ML and Causal Inference I: HTE

PS690 Computational Methods in Social Science

Jiawei Fu
Department of Political Science
Duke University

Sep 16, 2025

Overview

Framework of Potential Outcomes

- We hope to study the causal effect of a binary treatment on an outcome Y.
- Consider a finite population with size n. The treatment vector is $Z = (Z_1, ..., Z_n)$. The potential outcome for individual i is $Y_i(Z)$.

Assumption (No Interference)

Individual i's potential outcomes do not dependent on other individual's treatments.

- Therefore, we can simplify $Y_i(Z)$ as $Y_i(Z_i)$. In other words, for each individual i, there exists potential outcomes: $(Y_i(1), Y_i(0))$.

Assumption (Consistency)

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

- Therefore, the observed outcome is $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$
- People often call Stable Unit Treatment Value Assumption (STUVA) assumption by combing no interference and consistency.

Framework of Potential Outcomes

- Individual causal effect (ITE): $\tau_i = Y_i(1) - Y_i(0)$
- We can only observe one potential outcome for each individual; therefore, it is hard to identify the ITE.
- The finite population average treatment effect (ATE) is

$$\tau = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)] = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0).$$

- Under completely randomized experiment, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp Z_i$.
- Then, finite population ATE can be unbiasedly estimated by difference-in-means estimator

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i) Y_i,$$

where $n_1 = \sum_{i=1}^{n} Z_i$ is the number of individuals are randomly assigned to treatment, and $n_0 = \sum_{i=1}^{n}(1 - Z_i)$ is the sample size of control group.

Framework of Potential Outcomes

- It is easy to show that $\mathbb{E}[\hat{\tau}] = \tau$. (show it; what is the randomness?)
- The variance is

$$Var(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n},$$

  where $S^2(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(Z) - \overline{Y}_i(Z))^2$ is the population variance of potential outcomes, and $S^2(\tau) = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \overline{\tau}_i)^2$ is the population variance of ITE.
- Define $S(1,0) = \frac{1}{n-1}\sum_{i=1}^{n}[Y_i(1) - \overline{Y}_i(1)][Y_i(0) - \overline{Y}_i(0)]$ as the population covariance.
- (Advanced) Because $2S(1,0) = S^2(1) + S^2(0) - S^2(\tau)$, the variance can also be written as

$$Var(\hat{\tau}) = (\frac{n_0}{n_1 n})S^2(1) + (\frac{n_1}{n_0 n})S^2(0) + \frac{2}{n}S(1,0)$$

- In ether case, the last term is not identified because we never observe the joint distribution of $(Y_i(1), Y_i(0))$.

Framework of Potential Outcomes

- However, we can unbiasedly estimate the first two terms.
- Let $\hat{S}^2(Z) = \frac{1}{n-1} \sum_{i=1}^{n} Z_i [Y_i(1) - \overline{Y}_i(1)]^2$ denote sample variances.
- We obtain the conservative variance estimator

$$\widehat{Var}(\hat{\tau}) = \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$$

- It is conservative because $\mathbb{E}[\widehat{Var}(\hat{\tau})] \geq Var[\hat{\tau}]$.
- When will we have equality?

Framework of Potential Outcomes

- Now, let us consider super-population framework.
- We assume the sample we have is randomly drawn from a super population.
- We still consider completely randomized experiment.
- Now, the super-population ATE is $\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]$ (can you see why the second equality hold?).
- Therefore, it is identifiable. Its form suggests the moment estimator $\hat{\tau}$, difference-in-means estimator as before.
- It is easy to show that $\hat{\tau}$ is unbiased to population ATE τ. (show it; what is the randomness here now?)
- The variance is $Var(\hat{\tau}) = \frac{Var(Y_i(1))}{n_1} + \frac{Var(Y_i(0))}{n_0}$.
- Therefore, the variance estimator $\widehat{Var}(\hat{\tau})$ is unbiased now.

Observation Studies

- From now on, we consider super-population framework, and observational studies.
- Causal inference with observational studies is challenging. It relies on strong assumptions.

Assumption (Ignorability)

$Y_i(z) \perp\!\!\!\perp Z_i | X_i \ (z = 0, 1)$

- The above assumption has many names: ignorability due to Rubin; unconfoundedness which is popular among epidemiologists; selection on observables which is popular among social scientists; conditional independence which is merely a description of the notation in the assumption.

Assumption (Strong Ignorability)

$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i | X_i$

Observation Studies

- Another important concept in the conversational study is propensity score $\pi(X_i) = \mathbb{P}[Z_i = 1|X_i]$.

