Recursive partitioning for heterogeneous causal effects

Jasmine. Hao[1]

[1]University of Hong Kong

ECON 6083: Machine Learning

Outline

Challenges in Policy Evaluation

Gold Standard: Randomized controlled experiments.
Limitations: Difficult to implement due to *financial*, *political*, or *ethical reasons*, or *small populations*.
Example: Unethical to prevent students from attending college to study labor market effects.
Observational Data: Many studies rely on *non-randomly assigned* policies.
Inference Challenges: Difficult to draw causal effects from *observational data*.

Strategies for Causal Inference

Identification Strategies: Various methods to draw causal inferences from *observational data.*
Key Methods: Regression discontinuity, synthetic control, differences-in-differences, network settings, and hybrid approaches.
Supplementary Analyses: Increasingly important to validate primary analyses.
Machine Learning: Promising approach for enhancing policy evaluation credibility.

New Developments in Program Evaluation

Potential Outcome Approach:

- Gained acceptance since the early 1990s as a framework for analyzing causal problems (Rubin Causal Model).
- Focuses on *potential outcomes* for different treatment levels, highlighting the *fundamental problem of causal inference* (Holland, 1986).

Unconfoundedness Assumption:

- Assumes all confounders are observed, allowing treatment to mimic random assignment (Rosenbaum and Rubin, 1983a).
- Methods like matching, reweighting, and propensity scores help estimate treatment effects (Imbens, 2004; Abadie and Imbens, 2006; Imbens and Rubin, 2015; Heckman and Vytlacil, 2007).

Complementary Approaches:

- Graphical models approach (Pearl, 2000) is widely used in other disciplines for causal inference.

Machine Learning Methods for Average Causal Effects: Overview

Background:

- Machine learning methods control for a large number of covariates flexibly (Hirano et al., 2001; McCaffrey et al., 2004; Wyss et al., 2014).
- These methods adapt to emphasize covariates important for reducing bias correlated with both outcomes and treatment indicators.

Double Selection Procedure:

- Proposed by Belloni et al. (2013), involves LASSO regression to select covariates correlated with outcome and treatment.
- Improves estimator properties for average treatment effect by combining selected covariates in a final regression.

Balancing and Adjustment for Covariates

Direct Balancing of Covariates:

- Focuses on weighting to balance covariates between treatment and control groups (Hainmueller, 2012; Graham et al., 2012, 2016; Zubizarreta, 2015; Imai and Ratkovic, 2014).
- Athey, Imbens, and Wager (2016) develop an estimator combining balancing with regression adjustment to predict counterfactual outcomes, reducing extrapolation needs.

Semiparametric Influence Functions:

- Approach builds on semiparametric literature (van der Vaart, 2000; Robins and Rotnitzky, 1995; van der Laan and Rubin, 2006).
- Chernozhukov et al. (2016) suggest using machine learning for nonparametric components and sample-splitting for enhanced properties.

Practical Considerations in Data Trimming

Data Trimming Practices:

- Procedures for trimming data to eliminate extreme values of estimated propensity scores are crucial (Crump et al., 2009).
- Helps to refine the model by removing outliers that could skew the estimation of causal effects.

Understanding Heterogeneous Causal Effects

Context:

- Policies or treatments may have varying costs and benefits across different settings.
- Insight into heterogeneous treatment effects is crucial for determining where the benefit/cost ratios are most favorable.

Challenges with Machine Learning Methods:

- Machine learning methods explore many covariates and subsets, which can lead to spurious findings of differences in treatment effect.
- Similar to clinical trials, where pre-analysis plans are required to prevent spurious discoveries from data dredging.

Addressing Multiple Hypothesis Testing in Heterogeneity Analysis

Exhaustive Search and Corrections:

- Researchers may exhaustively search for treatment effect heterogeneity and correct for multiple hypothesis testing issues.
- This involves considering a large number of hypotheses and adjusting for the possibility of false discoveries (List, Shaikh, and Xu, 2016).

Proposed Method by List et al.:

- Assign each covariate a "low" or "high" value and test treatment effects across these values.
- Use bootstrapping to account for correlation among test statistics, improving detection of true heterogeneity when covariates are correlated.
- Requires specification of all hypotheses and discretization of covariates in advance, limiting flexibility in exploration.

Limitations and Practical Considerations

Limitations of Current Approaches:

- Standard multiple testing corrections may reduce the power of tests to detect true heterogeneity due to the large number of covariates.
- List et al.'s method, while improving over standard approaches, still requires pre-defined hypotheses and discretization strategies, which may not fully explore potential heterogeneities.

Need for Flexible Methods:

- There's a growing need for methods that allow more flexible interaction among covariates without predefined hypotheses.
- Such flexibility could enhance the ability to uncover and understand nuanced variations in treatment effects across different groups.

