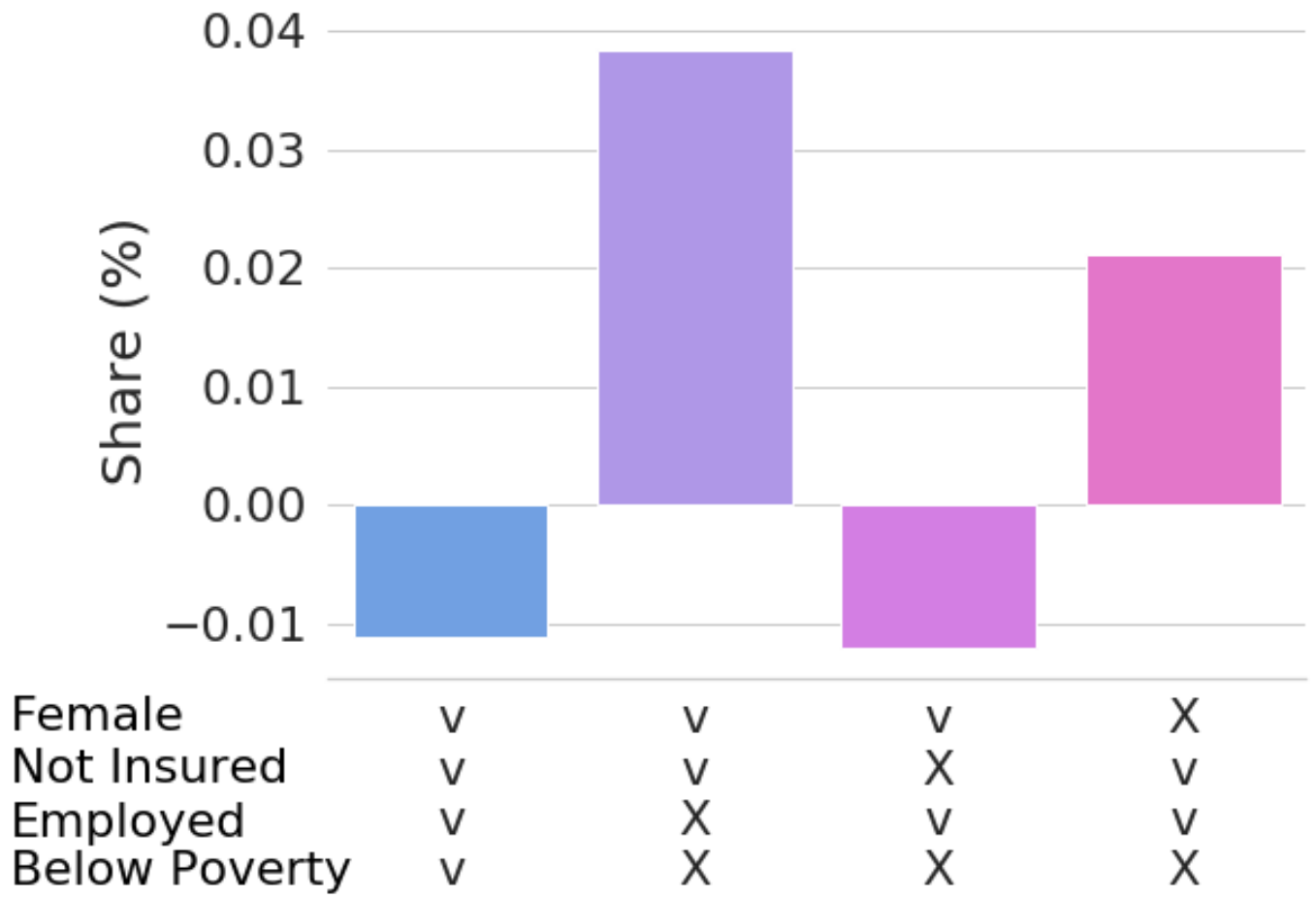
- Same procedure for direct method, IPW
- Similar central limit result follows as before

