STA 640 — Causal Inference

Chapter 4.2 Treatment Effect Heterogeneity: Machine Learning Approaches

Fan Li

Department of Statistical Science

Duke University

CATE estimation: basic interaction approaches

- Subgroup analysis: subgroups are pre-specified, static
- The literature has increasingly moved towards identify subgroups with significant effects post-analysis, dynamic
- As discussed earlier, under unconfoundedness

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

- This implies we can simply build an outcome model for $f(z, x) = \mathbb{E}[Y|Z = z, X = x]$
- Once we have estimates of this outcome model, we have estimates of the CATE $\widehat{\tau}(x) = \widehat{f}(1, x) - \widehat{f}(0, x)$

CATE estimation: Basic interaction approaches

- In principle, any outcome regression model (e.g. a simple linear regression) can used to calculate CATE
- The simplest approach is with a linear model

$$f(Z, X) = \beta_0 + \beta_x X + \beta_z Z + \beta_{zx} ZX$$

- Related approaches for other models, such as SVMs (Imai and Ratkovic, 2013)
- Easy to see that $\tau(x) = \beta_z + \beta_{zx} X$
- If we center X, then the ATE is simply

$$E(Y(1) - Y(0)) = \beta_z$$

- Otherwise the ATE is given by

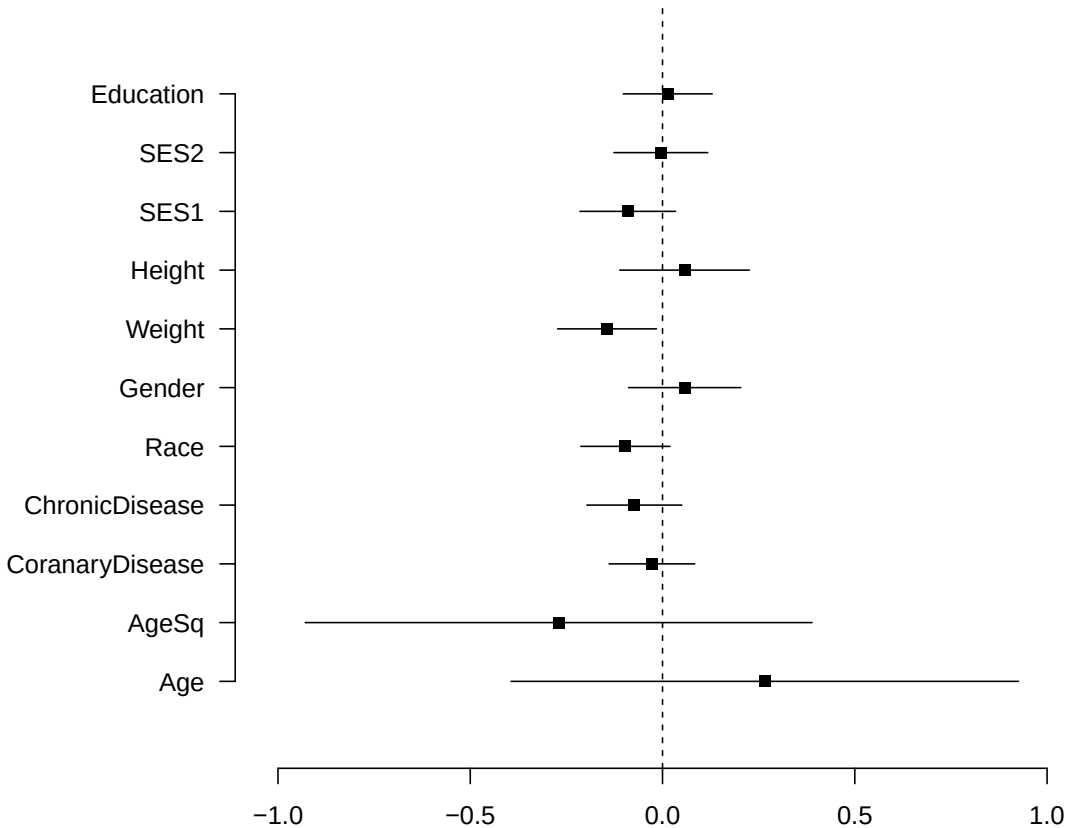
$$E(Y(1) - Y(0)) = \beta_z + \beta_{zx}\, \mathbb{E}(X)$$

CATE estimation: Basic interaction approaches

- Two main reasons why one might like this approach
  - Simple and easy to implement
  - Very interpretable
- A lot of questions are easy to answer in this framework
- Which covariates modify the treatment effect most
  - Examine magnitude of individual β_{zx} values
- Is there any treatment effect heterogeneity?
  - Amounts to testing $H_0 : \beta_{zx} = 0$

CATE estimation: Basic interaction approaches

- Below are estimates of β_{zx} from the NHANES analysis
- Overall ATE is estimated to be -0.08 (-0.19, 0.03)
  - More pronounced, negative effect in individuals with higher weight



CATE estimation: Basic interaction approaches

- A very related approach is to specify separate models in the treated and control groups

$$f(1, X) = \beta_{01} + \beta_{x1} X$$

$$f(0, X) = \beta_{00} + \beta_{x0} X$$

- The CATE is therefore

$$\tau(x) = \beta_{01} - \beta_{00} + (\beta_{x1} - \beta_{x0})x$$

- Treated individuals used to estimate $f(1, X)$ and vice-versa
- In linear models, these two approaches are identical
- Once we jump to nonlinear, flexible approaches these two will behave much differently

Flexible CATE estimators

- There has been a dramatic increase in semiparametric or nonparametric estimators of the CATE that utilize modern statistical learning tools
  - Bayesian nonparametric approaches
  - Machine learning (trees, regularized regressions, boosting, etc.)
- Throughout the rest of the lecture, we will review many of these approaches
  - Discuss pros and cons of each
- Some are left out, but this will cover many of the core ideas

High-dimensional causal analysis and machine learning

- High dimensional settings is common in CATE estimation, but also in ATE estimation
- Two types of high-dimensional settings:
  - A large number of covariates
  - A (propensity or/and outcome) model with a large number of parameters, regardless of the number of covariates, e.g. nonparametric and semiparametric models
- In both cases, machine learning (ML) are often used: dimension reduction, regularized inference
- ML methods are designed for prediction, how about causal, i.e. counterfactual prediction, task? Turns out to be not straightforward