Assumption (Overlap/Positivity)

$0 < \pi(X_i) = \mathbb{P}[Z_i = 1|X_i] < 1$

- If it is not hold, for example $\pi(X_i) = 1$, we always observe $Y_i(1)$, $Y_i(0)$ is not well defined. Basically, we require some residual randomness in the treatment conditional on the covariate.
- There is a very important result about propensity score: if conditional on high dimensional X can guarantee strong ignorability, then conditioning on lower dimension $\pi(X)$ can also remove all confounding induced by X.

Theorem

If $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\}|X_i$, *then* $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\}|\pi(X_i))$

Estimation via Outcome Regression

- Now, we will see that ATE $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ is identifiable, and introduce several estimators follow [Tsiatis et al., 2019].
- First, note that expectation of potential outcome is identifiable.

$$\begin{aligned}\mathbb{E}[Y_i(1)] &= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i]] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i, Z_i = 1]] \quad \text{ignorability and positivity} \\ &= \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 1]] \quad \text{STUVA}\end{aligned}$$

- Note that, in the final equation, the expectation is over the marginal distribution of X, not the conditional distribution of $X|Z = 1$.
- By analogous, $\mathbb{E}[Y_i(0)] = \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 0]]$
- Therefore, $\tau = \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 1]] - \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 0]]$.
- It shows that τ depends on the regression of observed outcome on covariates and treatment received.

Estimation via Outcome Regression

- We use $Q(x, z)$ to denote regression relationship $\mathbb{E}[Y_i|X_i = x, Z_i = z]$.
- Therefore $\tau = \mathbb{E}[Q(X_i, 1)] - \mathbb{E}[Q(X_i, 0)]$
- People often consider parametric model: $Q(x_i, z; \beta)$. For example,
  1. $Q(x, z; \beta) = \beta_1 + \beta_2 z + \beta_3^T x$
  2. $Q(x, z; \beta) = \beta_1 + \beta_2 z + \beta_3^T x + \beta_4 xz$
  3. $logit\{Q(x, z; \beta)\} = \beta_1 + \beta_2 z + \beta_3^T x + \beta_4 xz$
- We then obtain the estimate $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} [Q(x_i, 1; \hat{\beta}) - Q(x_i, 0; \hat{\beta})]$
- Note, if $Q(x, z; \beta) = f(x; \beta_1) + \beta_2 z$ (like equation 1), where $f(x; \beta_1)$ is an arbitrary function of x and β, then $\tau = \mathbb{E}[Q(X_i, 1)] - \mathbb{E}[Q(X_i, 0)] = \beta_2$.
- Therefore, for models with no treatment-covariate interaction terms, the estimator for ATE can be directly from the fit of the model Q, and no need to fit two and get difference.

Estimation via Outcome Regression

- In general, we can use other more complex models to estimate the causal effects.
- Define $\mu_1(X_i) = \mathbb{E}[Y_1(1)|X_i] = \mathbb{E}[Y|Z = 1, X]$ and $\mu_0(X_i) = \mathbb{E}[Y_i|Z = 0, X_i]$
- We can build two predictors $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$ based on the treated and control data, respectively.
- They can even be estimated non-parametrically. You can use any ML methods.
- Then, $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$.
- The biggest problem with the outcome regression approach is its sensitivity to the specification of the outcome model.

Estimation via Propensity Score

- Let us see the second Propensity Score approach.
- Recall, we use $\pi(X_i) = \mathbb{P}[Z_i = 1|X_i]$ to denote propensity score.
- We want to show that $\mathbb{E}[Y_i(1)] = \mathbb{E}[\frac{Z_i Y_i}{\pi(X_i)}]$ and $\mathbb{E}[Y_i(0)] = \mathbb{E}[\frac{(1-Z_i)Y_i}{1-\pi(X_i)}]$.
- Therefore, $\tau = \mathbb{E}[\frac{Z_i Y_i}{\pi(X_i)} - \frac{(1-Z_i)Y_i}{1-\pi(X_i)}]$

$$
\begin{aligned}
\mathbb{E}[\frac{Z_i Y_i}{\pi(X_i)}] &= \mathbb{E}[\mathbb{E}[\frac{Z_i Y_i(1)}{\pi_i(X_i)}|X_i]] \\
&= \mathbb{E}[\frac{1}{\pi_i(X_i)}\mathbb{E}[Z_i Y_i(1)|X_i]] \\
&= \mathbb{E}[\frac{1}{\pi_i(X_i)}\mathbb{E}[Z_i|X_i]\mathbb{E}[Y_i(1)|X_i]] \quad \text{Strong Ignorability} \\
&= \mathbb{E}[\frac{1}{\pi_i(X_i)}\pi(X_i)\mathbb{E}[Y_i(1)|X_i]] \\
&= \mathbb{E}[Y_i(1)]
\end{aligned}
$$