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



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

Fit a single model for all treatment options

$$Y \leftarrow (X, A)$$

$$\min_{\theta} \mathbb{E}\left[(Y - f_\theta(A, X))^2 \right]$$

Then $\hat{\tau}(x) := f_\theta(1, x) - f_\theta(0, x)$

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

T-Learner

Fit two models, for treatment & control

$$\min_{\theta} \quad \mathbb{E}\left[(Y - \mu_{\theta,1}(X))^2 \mid A = 1 \right]$$

$$\min_{\theta} \quad \mathbb{E}\left[(Y - \mu_{\theta,0}(X))^2 \mid A = 0 \right]$$

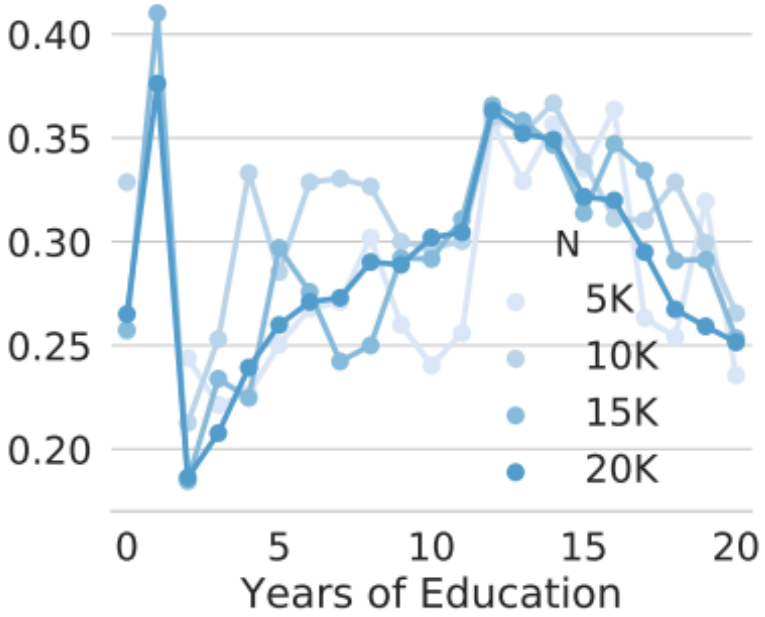
$$\hat{\tau}(x) := \hat{\mu}_{\theta,1}(x) - \hat{\mu}_{\theta,0}(x)$$

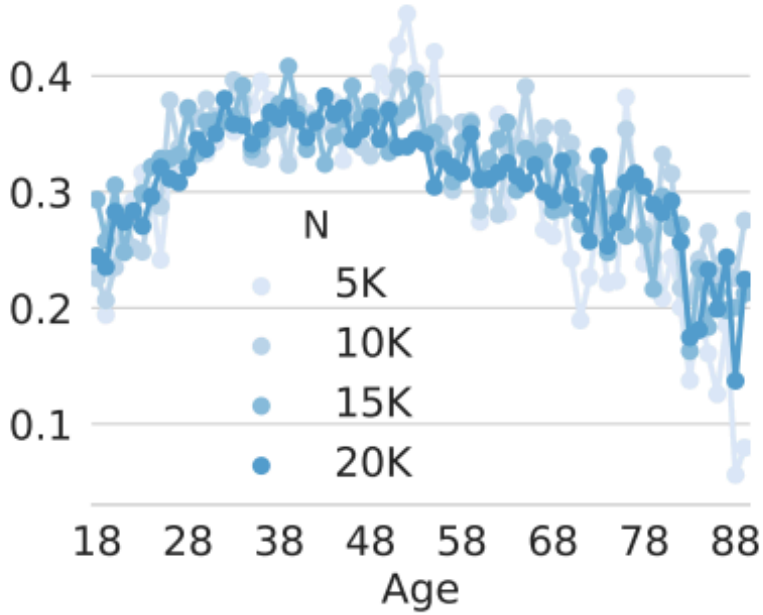
- Can fit different models over treatment options