Estimating Heterogeneity in Causal Effects (Athey and Imbens, 2016)

- **Objective:** Estimate heterogeneity in causal effects in studies.
- **Approach:** Data-driven partitioning based on treatment effect magnitude .
- **Confidence Intervals:** Construct valid confidence intervals without "sparsity" assumptions.
- **"Honest" Estimation:** Use separate samples for partition construction and effect estimation.
- **Methodology:** Modified regression tree methods for fitting and honesty of the treatment effect.
- **Challenge:** Address missing "ground truth" in cross-validation.
- **Results:** 90% confidence interval coverage for honest estimation; 7-22% mean squared error reduction.

Model Setup

- **Units:** N units, indexed by $i = 1, \ldots, N$.
- **Potential Outcomes:** Pair $(\mathbf{Y}_i(0), \mathbf{Y}_i(1))$ for each unit.
- **Causal Effect:** Unit-level effect defined as $\tau_i = \mathbf{Y}_i(1) - \mathbf{Y}_i(0)$.
- **Treatment Indicator:** Binary variable $\mathbf{W}_i \in \{0, 1\}$.
  - $\mathbf{W}_i = 0$: Control treatment.
  - $\mathbf{W}_i = 1$: Active treatment.
- **Observed Outcome:** $\mathbf{Y}_{i,\text{obs}} = \mathbf{Y}_i(\mathbf{W}_i)$.
- **Features:** K-component vector $\mathbf{X}_i$ of pre-treatment variables.

Data

- **Data:** Triple $(\mathbf{Y}_{i,\text{obs}}, \mathbf{W}_i, \mathbf{X}_i)$ as an i.i.d. sample.
- **Assumptions:**
  - Observations are exchangeable.
  - No interference (stable unit treatment value assumption).
- **Marginal Treatment Probability:** $p = \Pr(\mathbf{W}_i = 1)$.
- **Propensity Score:** $e(\mathbf{x}) = \Pr(\mathbf{W}_i = 1 | \mathbf{X}_i = \mathbf{x})$.

Assumptions and Implications: Unconfoundedness I

- **Unconfoundedness Assumption:** We maintain the assumption of randomization conditional on the covariates, formalized as follows:

  Assumption 1 (Unconfoundedness): $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i$

  - This assumption denotes the (conditional) independence of two random variables. It's met in randomized experiments without covariate conditioning and can be justified in observational studies if all relevant variables are observed.

Assumptions and Implications: Unconfoundedness II

- **Complete Randomization:** For simplicity, we maintain the stronger assumption:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1), X_i)$$

- **Conditional Average Treatment Effect:** $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$.
- **Focus:** Our main focus is on obtaining accurate estimates of and inferences for $\tau(x)$. We aim to find estimators $\hat{\tau}(\cdot)$ that are based on partitioning the feature space and do not vary within the partitions.

1

[1] Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41-55.

Conditional Average Treatment Effects and Partitioning

- **Conditional Average Treatment Effect:** $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$.
- **Population Average Treatment Effect:** Significant portion of causal inference literature focuses on estimating $E[Y_i(1) - Y_i(0)]$.
- **Main Focus:** Obtaining accurate estimates of and inferences for the conditional average treatment effect $\tau(x)$.
- **Partitioning Feature Space:** Interested in estimators $\hat{\tau}(\cdot)$ that are based on partitioning the feature space and do not vary within the partitions.

[1] References include: Imbens G, Rubin D (2015), Abadie A, Imbens G (2006), Pearl J (2000), Rosenbaum P (2002).

Honest Inference for Population Averages

- **Departure from CART:** Our approach departs from conventional Classification and Regression Trees (CART) in two fundamental ways:
  1. **Focus on Treatment Effects:**
     - ★ We focus on estimating *conditional average treatment effects* rather than predicting outcomes.
     - ★ Conventional regression tree methods are not directly applicable because causal effects at the unit level are not observed for any unit.
  2. **Honest Estimation:**
     - ★ We separate the tasks of constructing the partition and estimating effects within leaves, using separate samples for each.
     - ★ This is called *honest estimation*.
- **Contrast with Adaptive Estimation:**
  - ▶ Honest estimation contrasts with the *adaptive estimation* used in conventional CART, where the same data are used for both the building of the partition and the estimation of leaf effects.

Key Concepts and Functions

- **Partitioning (Π):**
  - A partitioning of the feature space $\mathbb{X}$.
  - Defined as $\Pi = \{\ell_1, \ldots, \ell_{\#(\Pi)}\}$ with $\cup_{j=1}^{\#(\Pi)} \ell_j = \mathbb{X}$.
- **Space of Partitions (P):**
  - Denoted as P, with $\ell(x; \Pi)$ representing the leaf $\ell \in \Pi$ containing x.
- **Space of Data Samples (S):**
  - Represents samples drawn from a population.
- **Algorithm (π):**
  - Constructs a partition based on a sample $S \in \mathbb{S}$.
  - Example: Splits if the difference in average outcomes exceeds a threshold c.