Estimation via Propensity Score

- Then, a natural estimator for $\mathbb{E}[Y_i(1)] = \mathbb{E}[\frac{Z_i Y_i}{\pi(X_i)}]$ is IPW estimator: $\frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{\pi(X_i)}$
- Therefore, $\hat{\tau}^{IPW} = \frac{1}{n}\sum_{i=1}^{n}[\frac{Z_i Y_i}{\pi(X_i)} - \frac{(1-Z_i)Y_i}{1-\pi(X_i)}] = \frac{1}{n}\sum_{i=1}^{n}[\frac{Z_i-\pi_i}{\pi_i(1-\pi_i)}]Y_i$.
- It is also called the Horvitz-Thompson (HT) estimator.
- If $\pi(X_i)$ is not known, we can model it as $logit(\pi(x;\gamma)) = \gamma_1 + \gamma_2 x$ or $\pi(x) = \frac{exp(\gamma_1+\gamma_2 x)}{1+exp(\gamma_1+\gamma_2 x)}$.
- Then $\hat{\tau}^{IPW} = \frac{1}{n}\sum_{i=1}^{n}[\frac{Z_i Y_i}{\pi(X_i;\hat{\gamma})} - \frac{(1-Z_i)Y_i}{1-\pi(X_i;\hat{\gamma})}]$
- We can view that $\frac{1}{\pi(X_i)}$ and $\frac{1}{1-\pi(X_i)}$ as the inverse probability weight; it is a weighted average.

Estimation via Propensity Score

- Note, counterintuitively, even if you know true value of γ, using MLE $\hat{\gamma}$ in $\hat{\tau}^{IPW}$ will be more efficient.
- For example, consider a randomized study that $\pi(x;\gamma) = \gamma = \frac{1}{2}$.
- We can also estimate $\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.
- Then,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\pi(X_i;\gamma)} = \sum_{i=1}^{n}\frac{Z_i Y_i}{n/2} = \hat{\mu}_1$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\pi(X_i;\hat{\gamma})} = \frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i} = \overline{Y}_1$$

Estimation via Propensity Score

- Both of then are consistent for μ_1.
- Consider $\sqrt{n}(\hat{\mu}_1 - \mu_1) = 2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i Y_i - \mu_1/2)$ first.
- Define $\sigma_1^2 = Var[Y_i|Z_i = 1]$

$$\begin{aligned} Var[Z_i Y_i] &= \mathbb{E}[Var(Z_i Y_i|Z_i)] + Var[\mathbb{E}[Z_i Y_i|Z_i]] \\ &= \mathbb{E}[Z_i \sigma_1^2] + Var[Z_i \mu_1] \\ &= \frac{\sigma_1^2}{2} + \frac{\mu_1^2}{4} \end{aligned}$$

- Then, by CLT, $\sqrt{n}(\hat{\mu}_1 - \mu_1) \to N(0, 2\sigma_1^2 + \mu_1^2)$

Estimation via Propensity Score

- Consider $\sqrt{n}(\overline{Y}_1 - \mu_1) = [\frac{1}{n}\sum_{i=1}^n Z_i]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i(Y_i - \mu_1)$ now.

$$\begin{aligned} Var[Z_i(Y_i - \mu_1)] &= \mathbb{E}[Z_i Var(Y_i - \mu_1|Z_i)] + Var[Z_i\mathbb{E}[Y_i - \mu_1|Z_i]] \\ &= \mathbb{E}[Z_i\sigma_1^2] + 0 \\ &= \frac{\sigma_1^2}{2} \end{aligned}$$

- Then, by CLT, $\sqrt{n}(\overline{Y}_1 - \mu_1) \to N(0, 2\sigma_1^2)$
- Thus, usual estimator $\overline{Y}_1$, estimated propensity score, is more efficient than $\hat{\mu}_1$, which uses the know propensity score.