CATE estimation: S-Learners

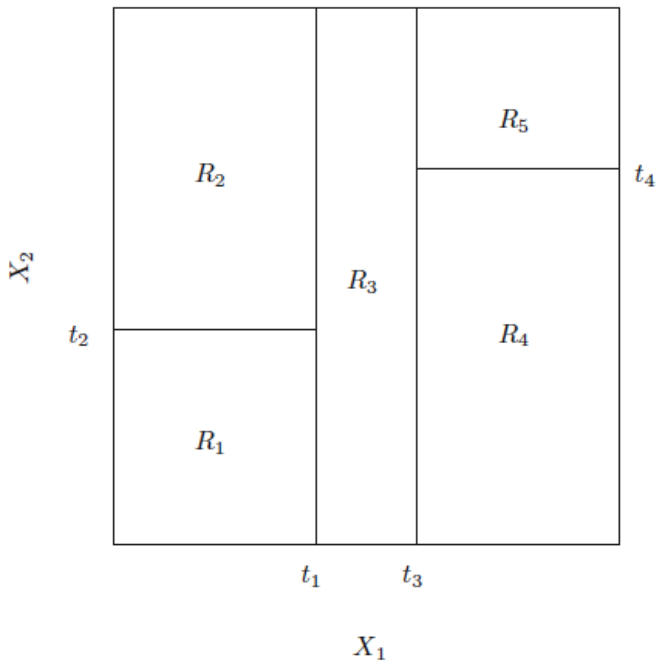
- One class of outcome modeling approach is sometimes referred to as S-learners (S refers to single)
- Exploit the fact that

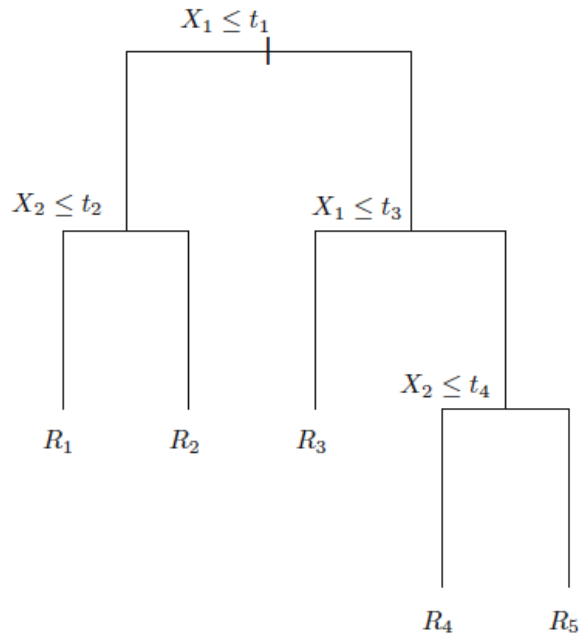
$$\tau(x) = f(1,x) - f(0,x)$$

- Focus solely on flexible estimation of $f(z,x)$
  - CATE estimation is automatic after this
- There are countless machine learning approaches to estimating $f(z,x)$
- One of the seminal papers in this regard is by Jennifer Hill (2011)

Brief review of regression trees

- Regression trees partition the covariate space into non-overlapping regions
- Predictions in each region based solely on data that falls in that region, R_j





James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.

BART approaches to CATE estimation

- Main idea in Hill (2011) is to use BART to estimate $f(z, x)$
- BART assumes that

$$f(z, x) = \sum_{t=1}^{T} g(x, z; \mathcal{T}_t, \mathcal{M}_t)$$

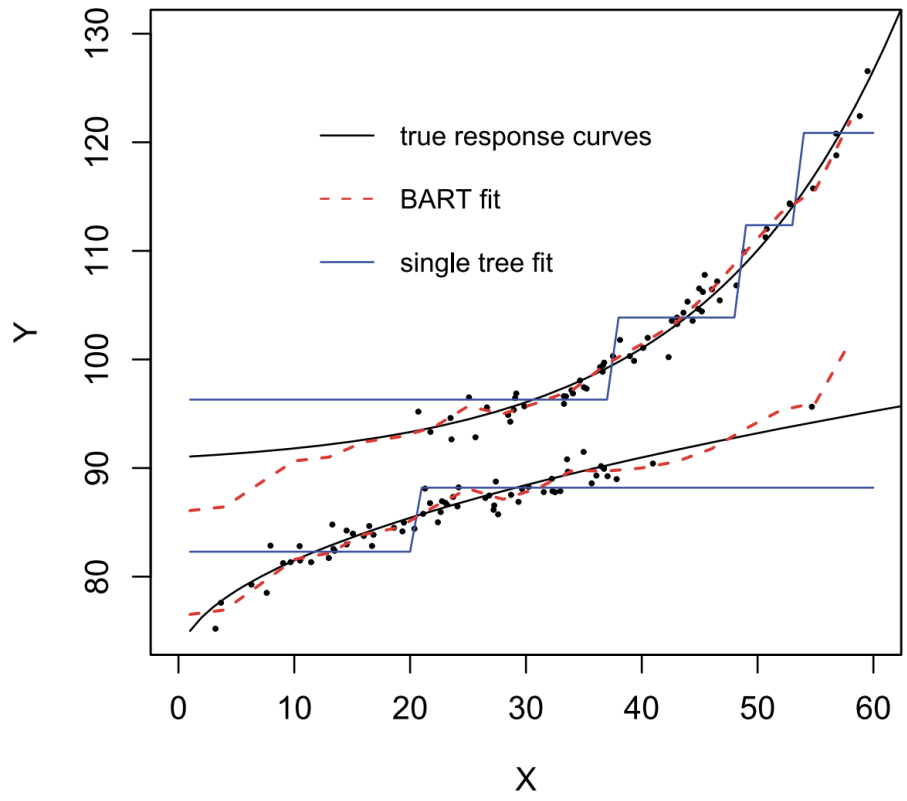
- Here, $g(x, z; \mathcal{T}_t, \mathcal{M}_t)$ is a tree that partitions the space of x and z
  - $\mathcal{T}_t$ represents the tree structure (where splits are)
  - $\mathcal{M}_t$ are parameters for predictions in each terminal node of the tree
- $\mathcal{M}_t = (\mu_{t1}, \ldots \mu_{tL_t})$ where L_t is the number of terminal nodes

BART approaches to CATE estimation

- BART is a Bayesian approach, and certain priors are placed on the parameters of the tree
- The prior probability of splitting decreases with tree depth
  - Probability of splitting at node depth k is $\gamma(1+k)^{-\beta}$ with $\gamma, \beta > 0$
- Shrinkage of mean parameters in each terminal mode are shrunk by a factor of T
  - $\mu_{tl} \sim \mathcal{N}(0, \sigma_u^2/T)$
- My experience is that this greatly outperforms random forests
  - Inference also easy in the Bayesian paradigm
  - Effectively tuning parameter free (defaults work well)
  - For more details, read Chipman et al. (2010)

BART approaches to CATE estimation

- Also much better than using a single regression tree
  - Not surprising given performance of boosting or RFs compared to a single tree



Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." Journal of Computational and Graphical Statistics 20.1 (2011): 217-240.