Simple Algorithm for Splitting

Simple Algorithm for Splitting:

- An algorithm $\pi(S)$ that decides to split based on the difference in average outcomes:

$$\pi(S) = \begin{cases} \{\{L, R\}\} & \text{if } \bar{Y}_L - \bar{Y}_R \leq c, \\ \{\{L\}, \{R\}\} & \text{if } \bar{Y}_L - \bar{Y}_R > c. \end{cases}$$

- $\bar{Y}_L$ and $\bar{Y}_R$ are the average outcomes in the left and right subsamples, respectively.
- c is a predefined threshold for splitting.

Potential Bias in Adaptive Estimation

Potential Bias in Adaptive Estimation:

- $\bar{Y}_L - \bar{Y}_R$ is generally an unbiased estimator for the difference in population conditional means $\mu(L) - \mu(R)$.
- However, conditioning on $\bar{Y}_L - \bar{Y}_R \geq c$ in a sample, we expect $\bar{Y}_L - \bar{Y}_R$ to be larger than the population analog.
- This illustrates the potential bias in leaf estimates from adaptive estimation.

Conditional Mean Function and Estimated Counterpart

Conditional Mean Function:

- Given a partition Π, the conditional mean function $\mu(x;\Pi)$ is defined as:

$$\mu(x;\Pi) \equiv \mathbb{E}[Y_i|X_i \in \ell(x;\Pi)] = \mathbb{E}[\mu(X_i)|X_i \in \ell(x;\Pi)],$$

  which is a step-function approximation to $\mu(x)$.

Estimated Counterpart:

- Given a sample S, the estimated counterpart $\hat{\mu}(x;S,\Pi)$ is:

$$\hat{\mu}(x;S,\Pi) \equiv \frac{1}{\#(i \in S : X_i \in \ell(x;\Pi))} \sum_{i \in S: X_i \in \ell(x;\Pi)} Y_i,$$

  which is unbiased for $\mu(x;\Pi)$.
- We index this estimator by the sample to be precise about which sample is used for estimation of the regression function.

Comparison between Adaptive and Honest Approaches

- **Adaptive Approach:**
  - Uses the training data for model selection.
  - Spurious correlations between covariates and outcomes can lead to biases in the selected model.
  - Biases disappear only slowly as the sample size grows.
  - Additional assumptions like "sparsity" may be needed for consistency or asymptotic normality of predictions.
- **Honest Approach:**
  - Does not use the same information for selecting the model structure as for estimation.
  - Involves splitting the training sample into two parts: one for constructing the model and another for estimating effects.
  - Places no restrictions on model complexity.
  - Asymptotic properties of treatment effect estimates are the same as if the partition had been exogenously given.
  - Loss of precision due to sample splitting is offset by the elimination of bias.

Criterion for Comparison and Adjustment for Prediction

- **Criterion for Comparison:**
  - Focus on Mean Squared Error (MSE) for comparing estimators.
  - Adjustments made as necessary for specific cases.
- **Adjustment for Prediction Case:**
  - MSE adjusted by $E[Y_i^2]$.
  - Does not affect ranking of estimators.

MSE Definition

- For partition Π, MSE is defined as:

$$MSE_\mu(S^{te}, S^{est}, \Pi) \equiv \frac{1}{\#(S^{te})} \sum_{i \in S^{te}} \left[(Y_i - \hat{\mu}(X_i; S^{est}, \Pi))^2 - Y_i^2 \right],$$

where S^{te} is the test sample, and S^{est} is the estimation sample.

Expected MSE (EMSE) and Estimators for EMSE

- **Expected MSE (EMSE):**
  - Expected MSE is the expectation of $MSE_\mu(S^{te}, S^{est}, \Pi)$ over test and estimation samples:

$$\text{EMSE}_\mu(\Pi) \equiv E_{S^{te},S^{est}}\left[MSE_\mu(S^{te}, S^{est}, \Pi)\right],$$

    where the test and estimation samples are independent.
- **Estimators for EMSE:**
  - Various estimators considered for the adjusted EMSE.
  - MSE functions evaluated at units in test sample S^{te}, based on estimation sample S^{est} and tree Π.

Ultimate Goal

- **Ultimate Goal:**
  - Construct and assess algorithms $\pi(\cdot)$ that maximize the honest criterion

$$Q_H(\pi) \equiv -\mathbb{E}_{S^{te},S^{est},S^{tr}}[MSE_\mu(S^{te}, S^{est}, \pi(S^{tr}))].$$

  - Focus on maximizing criterion functions, typically involving the negative of MSE expressions.