Doubly Robust Estimation for Causal Effects

- [Tsiatis, 2006] shows that all consistent and asymptotically normal estimators for ATE τ are asymptotically equivalent to the follow estimator

$$\hat{\tau}^{AIPW} = \frac{1}{n}\sum_{i=1}^{n}[\frac{Z_i Y_i}{\pi(X_i;\hat{\gamma})} - \frac{(1-Z_i)Y_i}{1-\pi(X_i;\hat{\gamma})} - (Z_i - \pi(X_i;\hat{\gamma}))h(X_i)]$$

  where $h(X_i)$ is any arbitrary function of X_i.
- The IPW estimator is a special case with $h(X) = 0$.
- The additional term in the estimator is often said to augment the simple IPW estimator so as to increase efficiency. Therefore, the estimator is called AIPW.
- Among all AIPW estimators, the most efficient one is obtained by choosing $h(X_i) = \frac{\mathbb{E}[Y_i|X_i,Z=1]}{\pi(X_i)} + \frac{\mathbb{E}[Y_i|X_i,Z=0]}{1-\pi(X_i)}$.
- We know we can model and estimate $\mathbb{E}[Y_i|X_i, Z_i = 1]$ through outcome regression.

Doubly Robust Estimation for Causal Effects

- Therefore, we get double robust estimator

$$\hat{\tau}^{DR} = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{Z_i Y_i}{\pi(X_i;\hat{\gamma})} - \frac{(1-Z_i)Y_i}{1-\pi(X_i;\hat{\gamma})} - \frac{Z_i - \pi(X_i;\hat{\gamma})}{\pi(X_i;\hat{\gamma})}\mu_1(X_i;\hat{\beta}_1) - \frac{Z_i - \pi(X_i;\hat{\gamma})}{1-\pi(X_i;\hat{\gamma})}\mu_0(X_i;\hat{\beta}_0)\Big]$$

- It argument the IPW by the imputed outcomes.
- We can also write it as

$$\hat{\tau}^{DR} = \frac{1}{n}\sum_{i=1}^{n}\Big[\mu_1(X_i;\hat{\beta}_1) - \mu_0(X_i;\hat{\beta}_0) + \frac{Z_i(Y_i - \mu_1(X_i;\hat{\beta}_1))}{\pi(X_i;\hat{\gamma})} - \frac{(1-Z_i)(Y_i - \mu_0(X_i;\hat{\beta}_0))}{1-\pi(X_i;\hat{\gamma})}\Big]$$

- It augments the outcome regression estimator by inverse propensity score weighting terms of the residuals.

Doubly Robust Estimation for Causal Effects

- It shows that one need model correctly only one of the propensity score or the outcome regression to render $\hat{\tau}^{DR}$ a consistent estimator.
- Thus, it has built-in protection against, or is robust to, misclassifications. We have two tries to develop correct model. Therefore, it is referred to as doubly robust.
- Statistically, it turns out that AIPW not only inherits robustness properties from both—it improves on both by using IPW to mitigate errors in the regression estimator and vice-versa (learn more in the next lecture).

Doubly Robust Estimation for Causal Effects

- First, let us look at the Weak Double robustness: AIPW is consistent if either $\hat{\mu}_1(X_i)$ are consistent or $\hat{\pi}(X_i)$ is consistent.
- Consider $\hat{\mu}_1(X_i)$ is consistent, i.e. $\hat{\mu}_z(X_i) \approx \mu_z(X_i)$, then,

$$\hat{\tau}^{DR} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{Z_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1-Z_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)}]$$

- The second part ≈ 0 because $\mathbb{E}[Y_i - \hat{\mu}_z(X_i)|X_i, Z_i] \approx 0$ under unconfoundedness.
- Thus even if we use inconsistent propensity score weights $\frac{1}{\hat{\pi}}$ and $\frac{1}{1-\hat{\pi}}$, they are multiplied by roughly mean-zero error terms and so asymptotically they do not bias the estimator.

Doubly Robust Estimation for Causal Effects

- Conversely, now suppose $\hat{\pi}(X_i)$ is consistent, i.e. $\hat{\pi}(X_i) \approx \pi_i(X_i)$, then

$$\hat{\tau}^{DR} = \frac{1}{n}\sum_{i=1}^{n}[\frac{Z_i Y_i}{\pi(X_i;\hat{\gamma})} - \frac{(1-Z_i)Y_i}{1-\pi(X_i;\hat{\gamma})} - \frac{Z_i - \pi(X_i;\hat{\gamma})}{\pi(X_i;\hat{\gamma})}\mu_1(X_i;\hat{\beta}_1) - \frac{Z_i - \pi(X_i;\hat{\gamma})}{1-\pi(X_i;\hat{\gamma})}\mu_0(X_i;\hat{\beta}_0)]$$

- The second part ≈ 0 because $\mathbb{E}[\frac{Z_i}{\hat{\pi}(X_i)} - 1|X_i] \approx 0$.
- Thus, even if we use inconsistent regression adjustments $\hat{\mu}_z(X_i)$, they will be multiplied by roughly mean-zero noise terms that asymptotically cancel their contribution.