BART approaches to CATE estimation

- This approach is flexible, automatic, and easy to use
- There are some potential drawbacks
- Putting a BART prior distribution on the response surface $f(z, x)$ has unknown implications for the parameter of interest, $\tau(x)$
- Generally speaking, especially in flexible models, we should be careful about the implications of our prior specification on the parameter of interest
  - Do we expect the CATE to be as complex as $f(z, x)$?

BART approaches to CATE estimation

- These issues were addressed in Hahn et al. (2020) - Bayesian Causal Forest (BCF)
- Main idea is to re-parameterize

$$f(z, x) = \mu(x) + \tau(x)z$$

- Nonparametric extension of the basic interaction approaches we saw earlier
- $\mu(x)$ adjusts for confounding by X
- $\tau(x)$ allows for heterogeneity of the treatment effect
- Separate BART prior distributions placed on these two functions
  - Can use simpler trees for $\tau(x)$

BART approaches to CATE estimation

- The authors further advocate for inclusion of the propensity score

$$f(z, x) = \mu(x, \widehat{e}(x)) + \tau(x)z$$

- This improves our ability to adjust for confounding
- Avoids an issue called regularization induced confounding (bias)
  - Unintended bias that occurs when we are not careful about how we implement regularization or shrinkage in high-dimensional or nonparametric situations
  - Our model might indirectly shrink degree of confounding bias to zero, which is bad when there is severe confounding

Choice of priors

- BART is a special case of the Bayesian nonparametric model. There are others, e.g. Gaussian Process (GP), Dirichlet Process mixture (DPM)
- Which one to choose? Depends on the degree of overlap
- A desirable prior should accurately reflect uncertainty for various degree of overlap
- Simulation evidence:
  - In regions with good overlap, all methods perform similarly
  - In regions with poor overlap, choose a robust prior that adaptively captures uncertainty according to different degree of overlap.
  - With poor overlap, BART appears to struggle whereas GP and DPM do better in reflecting uncertainty

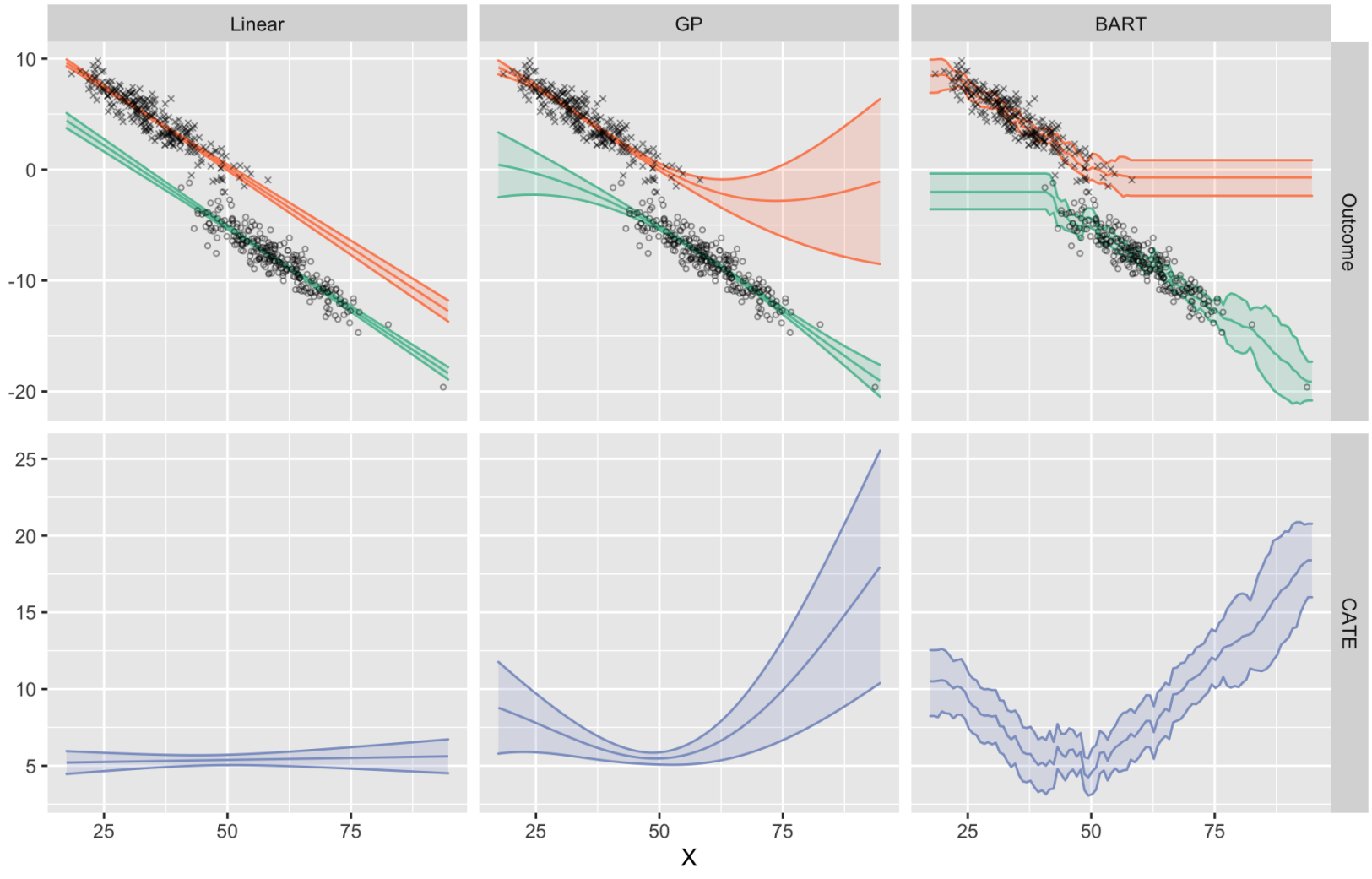
Choice of Priors: a simulated example

- (An example first due to Surya Tokdar, details in Li et al. (2022) review paper)
- A study with 250 treated and 250 control units
- A single covariate X following Gamma distribution with mean 60 in the control and 35 in the treatment group, and with SD 8 in both groups.
- A true outcome model with constant treatment effects:

$$Y_i(z) = 10 + 5z - 0.3X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

- here the CATE $\tau(x) = 5$ for all x. Covariate overlap is good between the groups in the middle of the range of X (around 40 to 50), but deteriorates towards the tails of X.