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





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

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

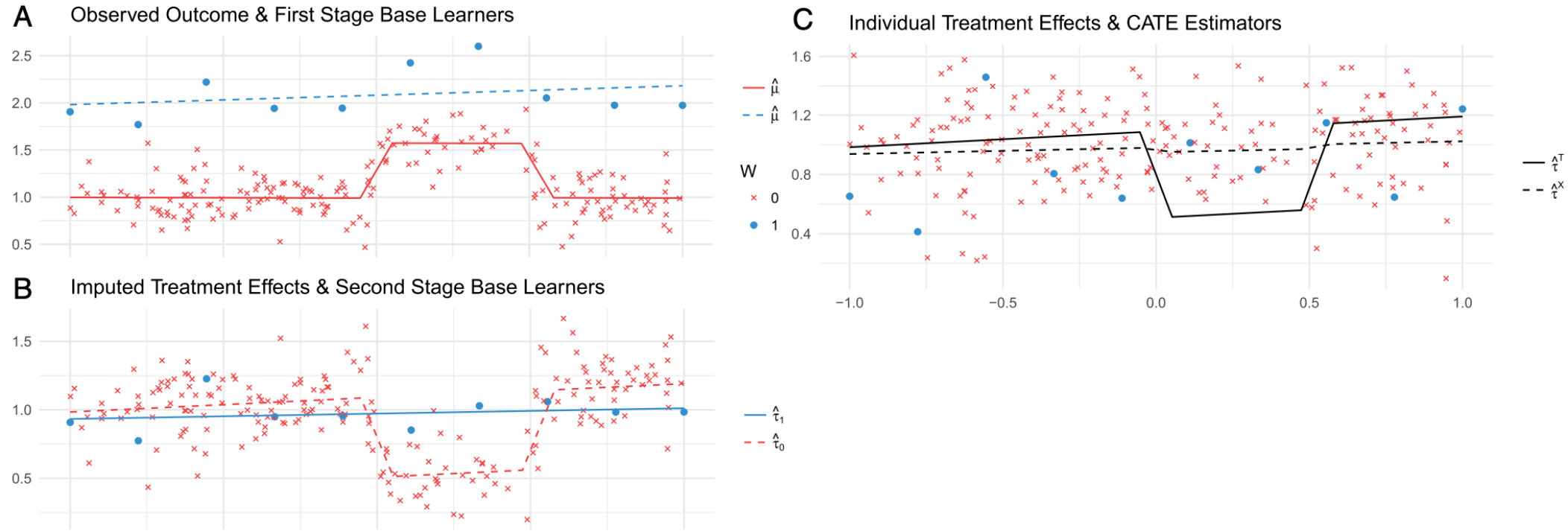
- Fit another set of models $\hat{\tau}_1, \hat{\tau}_0$ on the two category of imputed values, take

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

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)

Robinson's decomposition Define $\varepsilon(a) := Y(a) - (m^*_a(x) + a\tau(x))$, $\varepsilon := \varepsilon(A)$.

From no unobserved confounding,

$$\begin{aligned}
\mathbb{E}[\varepsilon \mid X] &= e^*(x)\mathbb{E}[\varepsilon(1) \mid X, A=1] + (1-e^*(x))\,\mathbb{E}[\varepsilon(0) \mid X, A=0] \\
&= e^*(x)\,\mathbb{E}[Y(1) - m_1^*(x) \mid X, A=1] + (1-e^*(x))\,\mathbb{E}[Y(1) - m_0^*(x) \mid X, A=0] \\
&= e^*(x)\,\mathbb{E}[Y(1) - m_1^*(x) \mid X] + (1-e^*(x))\,\mathbb{E}[Y(1) - m_0^*(x) \mid X] \\
&= 0
\end{aligned}$$

By def of ε, $\quad Y = m_0(x) + A\tau(x) + \varepsilon \quad \cdots (*)$

Define $m^*(x) := \mathbb{E}[Y \mid X]$. Taking $\mathbb{E}[\cdot \mid X]$ on both sides

$$m^*(x) = m_0(x) + e^*(x)\tau(x).$$

R-Learner

Nie and Wager (2020)

Subtracting this from (*), we arrive at

$$Y - m^*(x) = (A - e^*(x))\tau(x) + \varepsilon .$$

So if we fit $\hat{m}, \hat{e}$ on holdout data, we can now solve

$$\min_{\theta} \mathbb{E}\left[\left(Y - \hat{m}(x) - (A - \hat{e}(x))\tau_\theta(x)\right)^2 \right]$$

to get $\hat{\tau}_\theta(x)$.