Doubly Robust Estimation for Causal Effects

- There is also a much more interesting and useful class of "strong" double robustness results for AIPW.
- In general, it claims that if we use estimators $\hat{\mu}_z(x)$ and $\hat{\pi}(x)$ that are both consistent with root-mean squared error (RMSE) decaying faster than $n^{-\phi_\mu}$ and $n^{-\phi_\pi}$ respectively, and if $\phi_\mu + \phi_\pi \geq 1/2$, then

$$\sqrt{n}(\hat{\tau}^{AIPW} - \pi) \to N(0, V),$$

  where $V = Var(\tau(X_i)) + \mathbb{E}[\frac{\sigma_0^2(X_i)}{1-\pi(X_i)}] + \mathbb{E}[\frac{\sigma_1^2(X_i)}{\pi(X_i)}]$
- This is a very exciting result! Suppose that we use ML methods to estimate $\hat{\mu}_z(x)$ and $\hat{\pi}(x)$, we know that typically they have a slow rate, slower than $n^{-1/2}$ due to regularization. However, AIPW can make the bias substantially smaller than what regression or IPW estimators could achieve on their own.
- To be specific, if their decay rater is faster than $n^{-\phi_\mu}$ and $n^{-\phi_\pi}$, the bias of AIPW decays faster than $n^{-(\phi_\mu+\phi_\pi)}$; and in particular $\phi_\mu + \phi_\pi \geq 1/2$, then the bias is lower-order on the $1/\sqrt{n}-$scale. We will learn more about it in DML.

Conditional ATE

- Define conditional average treatment effect (CATE) as $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$.
- Actually, We have seen before, $\mathbb{E}[Y_i(Z)|X_i = x]$ is identifiable under ignorability and overlap:

$$\mu_z(x) := \mathbb{E}[Y_i(z)|X_i = x] = \mathbb{E}[Y_i(z)|Z_i = z, X_i = x] = \mathbb{E}[Y_i|Z_i = z, X_i = x].$$

- Our target causal parameter of interest is actually a function or the value of a function at a particular point.
- Suppose there exists high-dimensional controls W, and one only care about the CATE with some low-dimensional covariate X, and unconfoundedness holds with W, then the Identifying equation is: $\tau_0(X) = \mathbb{E}[\mathbb{E}[Y|D = 1, W] - \mathbb{E}[Y|D = 0, Z]|X]$

Meta-Learner

- We aim at constructing an estimate $\hat{\tau}(X)$ of the true CATE function $\tau_0(X)$.
- The key is to decompose the estimation of the CATE into a sequence of regression problems.
- This approach has been coined meta-learning since we are trying to treat ML techniques as a black-box oracle that solves any regression problem and we are trying to build on top of that oracle to learn the CATE.

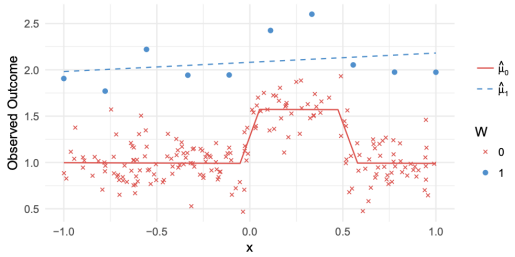
S-learner

- S-learner, also known as a "single learner".
- fit a single model for the response as a function of X and the treatment Z: $\mu(z, x) = \mathbb{E}[Y_i | Z_i = z, X_i = x]$
- Then estimate CATE by $\hat{\tau}(x) = \hat{\mu}(1, x) - \hat{\mu}(0, x)$
- If using regularization (which is quite common given high-dimensional controls), it tends to bias the treatment effect towards zero.
- For this reason, it seems natural to weaken this regularization bias on the treatment. This can be achieve by fitting two separate models: $T - learner$.

T-learner

- T-learner, also know as Two learners.
- Using control group to estimate $\mu_0(x) = \mathbb{E}[Y_i|Z_i = 0, X_i]$, and using treatment group to estimate $\mu_1(x) = \mathbb{E}[Y_i|Z_i = 1, X_i]$
- Then estimate CATE by $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$
- Performs well if the treatment effect is very complicated.
- Problem: Regularization bias: Given that we fit $\hat{\mu}_1$ and $\hat{\mu}_0$ separately, these two functions may end up being regularized in different ways from each other, creating artifacts in the learned CATE estimate.