Outline

Adaptive Target

- In the conventional CART approach, the target is:

$$Q^C(\pi) \equiv -\mathbb{E}_{S^{te},S^{tr}}[MSE_\mu(S^{te}, S^{tr}, \pi(S^{tr}))].$$

- The same training sample S^{tr} is used for both constructing the tree and estimating the conditional means.
- Note that maximizing criterion functions typically involve the negative of MSE expressions.

Key Difference

- In the honest approach, different samples S^{tr} and S^{est} are used for tree construction and conditional means estimation, unlike the adaptive approach.

Costs and Benefits

- *Cost*: Sample size. Using some data for estimation leaves fewer units for the training dataset, leading to higher expected MSE.
- *Advantage*: Honest estimation avoids the issue of spurious extreme values of Y_i being placed into the same leaf as other extreme values by the algorithm $\pi(\cdot)$.
- The adaptive estimation issue results in poorer coverage properties of confidence intervals compared to the honest methods.

CART Algorithm

- Consists of two parts: initial tree building and cross-validation for complexity parameter selection.
- Both parts are based on the MSE criterion functions.

Tree-Building Phase

- Recursively partitions the training sample using an "in-sample" goodness-of-fit criterion $-MSE_\mu(S^{tr}, S^{tr}, \Pi)$.
- Conventional criterion can lead to overfitting.
- Mitigated by using cross-validation to select a penalty on tree depth.
- Goodness-of-fit criterion will always improve with additional splits but may increase expected MSE with smaller leaf sizes.

Penalty Term

- Adding a penalty term (equal to a constant times the number of splits) to the criterion ensures only significant improvements in goodness-of-fit are considered.
- The term is selected to maximize the goodness-of-fit criterion in cross-validation samples.

Cross-Validation Approach

- The training sample is repeatedly split into $S^{tr,tr}$ (for building a new tree and estimating conditional means) and $S^{tr,cv}$ (for evaluating estimates).
- A leaf-cost representing penalty parameter is used for pruning.
- The optimal penalty parameter is chosen by evaluating trees associated with each penalty value.
- The cross-validation goodness-of-fit criterion is $-MSE_{\mu}(S^{tr,cv}, S^{tr,tr}, \Pi)$.

Addressing Overfitting

- The cross-validation criterion directly addresses overfitting by penalizing too-extreme estimates of leaf means since $S^{tr,cv}$ is independent of $S^{tr,tr}$.
- Smaller leaf penalty leading to deeper trees and noisier estimates results in larger average MSE across cross-validation samples.

Outline

Honest Splitting: Modifying CART

- **Independent Sample Usage:**
  - Use an independent sample S^{est}, not S^{tr}, to estimate leaf means (first modification).
- **Modified Criteria:**
  - Adjust splitting and cross-validation criteria to account for unbiased estimates using S^{est} for leaf estimation (second modification).
  - Helps to eliminate one aspect of overfitting.
- **Variance Consideration:**
  - Treat S^{est} as a random variable during the tree-building phase.
  - Explicitly account for the increased variance in leaf estimates due to finer partitions.

Developing Criteria: Expansion of $-\text{EMSE}_\mu(\Pi)$

To begin developing our criteria, let us expand $-\text{EMSE}_\mu(\Pi)$:

$$\begin{aligned}
-\text{EMSE}_\mu(\Pi) &= -\mathbb{E}_{(Y_i,X_i),S^{est}}\left[(Y_i - \mu(X_i;\Pi))^2 - Y_i^2\right] \\
&\quad - E_{X_i,S^{est}}\left[(\hat{\mu}(X_i;S^{est},\Pi) - \mu(X_i;\Pi))^2\right] \\
&= \underbrace{\mathbb{E}_{X_i}\left[\mu^2(X_i;\Pi)\right]}_{\text{Goodness-of-fit term}} - \underbrace{\mathbb{E}_{S^{est},X_i}\left[\mathbb{V}\left[\hat{\mu}(X_i;S^{est},\Pi)\right]\right]}_{\text{Variance term}},
\end{aligned}$$

where we exploit the equality $E_S\left[\hat{\mu}(x;S,\Pi)\right] = \mu(x;\Pi)$.