Choice of Priors: a simulated example



T-learner

- An extension of these ideas that is even more flexible is the T-learner (T refers to “two")
- The previous approach used all of the data to fit one model

$$E(Y \mid Z = z, X = x) = f(z, x)$$

- A T-learner fits separate models to the treated and control groups

$$E(Y \mid Z = 1, X = x) = f_1(x)$$
$$E(Y \mid Z = 0, X = x) = f_0(x)$$

and the CATE is simply

$$\tau(x) = f_1(x) - f_0(x)$$

T-learner

- A couple advantages to this approach
  - Extremely flexible
  - Works well when $f_z(x)$ differs greatly across $z = 0, 1$
- Some drawbacks as well
  - Too flexible! Highly variable
  - Difficult to estimate $f_z(x)$ when treatment group z has few individuals
  - Again no control of $\tau(x)$

T-learner

- Suppose we estimate $f_z(x)$ separately in each group and we have that

$$\text{Var}(\widehat{f_1}(x)) = v_1, \quad \text{Var}(\widehat{f_0}(x)) = v_0$$

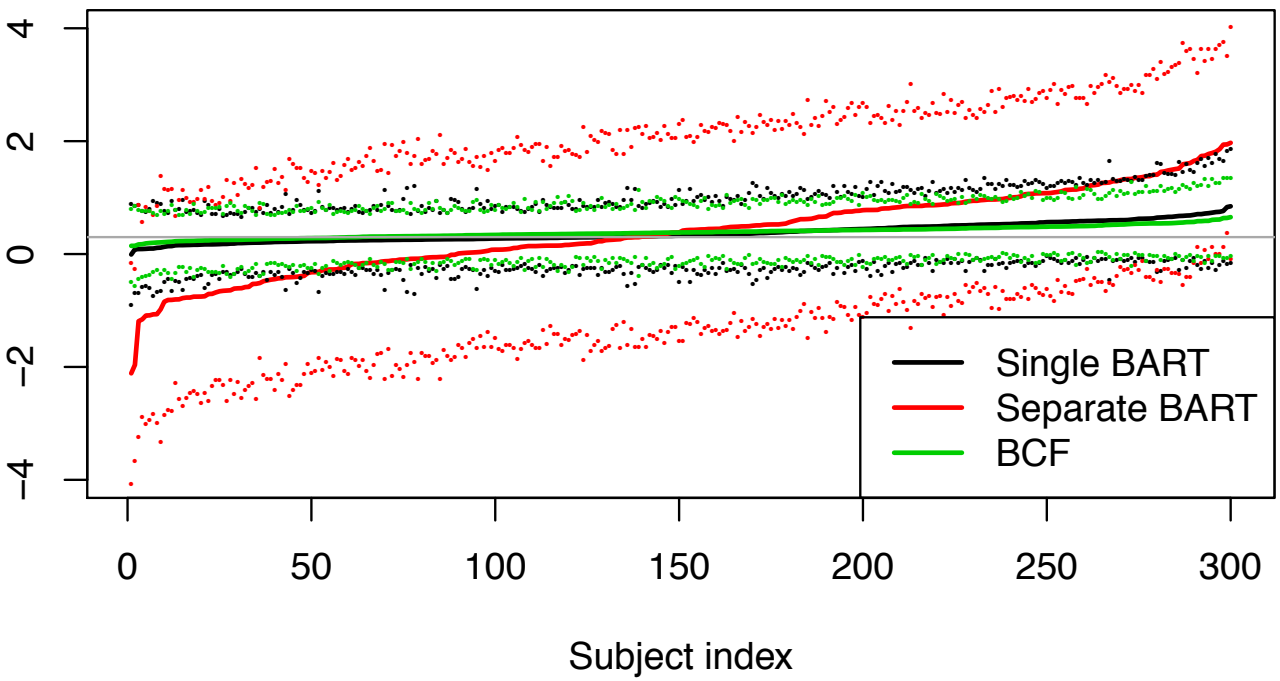
- Due to independence of individuals

$$\text{Var}(\widehat{\tau}(x)) = v_0 + v_1$$

- The variance of the treatment effect is greater than both of the individual functions!
  - Does this coincide with our prior knowledge about the treatment effect function?
  - We generally expect the treatment effect to be as simple, or simpler than $f_z(x)$

T-learner

- Below are estimates and confidence intervals for $\tau(X_i)$ for $i = 1, \ldots n$ in a simulated data set with no heterogeneity
- Separate BART models leads to extremely wide intervals and variable estimates



T-learner

- There are many ways to address this problem
- One way is to impose some structure on $f_z(x)$
  - Put shrinkage directly on $\tau(x)$ as in Hahn et al. (2020)
  - R-learners, which use a specific loss function and a penalty on $\tau(x)$
  - Multi-task learners put shared structure on $f_1(x)$ and $f_0(x)$, e.g. a Gaussian Process (Alaa et al. 2017)
- Another line of approaches constructs pseudo-outcomes and regresses them against X
  - Connections to IPW and DR estimators
- Some approaches directly estimate the CATE
  - Causal forests, related tree-based approaches

Pseudo-outcomes

- The R-learner can be thought of as running a regression on a transformed outcome and covariate
  - Residualized outcome and treatment
  - Allow for coefficient in front of treatment to vary by X_i
- There are a number of other approaches that fit into the scope of transformed outcome regression
- These approaches run a regression on X, but use a special outcome that allows us to estimate $\tau(x)$
- Close connection with IPW and DR estimators from earlier lectures

Pseudo-outcomes

- Remember for estimating the ATE, we had that

$$\mathbb{E}\left[\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right] = \tau^{\text{ATE}}.$$

- This motivated the IPW estimator, which is a sample average of this quantity

$$\hat{\tau}_{ipw} = \frac{1}{N}\left\{\sum_{i=1}^{N}\frac{Y_i Z_i}{e(X_i)} - \sum_{i=1}^{N}\frac{Y_i(1-Z_i)}{1-e(X_i)}\right\}$$

Pseudo-outcomes

- It turns out that

$$\mathbb{E}\left[\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\middle| X = x\right] = \tau(x)$$

- We can define the transformed outcome as

$$O_i = \frac{Y_i Z_i}{e(X_i)} - \frac{Y_i(1-Z_i)}{1-e(X_i)}$$

- Essentially, O_i is an unbiased estimator of an individual's treatment effect
  - Hence $E(O_i) = \tau^{ATE}$ and $E(O_i \mid X_i = x) = \tau(x)$

Pseudo-outcomes

- IPW is not the only choice, we can also use the DR transformed outcome

$$O_i = \left\{ \frac{Z_i Y_i}{e(X_i)} - \frac{Z_i - e(X_i)}{e(X_i)} m_1(X_i) \right\} - \left\{ \frac{(1 - Z_i) Y_i}{1 - e(X_i)} + \frac{Z_i - e(X_i)}{1 - e(X_i)} m_0(X_i) \right\}$$