Problem of T-learner



- To estimate $\mu_1(x)$, we must be careful not to overfit the data since we observe only 10 data points.
- We might decide to use a linear model (dashed line)
- For control group, we end up a nonlinear function.
- T-learner would jump at 0 and 0.5, while the true $\tau(x)$ is constant.

X-learner

- X-learner, Cross Learner, uses two important observations.
- First, Conditional Average Treatment Effect on the Treated (CATT) is equal to the Conditional Average Treatment Effect on the Control (CATC):

$$CATT = \mathbb{E}[Y(1) - Y(0)|X, Z = 1] = \mathbb{E}[Y(1) - Y(0)|X] = CATE,$$

similar for CATC.

X-learner

- Second, for CATT and CATC, we only need to identify one counterfactual outcome.

$$CATT = \mathbb{E}[Y - \mathbb{E}[Y|X, Z = 0]|X, Z = 1]$$
$$CATC = \mathbb{E}[\mathbb{E}[Y|X, Z = 1] - Y|X, Z = 0]$$

- This yields two ways of identify the CATE, and any convex combination of these two will also be a valid identification strategy for the CATE.
- If π_i is small for some X, which means there are more data in the control group, we can focus on CATT, where we only need to estimate counterfactuals under control.
- Similar for π_i is large for some X.

X-learner [Künzel et al., 2019]

- In summary, identical to the T-learner, estimate $\mu_0(x)$ and $\mu_1(x)$ first.
- Impute missing potential outcomes for individual i by $\mu_{1-z}(X_i)$, and compute ITE as $\hat{\tau}_i$
- Employ any supervised learning or regression method(s) to estimate conditional effects in two groups: using the imputed treatment effects as the response variable in the treatment group to obtain $\hat{\tau}_1(x)$, and similarly in the control group to obtain $\hat{\tau}_0(x)$
- The CATE is estimated as a weighted average: $\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$, where $g(x) \in [0, 1]$ is a weight function. A good choice is propensity score $\hat{\pi}(X_i)$.
- The weighting serves to pull the final estimate closer to the estimated effect that relies on the conditional mean function estimated in the larger group.

DR-Learner

- The above approaches rely fully on accurate outcome modeling.
- If the model cannot capture the real complicity of the true outcomes $\mathbb{E}[Y|Z, X]$, meta-learners will suffer from large estimation errors.
- Use the similar idea of DR estimator in causal inference, we can get robust and efficient estimation based on Doubly-Robust approach.
- After get $\mu_0(x)$ and $\mu_1(x)$, construct Pseudo-outcome

$$\hat{\tau}^{DR}(X_i) = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{Z_i[Y_i - \hat{\mu}_1(X_i)]}{\hat{\pi}(X_i)} - \frac{(1 - Z_i)[Y_i - \hat{\mu}_1(X_i)]}{1 - \hat{\pi}(X_i)}$$

- Then, $\hat{\tau}^{DR}(X_i)$ is regressed on the observed covariates to obtain the DR-learner's final CATE estimate $\hat{\tau}(x)$.

Comparison

- S/T-Learner: they heavily rely on correct outcome modelling.
- Although the S-learner and T-learner can perform well in particular settings, simulation studies found them to be overall outperformed by the pseudo-outcome methods
- If we believe that the CATE function is simpler than the response functions under treatment or control, X-learner should be preferred.
- DR-Learner: possesses doubly robust properties. However, contrary to the S/T/X-learners, when the true DGP has extreme propensities in parts of the covariate space, then the DR-Learner can have high variance and become unstable.
- All-in-all, one should note that there is no clear winner among the X-, R- and DR-Learner methods and each can potentially be the best performer in different contexts.
- There are many other learners! Probably, in the future, you will have your own learner.

Test of CATE

- Given selected CATE model, we want to run formal statistical tests that validate whether the model that we chose contains any signal of treatment effect heterogeneity.
- In other words, we need a formal statistical test of the presence of treatment effect heterogeneity.
- A core difficulty of applying ML tools to the estimation of heterogeneous causal effects is that, while they are successful in prediction empirically, it is much more difficult to obtain uniformly valid inference.
- We can still use methods introduced in previous post-selection inference lecture for LASSO types estimators. But most of them require some strong assumptions.
- Here, we introduce a generic method which can be applied to all ML methods.

Best Linear Predictor (BLP) [Chernozhukov et al., 2017]

- An approach focusing on key features of $\tau(X)$ rather than $\tau(X)$ itself, because ML typically does not give consistent estimators of $\tau(X)$ without some form of sparsity assumption in high-dimensional settings.
- If the CATE model is good, the best linear predictor of the true CATE $\tau(X)$ using $\hat{\tau}(X)$ as features should yield a statistically significant coefficient. In fact, in an ideal world this coefficient should be 1.
- This can also serve as omnibus evaluation of the quality of the ML estimates via calibration.