Estimating -EMSE$_\mu(\Pi)$ and Variance

- Objective: Estimate -EMSE$_\mu(\Pi)$ based on the training sample S^{tr} and the size of the estimation sample N^{est}.
- Unbiased Estimator: Within each leaf, there is an unbiased estimator for the variance of the estimated mean, $\hat{\mu}(x; S^{est}, \Pi)$.
- Variance Estimation: To estimate the variance on the training sample, use

$$\hat{\mathbb{V}}(\hat{\mu}(x; S^{est}, \Pi)) \equiv \frac{S^2_{S^{tr}}(\ell(x; \Pi))}{N^{est}(\ell(x; \Pi))},$$

  where $S^2_{S^{tr}}(\ell)$ is the within-leaf variance.
- Expected Variance: Weight the variance estimator by the leaf shares p_ℓ to estimate the expected variance.
- Approximation: Assuming approximately equal leaf shares in estimation and training samples, approximate the expected variance as

$$\hat{\mathbb{E}}\left[\mathbb{V}\left[\hat{\mu}^2(X_i; S^{est}, \Pi)\right]\right]_{i \in S^{te}} \equiv \frac{1}{N^{est}} \cdot \sum_{\ell \in \Pi} S^2_{S^{tr}}(\ell).$$

Estimating the Average of the Squared Outcome

To estimate the average of the squared outcome $\mu^2(x;\Pi)$ (the first term of the target criterion), we can use:

$$\hat{\mathbb{E}}\left[\mu^2(x;\Pi)\right] = \hat{\mu}^2(x;S^{tr},\Pi) - \frac{S^2_{S^{tr}}(\ell(x;\Pi))}{\mathrm{N}^{tr}(\ell(x;\Pi))},$$

where $\hat{\mu}^2(x;S^{tr},\Pi)$ is the square of the estimated means in the training sample, and $S^2_{S^{tr}}(\ell(x;\Pi))$ is an estimate of its variance.

Unbiased Estimator for $\text{EMSE}_\mu(\Pi)$

Combining these estimators leads to the following unbiased estimator for $\text{EMSE}_\mu(\Pi)$:

$$-\widehat{\text{EMSE}}_\mu(S^{tr}, \text{N}^{est}, \Pi) \equiv \frac{1}{\text{N}^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(X_i; S^{tr}, \Pi) - \left(\frac{1}{\text{N}^{tr}} + \frac{1}{\text{N}^{est}} \right) \sum_{\ell \in \Pi} S^2_{S^{tr}}(\ell).$$

Comparing this to the criterion used in the conventional CART algorithm:

$$-\text{MSE}_\mu(S^{tr}, S^{tr}, \Pi) = \frac{1}{\text{N}^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(X_i; S^{tr}, \Pi),$$

the difference comes from the terms involving the variance. The honest criterion penalizes small leaf size by how the within-leaf MSE is weighted.

Estimation of the First Term of the Target Criterion

- The first term of the target criterion, $\hat{E}[\mu^2(x;\Pi)]$, is estimated using the square of the estimated means in the training sample, $\hat{\mu}^2(x;S^{tr},\Pi)$, minus an estimate of its variance.
- Formula:

$$\hat{E}[\mu^2(x;\Pi)] = \hat{\mu}^2(x;S^{tr},\Pi) - \frac{S^2_{S^{tr}}(\ell(x;\Pi))}{\mathrm{N}^{tr}(\ell(x;\Pi))}$$

Honest Estimator v.s. Adaptive CART

- Combining these estimators leads to an unbiased estimator for $\text{EMSE}_\mu(\Pi)$:

$$-\widehat{\text{EMSE}}_\mu(S^{tr}, \text{N}^{est}, \Pi) = \frac{1}{\text{N}^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(X_i; S^{tr}, \Pi) - \frac{1}{\text{N}^{tr} + \text{N}^{est}} \sum_{\ell \in \Pi} S^2_{S^{tr}}(\ell(x; \Pi))$$

- Comparing with the conventional CART criterion, $-MSE_\mu(S^{tr}, S^{tr}, \Pi) = \frac{1}{\text{N}^{tr}} \sum_{i \in S^{tr}} \hat{\mu}^2(X_i; S^{tr}, \Pi)$, the difference comes from the terms involving the variance.
- $S^2_{S^{tr}}(\ell(x; \Pi))$ is proportional to the MSE within the associated leaf. The difference between the adaptive and honest criteria is how the within-leaf MSE is weighted. The honest criterion penalizes small leaf size.

Outline

Honest Cross-Validation: Overview

Approximate Unbiasedness:

- $\hat{\text{EMSE}}_{\mu}(S^{tr}, \text{N}^{est}, \Pi)$ is approximately unbiased when Π is fixed.
- Evaluating splits during recursive partitioning on S^{tr} compromises this unbiasedness.

Overstating Goodness of Fit:

- Initial splits tend to group observations with similar, extreme outcomes.
- This often results in overstating the goodness of fit as the decision tree grows deeper.

Honest Cross-Validation: Criteria and Practical Considerations

Role of Cross-Validation:

- Remains crucial in honest estimation, albeit less critical than in conventional CART.

Honest Estimation Criterion:

- Utilizes $\widehat{\text{EMSE}}_{\mu}(\mathcal{S}^{tr,cv}, \text{N}^{est}, \Pi)$, focusing on the cross-validation sample $\mathcal{S}^{tr,cv}$.
- Provides an unbiased estimate for a fixed Π, though it may exhibit higher variance from the small size of the cross-validation sample.