- If either $e(X_i) = P(Z = 1 | X = X_i)$ or $m_z(X_i) = E(Y | Z = z, X = X_i)$ then

$$E(O_i | X = x) = \tau(x)$$

- This gives us doubly robust estimators of the CATE!
  - Assuming that we correctly specify the $\tau(\cdot)$ function as well

Pseudo-outcomes

- As with the R-learner, this lends itself to a two-stage estimation strategy
- First we need to construct estimates of the PS and outcome regression functions
  - Same as for ATE estimation
  - Use these to create O_i
- Then, once we have O_i, we can run a standard regression of O_i against X_i
  - Using any technique you want!
  - Inference is easier if you use a parametric model here

Pseudo-outcomes

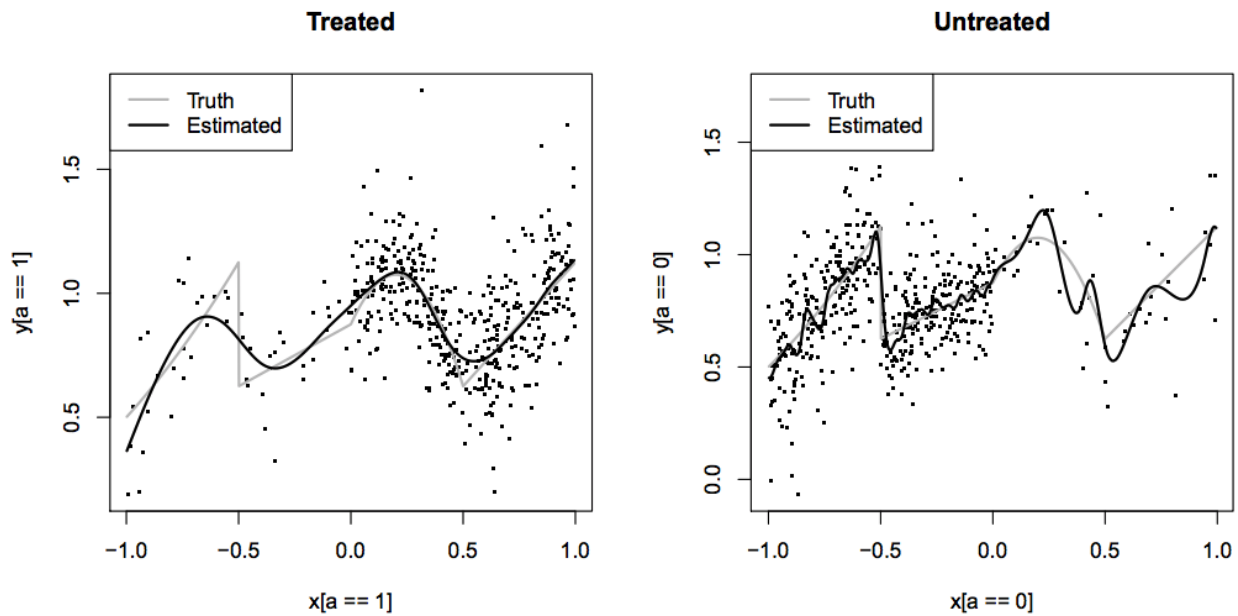
- Many nice features of this approach
- Solves the problem of the T-learner that the CATE is too complicated
  - Second stage estimates can be as simple as you want, regardless of the complexity of $e(X)$ or $m_z(X)$
- Doubly robust version allows us to estimate $\tau(x)$ at a faster convergence rate than either of $e(X)$ or $m_z(X)$ if they're both correctly specified (later)
- We can estimate a wide variety of estimands, not just $\tau(x)$

Pseudo-outcomes

- Not always interested in $E(Y(1) - Y(0)|X = x)$
- What if we only care about heterogeneity by a particular covariate, X_j?
  - Very interpretable estimand that is relevant in many studies
- The pseudo-outcome framework accommodates this easily
  - The second stage regression can simply be a univariate one
- More generally we can learn heterogeneity by V instead of X, where V need not be a subset of X
  - Typically $V \subset X$

Pseudo-outcomes

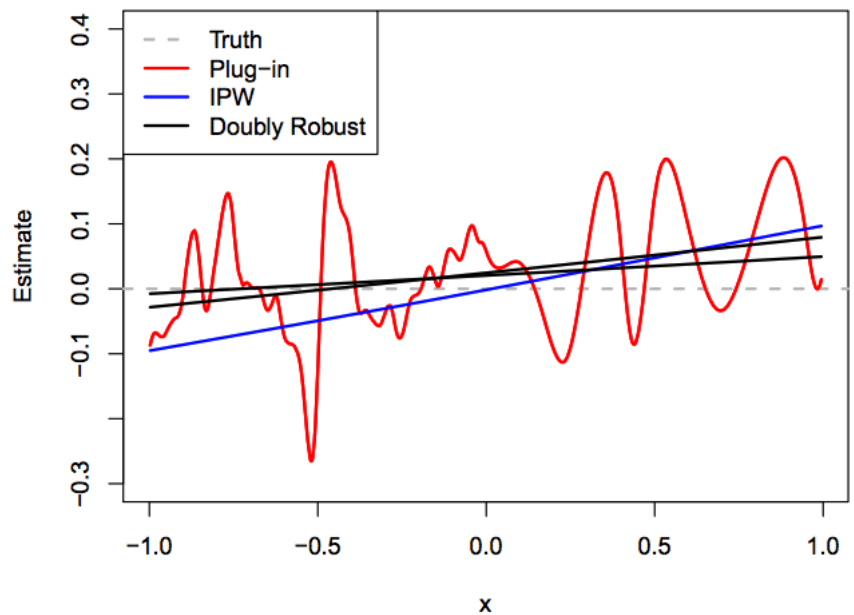
- Here is an illustration of where the individual regression functions are complex, but the difference between them is zero
- T-learner would simply take the difference between these two estimates



Kennedy, EH. "Towards optimal doubly robust estimation of heterogeneous causal effects." (2023)

Pseudo-outcomes

- Now here are the estimates of the CATE using the T-learner (Plug-in) and some pseudo-outcome approaches
- Pseudo-outcome approaches adapt to the simplicity of the problem much better



Kennedy, EH. "Towards optimal doubly robust estimation of heterogeneous causal effects." (2023)

Pseudo-outcomes

- Strong theoretical support
- If we define the oracle risk as

$$R^*(x) = \mathbb{E}\left[\{\widetilde{\tau}(x) - \tau(x)\}^2\right]$$

with $\widetilde{\tau}(x)$ comes from regressing the true $Y_i(1) - Y_i(0)$ against X_i