Best Linear Predictor (BLP) [Chernozhukov et al., 2017]

- Formally, the BLP of $\tau(X)$ by $\hat{\tau}(X)$ is the solution to

$$\min_{b_1, b_2} \mathbb{E}[\tau(X) - b_1 - b_2 \hat{\tau}(X)]^2,$$

which, if exists, is defined to (equivalent to)

$$BLP[\tau(X)|\hat{\tau}(X)] := \beta_1 + \beta_2(\hat{\tau}(X) - \mathbb{E}\hat{\tau}(X))$$

where $\beta_1 = \mathbb{E}\tau(X)$ (the ATE), and $\beta_2 = \frac{Cov[\tau(X), \hat{\tau}(X)]}{Var(\tau(X))}$.

- If this test of whether β_2 is nonzero comes up as significant, then this implies that there exists treatment heterogeneity.

Best Linear Predictor (BLP) [Chernozhukov et al., 2017]

- We do not know $\tau(X)$ exactly, how to estimate β_2?
- We use HT-transformed outcome Y^{HT}.
- We have shown that it is an unbiased signal of CATE $\tau(X)$.
- It follows that $BLP[\tau(X)|\hat{\tau}(X)] = BLP[Y^{HT}|\hat{\tau}(X)]$.
- Therefore, we can run the regression of the transformed outcome on $(1, \hat{\tau}(X) - \mathbb{E}_n\hat{\tau}(X))$.
- To reduce nose, we can also add some other controls. Read [Chernozhukov et al., 2017] for more information.
- Note that we also need split samples: using auxiliary data to obtain $\hat{\tau}(X)$ through ML methods. Then, under regularity assumptions, the inference is valid.

Causal Tree

- Hope to estimate: $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$
- Depart from conventional CART:
  - Focus on CATE rather than predicting outcomes
  - For (causal) inference, we require (valid) statistical inference (like confidence intervals and p-values) within estimators. This is achieved through honest estimation.
- We call a tree honest if, for each training example i, it only uses the response Y_i to estimate the within-leaf treatment effect or to decide where to place the splits, but not both.
- Why does it give a valid inference? Once the tree is constructed, by using distinct sample to estimate, standard treatment-effect inference applies exactly as if the partition had been prespecified.
- In addition, honest estimation also avoids overfitting. Why?
- Suppose we split if average outcomes exceed a threshold c (that is $\overline{Y}_L - \overline{Y}_R > c$). Then, given split, the differences $\overline{Y}_L - \overline{Y}_R$ is larger than the population analog.
- Therefore, we do not need to add a penalty term as conventional decision trees.

Causal Tree

- Recall, CART regression tree predicts $\mu(x) = \frac{1}{|\{i:X_i \in L(x)\}|} \sum_{i:X_i \in L(x)} Y_i$
- Here, we focus on causal effects. Therefore, causal tree estimates treatment effect for any $x \in L$

$$\hat{\tau}(x) = \frac{1}{|\{i : Z_i = 1, X_i \in L\}|} \sum_{i:Z_i=1, X_i \in L} Y_i - \frac{1}{|\{i : Z_i = 0, X_i \in L\}|} \sum_{i:Z_i=0, X_i \in L} Y_i$$

- For each tree, the splits of the tree are chosen by maximizing the variance of $\hat{\tau}(X)$.
- code: https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/hte-i-binary-treatment.html

GRF

- Generalized random forest actually is a general framework for using random forests to estimate local parameters defined by moment conditions.

$$\mathbb{E}[m(Z_i; \tau_0(X_i), \eta_0)|X_i = x] = 0$$

- It can be generalized to estimate other quantities besides heterogeneous treatment effects.
- GRF casts forests as a type of adaptive locally weighted estimators that first use a forest to calculate a weighted set of neighbors for each test point x, and then solve a plug-in version of the estimating equation using these neighbors.
- That is, using forest to calculate weights $\alpha_i(x)$, which measures the similarity of x to each other X_i, and the $\hat{\tau}(x)$ is estimated as

$$\sum_{i=1}^{n} \alpha_i(x) m(Z_i; \hat{\tau}(X_i), \hat{\eta}) = 0$$

- For those who have interests, please read the paper [Athey et al., 2019].