Methodological Note:

- When applying $\widehat{\text{EMSE}}_{\mu}$, replace N^{tr} with $\text{N}^{tr,cv}$ to account for the cross-validation context.

Outline

Honest Inference for Treatment Effects I

- **Objective:**
  - Estimate **the average effect of conditional treatment** instead of the means of the conditional population.
- **Challenges:**
  - Treatment effects are not directly observable.
- **Data Structure:**
  - Observations are triples (Y_i^{obs}, X_i, W_i).
  - S^{treat} and $S^{control}$ denote subsamples of treated and control units, respectively.
- **Definitions:**
  - $\mu(w, x; \Pi)$: Population average outcome for treatment level w.
  - $\tau(x; \Pi) = \mu(1, x; \Pi) - \mu(0, x; \Pi)$: Average causal effect.

Honest Inference for Treatment Effects II

- **Estimations:**
  - $\hat{\mu}(w, x; S, \Pi)$: Estimated population average outcome.
  - $\hat{\tau}(x; S, \Pi) = \hat{\mu}(1, x; S, \Pi) - \hat{\mu}(0, x; S, \Pi)$: Estimated average causal effect.
- **MSE for Treatment Effects:**
  - $\text{MSE}_\tau(S^{te}, S^{est}, \Pi)$: Mean squared error for treatment effects.
  - $\text{EMSE}_\tau(\Pi) = E_{S^{te}, S^{est}}[\text{MSE}_\tau(S^{te}, S^{est}, \Pi)]$: Expected MSE.
- **Estimation Feasibility:**
  - $\text{MSE}_\tau(S^{te}, S^{est}, \Pi)$ is infeasible due to unobservable τ_i.
  - Estimation methods are developed to address this challenge.

Outline

Modifying Conventional CART for Treatment Effects I

- **Objective:**
  - Modifying conventional (adaptive) CART to estimate **heterogeneous treatment effects**.
- **Prediction Case:**
  - Using the fact that $\hat{\mu}$ is constant within each leaf, the **MSE** can be written as:

$$\begin{aligned}\text{MSE}_\mu(S^{te}, S^{tr}, \Pi) = &-\frac{2}{\text{N}^{tr}} \sum_{i \in S^{te}} \hat{\mu}(X_i; S^{te}, \Pi) \cdot \hat{\mu}(X_i; S^{tr}, \Pi) \\ &+ \frac{1}{\text{N}^{tr}} \sum_{i \in S} \hat{\mu}^2(X_i; S^{tr}, \Pi).\end{aligned}$$

Modifying Conventional CART for Treatment Effects II

- **Treatment Effect Case:**
  - Using the fact that $E_{S^{te}}[\tau_i | i \in S^{te} : i \in \ell(x, \Pi)] = E_{S^{te}}[\hat{\tau}(x; S^{te}, \Pi)]$, we construct an unbiased estimator of:

$$\widehat{\mathrm{MSE}}_{\tau}(S^{te}, S^{tr}, \Pi) \equiv -\frac{2}{\mathrm{N}^{tr}} \sum_{i \in S^{te}} \hat{\tau}(X_i; S^{te}, \Pi) \cdot \hat{\tau}(X_i; S^{tr}, \Pi) + \frac{1}{\mathrm{N}^{tr}} \sum_{i \in S^{te}} \hat{\tau}^2(X_i; S^{tr}, \Pi)$$

Modifying Conventional CART for Treatment Effects III

- **In-Sample MSE Criterion:**
  - Propose $\widehat{\text{MSE}}_\tau(S^{tr}, S^{tr}, \Pi) = -\frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\tau}^2(X_i; S^{tr}, \Pi)$ as an estimator of the infeasible goodness-of-fit criterion in the sample.
- **Cross-Validation:**
  - In the prediction case, use $-\widehat{\text{MSE}}_\mu(S^{tr,cv}, S^{tr,tr}, \Pi)$. For treatment effects, use an unbiased estimate of the infeasible analog, leading to $-\widehat{\text{MSE}}_\tau(S^{tr,cv}, S^{tr,tr}, \Pi)$.