- Then under certain conditions, the pseudo-outcome with the DR construction satisfies

$$\mathbb{E}\left[\{\widehat{\tau}(x)-\tau(x)\}^2\right] \leq R^*(x)+\mathbb{E}\left[\{\widehat{e}(x)-e(x)\}^2\right] \sum_{z=0}^{1} \mathbb{E}\left[\{\widehat{m}_z(x)-m_z(x)\}^2\right.$$

- However, transformed outcome inherits the extreme weights problem of IPW and the empirical performance is often sensitive to lack of overlap and inferior to other methods

Tree-based approaches

- Tree-based approaches are popular
- The overarching goal of these approaches is to find subsets of the data where the treatment effect varies the most
- No need to specify functional form for $\tau(x)$
  - Assumed constant within areas of covariate space
- Key papers in this area are Wager and Athey (2018), Athey et al. (2019), and Powers et al. (2018)

A quick remark on regression trees

- Before discussing causal trees, we need to discuss one aspect of regression trees
- How do we determine the tree structure?
  - Which covariates to split on?
  - What value of a covariate do we split at?
- In regression trees, we pick splits that reduce the MSE the most among all possible splits
  - Or Gini index / classification error for categorical outcomes
- Greedy algorithm that successively creates splits that improve the model the most

A quick remark on regression trees

- Suppose I'm at the top of a tree and haven't split yet
- My current prediction is $\widehat{Y}_i = \overline{Y}$ for all i
- Now we find the values of j and s that minimize

$$\sum_{i:X_i \in R_1(j,s)} (Y_i - \widehat{Y}_{R_1})^2 + \sum_{i:X_i \in R_2(j,s)} (Y_i - \widehat{Y}_{R_2})^2$$

where

$$R_1(j,s) = \{X | X_j < s\} \quad R_2(j,s) = \{X | X_j \geq s\}$$

- And predictions in these regions are just sample averages (within group)

$$\widehat{Y}_{R_1} = \frac{\sum_{i=1}^n 1(X_i \in R_1(j,s)) Y_i}{\sum_{i=1}^n 1(X_i \in R_1(j,s))} \quad \widehat{Y}_{R_2} = \frac{\sum_{i=1}^n 1(X_i \in R_2(j,s)) Y_i}{\sum_{i=1}^n 1(X_i \in R_2(j,s))}$$

Causal trees and forests

- Causal trees are constructed in a similar way
- Key difference: instead of splitting to reduce MSE the most, we split to maximize heterogeneity of the treatment effect (later)
  - This will lead us to finding areas of the covariate space with different treatment effects
- Another difference: in the terminal nodes
  - In regression trees, the estimates are sample averages of the outcome
  - In causal trees, the estimates are the treatment effects
- There are multiple causal tree algorithms, but we will mostly focus on the original one from Wager and Athey (2018)

Causal trees: estimate treatment effects in nodes

- Suppose we have a tree with terminal nodes or leaves given by $L_1(x), \ldots, L_K(X)$
- In leaf k, we can estimate the treatment effect as

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

- The hope is that within leaf k, individuals have similar covariate values and therefore the treatment is as if randomized
  - And therefore the difference in means estimator is unbiased

Causal forests: where to split?

- Now suppose that we're considering a split of a parent node into two separate nodes
- The estimated treatment effects in each new node are given by $\widehat{\tau}_l$ and $\widehat{\tau}_r$
- One approach to finding splits is to calculate heterogeneity of the treatment effect:

$$\frac{|\widehat{\tau}_l - \widehat{\tau}_r|}{\sqrt{\widehat{\text{Var}}(\widehat{\tau}_l) + \widehat{\text{Var}}(\widehat{\tau}_r)}}$$

  and choose the split that maximizes this
- Other approaches explored in Athey and Imbens (2016)

Causal trees: covariate adjustment

- This will perform well for estimating treatment effects if treatment is unconfounded within leaves
- As suggested in Powers et al. (2018), you can perform additional adjustment
  - Propensity score stratification within leaves
  - Other approaches to confounding adjustment certainly possible
  - Requires larger amount of data within leaves
- Can also incorporate propensity scores into choice of splits
  - Ensures individuals in same leaf have similar PS values
  - No longer maximizes heterogeneity of treatment effect

Causal forests: inference

- Inference in random forests models is typically very hard!
- Wager and Athey (2018) show how inference can be performed for random forests and causal forests
- Sample splitting is used such that
  1. Part of the data is used to find splits, i.e. tree structure
  2. Other part of the data is used to estimate treatment effects within leaves
- They show this leads to asymptotic normality of results with variance estimated by the infinitesimal jacknife (Wager et al. 2014)
- Implemented in the R package `grf`

Causal forests

- Throughout we've discussed creating splits for a single tree, but generally this is repeated a large number of times and results are averaged over all trees (as in random forests)
- We described the simplest type of causal forest
- Many extensions have been proposed that might perform better empirically
  - See Athey et al. (2019) for some ideas
- See also Powers et al. (2018) for other related algorithms such as boosting and MARS that are based on similar ideas

From causal trees to forests

- The most recent version of these causal forests (that I'm aware of) involves combining causal forests with the R-learner from earlier
- Can create a pseudo-outcome as in the R-learner (also using regression trees to estimate $e(X)$ and $m(X)$)
- As in the R-learner, minimize

$$\tau(\cdot) = \underset{\tau}{\operatorname{argmin}} \left\{ \widehat{L}_n(\tau(\cdot)) + \Lambda_n(\tau(\cdot)) \right\}$$

- And now, the splits of the tree can be chosen to minimize this quantity

Tune a tree model for causal inference

- In ML models, a crucial step is to use cross-validation to tune hyperparameters: split the data into training (build model) and testing data (check model)
- In prediction problems, the standard performance metric is prediction MSE
- Similarly for a estimator of a causal estimand, say a CATE estimator $\hat{\tau}(x)$, we may use a MSE:

$$L(\hat{\tau}) = E[(Y_i(1) - Y_i(0) - \hat{\tau}(X_i))^2].$$

- But wait... this is usually not possible in causal inference problems, because even in the test data we do not know the true causal effect (Rolling and Yang, 2014; Athey and Imbens, 2016)
- So we would need approximations to the truth

Honesty Criterion

Athey and Imbens, 2016; Athey, Tibshirani, Wager 2018

- Honesty criterion: a sample can only be used to estimate τ or decide how to build the model (e.g. where to place the splits in trees), but not both.
- Intuition: avoid using data twice
- Implementation: the study sample is divided into three subsamples: two for training (one for building the tree and one for estimating causal effects) and one for testing
- Wager and Athey (2018) devised two tree-based *honest* procedure to estimate CATE: (i) double-sample (outcome) tree, and (ii) propensity tree (discuss later)
- Honesty is important to achieve asymptotic normality and unbiasedness.