Forest-Based Local Estimation

- How to use forest get this adaptive weight?
- We grow a set of B trees indexed by $b = 1, 2, ..., B$.
- For each tree b, define $L_b(x)$ as the set of training examples falling in the same "leaf" as x.
- Define weights $\alpha_i(x)$ which capture the frequency with which the $i-$th training example falls into the same leaf as x:

$$\alpha_{bi}(x) = \frac{1(\{X_i \in L_b(x)\})}{|L_b(x)|}, \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\alpha_{bi}(x)$$

- Note: Within a leaf, each observation gets equal weight, and outside the leaf the weight is 0; and $\sum_{i=1}^{n}\alpha_{bi}(x) = 1$; Note: The weight of the forest $\alpha_i(x)$ is just the average over the trees; $\sum_{i=1}^{n}\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\alpha_{bi}(x) = 1$.
- Recall that for the classical forest, each tree predicts $\hat{\mu}_b(x) = \sum_{i=1}^{n}\frac{Y_i 1(X_i \in L_b(x))}{|L_b(x)|}$, and the final estimate is $\hat{\mu}_b(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b(x)$. α can be obtained by changing the order of summation.

Forest-Based Local Estimation

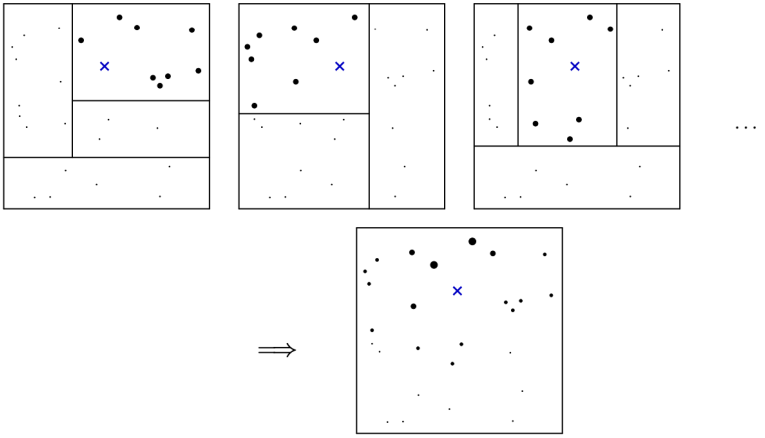


Figure: From [Athey et al., 2019]

GRF based Causal Forest

- GRF based Causal Forest is running a "forest"-localized version of residual-on-residual regression.
- For binary treatment, consider partially linear model

$$Y_i = \tau(X_i)Z_i + g(X_i) + \epsilon_i, \mathbb{E}[\epsilon_i|X_i, Z_i] = 0$$

- Estimate $m(X) = \mathbb{E}[Y|X]$ and $e(X) = \mathbb{E}[Z|X]$, and then calculate the residuals: $\tilde{Y} = Y - \mathbb{E}[Y|X]$ and $\tilde{Z} = Z - \mathbb{E}[Z|X]$
- Estimate $\tau(x)$ by a weighted residual-on-residual regression:

$$\tau(x) := lm(Y_i - \hat{m}^{-i}(X_i) \sim Z_i - \hat{e}^{-i}(X_i), weight = \alpha_i(x))$$

- Note that we can write the moment as $m[(\tilde{Y} - \tau_0(X)\tilde{Z})\tilde{Z}|X = x] = 0$
- Why residual-on-residual regression? How to estimate $m(X)$ and $e(X)$? Can we use ML methods? They are the ideas of Double ML. We will cover it in the next lecture.

GRF

- To produce asymptotically normal and unbiased predictions, except for honest, we also need
  1. balancedness: every split should leave at least a $\rho \geq 0.2$ fraction of the samples on each side.
  2. Random feature split: Every feature should have a probability of at least some probability to be chosen in each split.
  3. Fully grown: the tree should be grown fully such that the number of samples that fall in every leaf should be at most some small constant.
  4. subsampling: unlike typical random forest methods that use bootstrap subsamples to build each tree (i.e. of the same size as the original samples and drawn with replacement), these adapted forests use sub-samples without replacement and of a smaller size.

References

Athey, S., Tibshirani, J., and Wager, S. (2019).
Generalized random forests.
The Annals of Statistics, 47(2).

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2017).
Fisher-schultz lecture: Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in india.
arXiv preprint arXiv:1712.04802.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019).
Metalearners for estimating heterogeneous treatment effects using machine learning.
Proceedings of the national academy of sciences, 116(10):4156–4165.

Tsiatis, A. A. (2006).
Semiparametric theory and missing data.
Springer.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019).
Dynamic treatment regimes: Statistical methods for precision medicine.
Chapman and Hall/CRC.