Overview of Honest EMSE Modification

- Expands the $-\text{EMSE}_\tau(\Pi)$ to balance treatment effect heterogeneity against variance in leaf estimates.
- Expression for honest approach:

$$-\text{EMSE}_\tau(\Pi) = E_{X_i}\left[\tau^2(X_i;\Pi)\right] - \mathbb{E}_{S^{est},X_i}\left[\mathbb{V}\left(\hat{\tau}^2(X_i;S^{est},\Pi)\right)\right]$$

Estimation Methodologies for Honest Approach

- **Estimation for Splitting**:

$$-\widehat{\mathrm{EMSE}}_{\tau}(S^{tr}, \mathrm{N}^{est}, \Pi) = \frac{1}{\mathrm{N}^{tr}} \sum_{i \in S^{tr}} \hat{\tau}^2(X_i; S^{tr}, \Pi)$$
$$- \frac{1}{\mathrm{N}^{tr} + \mathrm{N}^{est}} \left(\sum_{\ell \in \Pi} \frac{S^{tr,2}_{\mathrm{treat}}(\ell)}{p} + \frac{S^{tr,2}_{\mathrm{control}}(\ell)}{1-p} \right)$$

- **Estimation for Cross-Validation** uses the cross-validation sample.

Honest CART Analogy and Criteria Impact

- Criteria analogous to honest version of CART, focusing on treatment effects rather than outcomes.
- **Rewards and Penalties**:
  - Rewards partitions that reveal strong heterogeneity in treatment effects.
  - Penalizes partitions that increase variance in leaf estimates.

Feature Selection and Criteria Distinction

- **Feature Selection**: Different terms in the criteria select features rewarding various treatment effect insights.
- **Distinction in Criteria**: More pronounced differences between adaptive and honest criteria for treatment effect estimation.
- **Role of Cross-Validation**: Uses $S^{tr,cv}$ sample for estimating treatment effects, ensuring robustness and validation of findings.

Causal Trees (CTs) Overview

- **CT-A (Adaptive)**: Uses $-\widehat{\text{MSE}}_\tau(S^{tr}, S^{tr}, \Pi)$ for both splitting and cross-validation.
- **CT-H (Honest)**: Employs $-\widehat{\text{EMSE}}_\tau(S^{tr}, \text{N}^{est}, \Pi)$ for splitting, with evaluation at cross-validation samples.

Characterization of ATE

- ATE can be represented in various ways, including
  1. Difference adjusted for covariates between treatment groups.
  2. Weighted average of outcomes.
  3. Through an influence or efficient score function.

$$\tau = E[\mu(1, Xi) - \mu(0, Xi)] \tag{1}$$

$$= E\left[\frac{YiWi}{e(Xi)} - \frac{Yi(1 - Wi)}{1 - e(Xi)}\right] \tag{2}$$

$$= E\Big[\frac{[Yi - \mu(1, Xi)]Wi}{e(Xi)} - \frac{[Yi - \mu(0, Xi)](1 - Wi)}{1 - e(Xi)}\Big] \tag{3}$$

$$+ E[\mu(1, Xi) - \mu(0, Xi)],$$

- $\mu(w, x) = E[Yi|Wi = w, Xi = x]$ and $e(x) = E[Wi|Xi = x]$.
- Choose representation based on desired estimation: conditional outcomes, propensity score, or both.

Transformed Outcome Trees (TOT)

- Uses transformed outcomes $Y_i^p = Y_i \cdot \frac{(W_i - p)}{(p \cdot (1-p))}$ for regression tree analysis.
- **TOT-A (Adaptive)**: Conventional CART with transformed outcomes. Estimates treatment effects using sample means.
- **TOT-H (Honest)**: Similar to TOT-A but with separate samples for estimation.
- **Advantages**: Simple implementation using standard CART.
- **Disadvantages**: Less efficient due to incomplete use of treatment indicator information.

Fit-Based Trees

- Focus on outcome's goodness of fit rather than treatment effect.
- Uses a linear model with treatment indicators for splitting.
- **Adaptive (F-A)**: Utilizes modified MSE function:

$$\text{MSE}_{\mu,W}(S^{te}, S^{est}, \Pi) = \sum_{i \in S^{te}} \left(Y_{i,obs} - \hat{\mu}_w(W_i, X_i; S^{est}, \Pi)\right)^2 - Y_i^2$$

- **Concerns**: Equal fit improvements may neglect meaningful treatment effect variations.

Squared T-Statistic Trees

- Splits based on the largest squared t-statistic (TS) to assess uniform treatment effects across splits:

$$T^2 = N \cdot \left(\frac{\overline{Y}_L - \overline{Y}_R}{\sqrt{\frac{S_L^2}{N_L} + \frac{S_R^2}{N_R}}} \right)^2$$

- **Cross-validation**: Uses standard fit measures or specific CT criteria to ensure depth without losing focus on treatment variations.
- **Concerns**: TS may overlook the benefit of fit improvement, potentially reducing variance in estimates.