Double/debiased machine learning

- A recurring idea in the previous methods is "double" learning: using ML for both outcome and propensity model, and combine
- A general theoretical framework is double/debiased ML by Chernozhukov et al.
- Recall: ML methods are effective for prediction, how about causal (i.e. counterfactual prediction) task?
- Good prediction performance of ML models does not automatically translate into good performance for estimation of "causal" parameters
  - Regularization bias: slower convergence rate
  - Overfitting bias: Capturing more than the relationship between Y and Z

An Earlier Example: Double Selection for Causal

- An earlier method is to use ML methods in double-robust (DR) estimators for ATE
- Main idea: specify ML models for both propensity score and outcome models (Farrell, 2015)
- With high-dimensional confounders, Belloni et al. (2014) proposes a double-selection procedure
  - Select confounders/covariates for the propensity score model and for the outcome model, e.g., by LASSO
  - Use least square estimation of the outcome with treatment indicator plus the union of selected confounders
- "Double-selection" gives $\sqrt{N}$ consistency of ATE, whereas "single-selection" cannot reach

Frisch-Waugh-Lovell (FWL) Theorem

- Consider a linear regression

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + U,$$

  with $E(U|Z, X) = 0$

- To obtain β_1, we can use OLS by concatenating Z, X
- Frisch-Waugh-Lovell theorem gives another consistent way to estimate β_1:
  1. Regress (linear) Y on X, obtain residual $\hat{U} = Y - \hat{Y}$
  2. Regress (linear) Z on X, obtain residual $\hat{V} = Z - \hat{Z}$
  3. Regress (linear) $\hat{U}$ on $\hat{V}$, obtain $\hat{\beta}_1$
- The proof is a classic in linear models textbooks, e.g. you can find it **here (click)**

Robinson decomposition

- Robinson (1988) generalized FWL theorem: replace the linear regressions in FWL to some nonparametric (e.g. kernel) regression
- Robinson's procedure:
  1. Kernel regression of Y on X, obtain residual $\hat{U} = Y - \hat{Y}$
  2. Kernel regression of Z on X, obtain residual $\hat{V} = Z - \hat{Z}$
  3. Regress (linear) $\hat{U}$ on $\hat{V}$, obtain $\hat{\beta}_1$
- Double machine learning (DML) further generalizes FWL and Robinson ideas to machine learning models

Canonical Example of DML: Partial Linear Model

Chernozhukov et al. 2018

- Intuition: The relationship between Y and X is usually more complex than the relationship between Y and T (echoing Bayesian causal forest parametrization)
- Idea: Use a ML model for $Y \sim X$ and a linear model for $Y \sim Z$
- Consider a causal partial linear model:

$$Y = Z\tau + m(X) + U$$

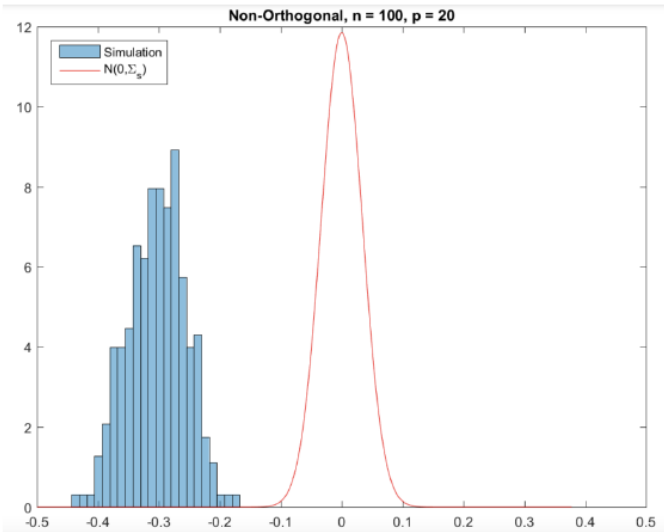
$$Z = e(X) + V$$

with $\mathbb{E}(U|X, Z) = 0$ and $\mathbb{E}(V|X) = 0$

  - Z: treatment
  - X: a high-dimensional vector of covariates/confounders $\mathbb{E}(U|X, Z) = 0$
  - τ: the causal estimand/parameter ATE; later τ can be extended to CATE $\tau(X)$

Näive prediction-based ML approach is Bad

- Predict Y using X and Z: $\hat{Y} = Z\hat{\tau} + \hat{m}(X)$
- For example, we can estimate τ and g iteratively:
  - Given initial parameters, run a ML model, e.g. random forest or boosting on $\hat{Y} - Z\hat{\tau}$ to fit $\hat{m}(X)$
  - Run OLS on $\hat{Y} - \hat{m}(X)$ to fit $\hat{\tau}$
  - Repeat until converge
- Good prediction performance of Y: small $(\hat{Y} - Y)^2$, but dist of the estimated causal parameter, $\hat{\tau} - \tau$, looks like this



Source of bias and solutions

- We can look at the asymptotic distribution of $\hat{\tau} - \tau$ (a clear derivation is here (click)
- Two sources of bias
  - Regularization bias: Machine learning methods employ regularization (e.g., L1 or L2 regularization) to reduce variance, but this often induces bias and slower convergence rates
    - How to solve? Double ML - using ML twice: once to learn Y on X, and once to learn Z on X, and then regress residual on residual – the FWL/Robinson style:
  - Overfitting bias
    - How to solve? Sample-splitting and cross-fitting.
- The key theory were developed in a series of papers by Chernouzhukov and co-authors, starting from Chernouzhukov et al. (2017, 2018)

DML Algorithm - Summary

In summary, for a given dataset $\{X_i, Z_i, Y_i\}_{i=1}^N$, DML follows this algorithm to estimate average treatment effect:

1. *Split sample*: random partition the data into k mutually exclusive parts: $\{I_k\}_{k=1}^K$. For each k, define $I_k^c = \{1, \ldots, N\} \backslash I_k$
2. *Estimate propensity model in training sample:* Train any (regularized) ML model M_z to predict Z from X (propensity) using auxiliary I_k^c
3. *Estimate outcome model in training sample:* Train any (regularized) ML model M_y to predict Y from X (outcome) using auxiliary I_k^c
4. *Estimate ATE in prediction sample:* obtain the residuals in I_k: $Z_R = Z - M_z(X)$, and $Y_R = Y - M_y(X)$, and regress (linearly) Y_R on Z_R to estimate ATE
5. *Aggregate over K folds:* Repeat 2-3 for $k = 2, \ldots, K$ so that DML uses the full data