Comparison of Tree-Based Estimators

Method	Approach	Splitting Criterion	Cross-validation and Pruning
CTs	Uses $-\widehat{\text{MSE}}_\tau$ and $-\widehat{\text{EMSE}}_\tau$ based on treatment effects.	Adaptive: $-\widehat{\text{MSE}}_\tau(S^{tr}, S^{tr}, \Pi)$, Honest: $-\widehat{\text{EMSE}}_\tau(S^{tr}, \text{N}^{est}, \Pi)$	Evaluated at the cross-validation samples.
TOT	Utilizes transformed outcomes for conventional CART to estimate treatment effects.	Both versions use the transformed outcome $Y_i^p = Y_i \cdot \frac{(W_i - p)}{(p \cdot (1-p))}$.	Same trees built; separate estimation sample for honest version.
F	Focuses on goodness of fit using a linear model with treatment indicators.	Uses goodness of fit (F) and modified MSE function.	Not specified; possibly uses conventional methods.
TS	Maximizes the squared t-statistic (TS) for treatment effects.	$T^2 = N \cdot \left(\frac{\overline{Y}_L - \overline{Y}_R}{\sqrt{\frac{S_L^2}{N_L} + \frac{S_R^2}{N_R}}} \right)^2$	Uses goodness-of-fit measures or CT-A and CT-H criteria.

Table: Comparison of Tree-Based Estimators in Treatment Effect Analysis

Comparison of Estimators: CT, F Criterion, and TS Criterion

Criterion	Focus	Emphasis
CT	Treatment Effect	Balanced
F	Goodness of Fit	Fit Improvement
TS	Treatment Effect	Pure Split Impact

Table: Comparison of Estimators

- **CT:** Incorporates benefits of improving fit along with the impact of the treatment effect.
- **F Criterion:** Focuses solely on improving the goodness of fit, measuring the overall fit improvement.
- **TS Criterion:** Primarily focuses on the heterogeneity in treatment effects without considering fit improvements.

Average Number of Leafs

Table 1. Simulation study

$N^{tr} = N^{est}$	Design 1		Design 2		Design 3	
Estimator	500	1,000	500	1,000	500	1,000
			No. of leaves			
TOT	2.9	3.2	2.9	3.5	3.6	5.4
F-A	6.1	13.1	6.3	13.0	6.2	13.0
TS-A	4.0	5.4	3.4	5.1	3.4	6.6
CT-A	4.0	5.5	3.2	3.7	3.5	5.4
F-H	6.0	12.9	6.3	13.0	6.3	13.1
TS-H	4.3	7.8	5.6	11.4	5.9	12.4
CT-H	4.2	7.6	5.6	11.4	6.1	12.5

Costs and benefits to honest estimation

Table 1. Simulation study

$N^{tr} = N^{est}$	Design 1		Design 2		Design 3	
Estimator	500	1,000	500	1,000	500	1,000
	Infeasible MSE divided by infeasible MSE for CT-H*					
TOT-H	1.554	1.938	1.089	1.069	1.081	1.042
F-H	1.790	1.427	1.983	2.709	1.502	2.085
TS-H	0.971	0.963	1.183	1.145	1.178	1.338
	Ratio of infeasible MSE: Adaptive to honest†					
TOT-A/TOT-H		1.021		0.754		0.717
F-A/F-H		0.491		0.985		0.993
T-A/T-H		0.935		0.841		0.918
CT-A/CT-H		0.929		0.851		0.785
	Coverage of 90% confidence intervals – adaptive					
TOT-A	0.82	0.85	0.78	0.81	0.69	0.74
F-A	0.89	0.89	0.83	0.84	0.82	0.82
TS-A	0.84	0.84	0.78	0.82	0.75	0.75
CT-A	0.83	0.84	0.78	0.82	0.76	0.79
	Coverage of 90% confidence intervals – honest					
TOT-H	0.90	0.90	0.90	0.89	0.89	0.90
F-H	0.90	0.90	0.90	0.90	0.90	0.90
TS-H	0.90	0.90	0.91	0.91	0.89	0.90
CT-H	0.89	0.90	0.90	0.90	0.89	0.90

*$\mathrm{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \pi^{\text{Estimator}}(\mathcal{S}^{tr}))/\mathrm{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \pi^{CT-H}(\mathcal{S}^{tr}))$.
†$\mathrm{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est} \cup \mathcal{S}^{tr}, \pi^{\text{Estimator-A}}(\mathcal{S}^{est} \cup \mathcal{S}^{tr}))/\mathrm{MSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \pi^{\text{Estimator-H}}(\mathcal{S}^{tr}))$.

For Further Reading I

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York:

Susan Athey and Guido Imbens.
Recursive partitioning for heterogeneous causal effects.
Proceedings of the National Academy of Sciences, 113(27):7353-7360, 2016.

Justin Grimmer, Solomon Messing, and Sean J. Westwood.
Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods.
Political Analysis, 25(4):413–434, 2017.

Sridhar Narayanan, Kirthi Kalyanam, and others.
Behavioral Targeting, Machine Learning and Regression Discontinuity Designs.
Tech Report, 2020.