DML for CATE

- Extend DML to CATE with a generalized partial linear model:

$$Y = Z\tau(X) + m(X) + U, \quad \mathbb{E}(U|X, Z) = 0$$
$$Z = e(X) + V, \quad \mathbb{E}(V|X) = 0$$

  with $\mathbb{E}(UV|X, Z) = 0$, where $\tau(X)$ is the CATE

- Same idea: regress residualized outcome $\hat{U}$ and treatment $\hat{V}$:

$$\hat{\tau}(X) = \text{argmin}_{\tau \in} E_n[(\hat{U} - \tau(X)\hat{V})^2]$$

- Difference choices of $\tau(X)$ in DML for CATE:
  - Reproducing Kernel Hilbert Space (Nie and Wager, 2021)
  - Random forest (Athey et al. 2019)
  - Sparse linear space (Chernozhukov et al.)
- A rich online source and package of DML is **here (click)**

R-learners

- R-learners use a clever parameterization of the problem to directly estimate and regularize the CATE
- Assuming a generalized partial linear model

$$Y_i = \mu(X_i) + \tau(X_i)Z_i + \epsilon_i$$

and if we take the conditional expectation of this, we obtain

$$m(X_i) = E(Y_i \mid X_i) = \mu(X_i) + \tau(X_i)e(X_i)$$

R-learners

- As first pointed out in Robinson (1988), these imply that

$$Y_i - m(X_i) = (Z_i - e(X_i))\tau(X_i) + \epsilon_i$$

- Which further implies that

$$\tau(\cdot) = \underset{\tau}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\left((Y_i - m(X_i)) - (Z_i - e(X_i))\tau(X_i) \right)^2 \right] \right\}$$

- Nie and Wager (2021) build on these ideas to estimate heterogeneous treatment effects

R-learners

- Their main idea is to estimate the CATE in the following way:

$$\tau(\cdot) = \underset{\tau}{\operatorname{argmin}} \left\{ \widehat{L}_n(\tau(\cdot)) + \Lambda_n(\tau(\cdot)) \right\}$$

where

$$\widehat{L}_n(\tau(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \left(\left(Y_i - \widehat{m}^{-i}(X_i) \right) - \left(Z_i - \widehat{e}^{-i}(X_i) \right) \tau(X_i) \right)^2$$

- $\Lambda_n(\tau(\cdot))$ is a penalty on the complexity of the CATE
  - Many options such as smoothness penalties, lasso, etc.

R-learners

- Note that we used $\widehat{m}^{-i}(X_i)$ and $\widehat{e}^{-i}(X_i)$ in the squared error loss
- These are estimates of the conditional mean outcome regression and propensity score with the i^{th} observation removed
  - Typically done using 5 or 10-fold cross validation, not leave one out
- This approach separated the problem into two separate stages
  - Estimating nuisance functions, $m(\cdot)$ and $e(\cdot)$
  - Estimation of $\tau(\cdot)$ conditional on nuisance function estimates
- Allows for separate penalization in these two steps
  - Allows for the CATE to be much simpler than the outcome regression functions

R-learners

- This approach directly addressed the problems of the T-learner
- They show this approach can be used with many modern machine learning type of estimators for the CATE
  - High-dimensional models
  - Gradient boosting
  - Neural networks
- Key idea: regress residuals of outcome on residuals of treatment – a special case of double/debiased machine learning

Another double approach: TMLE

- Targeted Maximum Likelihood Estimation or Targeted Minimum Loss Estimation (TMLE) (van der Laan and Rubin, 2006, and series of following work)

  1. Obtain a preliminary estimate of $\{\hat{m}_z^{(0)}(X)\}$ of the outcome $E\{Y(z)|X\}$ based on a ML algorithm (e.g. an ensemble learner), and fit a parametric (or ML) PS model to estimate PS $\hat{e}(X)$
  2. Fit a canonical generalized linear model for $E\{Y(z)|X\}$, with link function $h(\cdot)$, offset term $h\{\hat{m}^{(0)}(X)\}$, and the single covariate – IP weights: $Z_i/\hat{e}(X)$

- TMLE uses (inverse of) PS as the additional covariate: recall discussion earlier on regression with the clever covariate

Another double approach: TMLE

- The logistic model in step (3) is called a fluctuation working model
- Without the fluctuation model, the algorithm is simply an OR estimator based on $N^{-1} \sum_{i=1}^{N} \hat{m}^{(0)}(X_i)$
- TMLE uses (inverse of) PS to fluctuate the initial regression
  - can show that the score of the stabilized fluctuation model at zero fluctuation ($\hat{\epsilon}_n = 0$) spans the doubly robust estimating function (recall discussion earlier on regression with the clever covariate)
- This is a fully iterated DR estimator

Machine Learning and Causal Inference: Key Insights

- Machine learning greatly expands the toolbox for outcome modeling
- But machine learning does not magically solve the fundamental problem of causal inference
- The key issues in causal inference — overlap, balance, unconfoundedness — remain the same and requires more care
- To adapt machine learning methods to causal inference, one has to adapt to those key issues.
- Key insights:
  - Sample splitting: for building model and for estimating effects
  - Double learning: combine both PS model and outcome model for causal inference with high-dimensional data
  - Flexible (outcome) modeling

References

- Imai, Kosuke, and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation." The Annals of Applied Statistics 7.1 (2013): 443-470.
- Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." Journal of Computational and Graphical Statistics 20.1 (2011): 217-240.
- Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." The Annals of Applied Statistics 4.1 (2010): 266-298.
- Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)." Bayesian Analysis 15.3 (2020): 965-1056.
- Nie, Xinkun, and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects." Biometrika 108.2 (2021): 299-319.
- Kennedy, Edward H. "Towards optimal doubly robust estimation of heterogeneous causal effects." Electronic J Statistics 17(2): 3008-3049 (2023).
- Athey, Susan, and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113.27 (2016): 7353-7360.

References

- Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association 113.523 (2018): 1228-1242.
- Athey, Susan, Julie Tibshirani, and Stefan Wager. "Generalized random forests." The Annals of Statistics 47.2 (2019): 1148-1178.
- Wager, Stefan, Trevor Hastie, and Bradley Efron. "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife." The Journal of Machine Learning Research 15.1 (2014): 1625-1651.
- Powers, Scott, et al. "Some methods for heterogeneous treatment effect estimation in high dimensions." Statistics in medicine 37.11 (2018): 1767-1787.
- Rolling, C. A., Yang, Y. (2014). Model selection for estimating treatment effects. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 749-769.

References

- Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics, 189(1), 1-23.
- Belloni, A., Chernozhukov, V., Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2), 608-650.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. Econometrica, 85(1), 233-298.
- Robinson, P. M. Root-n-consistent semiparametric regression. Econometrica, 56(4):931-954, 1988.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. American Economic Review, 107(5), 261-65.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21(1), C1-C68.
- van der Laan, M. J., Rubin, D. (2006). Targeted maximum likelihood learning. The international journal of biostatistics, 2(1).