Imperial College Business School

Beyond OLS: Estimating Treatment Heterogeneity with Causal Forests and Double Machine Learning

Zhanghao Li
CID: 02432954

A report submitted in partial fulfilment of the requirements
for the MSc Business Analytics degree

August 2024

Abstract

Inferring heterogeneous treatment effects (HTE) from data is a crucial topic in both academia and industry. Accurate inference of HTE relies on precise estimation of treatment effects. While traditional machine learning (ML) models excel at prediction, they often function as black boxes and lack interpretability.

This paper first discusses the potential weaknesses and limitations of traditional econometric tools. It then introduces causal forest, an extension of random forest that offers better interpretability and ability to capture causal relationships within data. Furthermore, it demonstrates how double machine learning (DML) and instrumental variables (IV) can be integrated within the causal forest framework to address biases in parameter estimation, particularly confounding bias. Simulations were conducted at each step to validate the efficacy of the models in estimating treatment effects. Finally, it applies R-learner causal forest and instrumental causal forest to the General Social Survey (GSS) and the Oregon Health Insurance Experiment (OHIE) datasets. The results are compared with classical methods such as OLS, logistic regression, and 2SLS, and illustrate the advantages of causal forest and instrumental causal forest.

Our findings demonstrate the effectiveness of the causal forest in providing accurate estimation of conditional average treatment effect (CATE), group average treatment effect (GATE) and inferring HTE, highlighting its potential for more accurate and insightful analysis.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Martin, for his continuous support, invaluable guidance, and insightful feedback throughout my research. His encouragement and expertise have been instrumental in the completion of this thesis.

I am also grateful to the faculty and staff at Imperial College Business School for providing an enriching academic environment and the resources necessary for this research.

Special thanks to my colleagues and friends for their encouragement and for providing a stimulating and supportive environment. Their constructive comments and suggestions have greatly enhanced the quality of this work.

Finally, I would like to thank my family, my parents Jiangfeng Li and Juan Zhang for their unwavering support and understanding. Their love and patience have been a source of strength and motivation throughout my studies.

Contents

1 Introduction

Causal inference is a crucial topic in today's academic empirical research and real-world application. Given data, researchers seek to infer causal relationships and make accurate estimations. A key aspect of causal inference is estimating **heterogeneous treatment effects (HTE)**, which examines how treatment effects vary across different subgroups of samples. This concept has numerous applications, such as understanding how different patients react to the same drug in medical studies, customizing recommendations for customers in technology companies, and evaluating A/B tests. An important measure for inferring heterogeneity is **conditional average treatment effect (CATE)**, CATE represents expected treatment effect conditioned on specific covariates. If CATE varies across different samples, it can indicate heterogeneity. Estimating CATE is therefore valuable for inferring HTE.

Classical econometric methods for parametric estimation include ordinary least squares (OLS) and two-stage least squares (2SLS). They are efficient when the functional form is correctly specified and their assumptions hold or nearly hold. However, these methods have limitations because in the real world we don't know true functional form and their assumptions might be violated. Additionally, the coefficients in linear models are constants, which makes it challenging to estimate CATE and infer heterogeneity. Machine learning methods offer flexible functional forms with fewer assumptions and can regularize data effectively. However, traditional ML models often function as black boxes and lack interpretability.

Machine learning (ML) methods can be effectively applied to causal inference, leveraging their flexibility and prediction accuracy to enhance causal effect estimation (Athey 2019). The **causal tree**, introduced by Athey and Imbens (2016), is an innovative model derived from classification and regression trees (CART) (Breiman et al. 1984). Causal tree recursively partitions the feature space to maximize the variation of treatment effects across different subgroups, providing CATE estimation for each sample. To prevent overfitting, it employs an **honest split** approach. Building on causal tree, **causal forest** (Wager and Athey 2018) uses ensemble learning techniques, applying bagging (Breiman 1996) to average the results of multiple causal trees, thereby offering more stable CATE estimations.

While causal forest provides robust estimates for CATE, it does not adequately address common types of estimation bias. Athey, J. Tibshirani, and Wager (2019) introduced the **generalized random forest**, which incorporates the **double machine learning (DML)** technique (Chernozhukov, Chetverikov, et al. 2017) into the causal forest framework to address biases, with a particular focus on confounding bias. This enhanced approach significantly outperforms the initial version of causal forest (Wager and Athey 2018). Additionally, by incorporating instrumental variables (IV) with DML into the causal forest, they developed **instrumental causal forest**, effectively mitigating endogeneity issues.

1.1 Outline and Overview

In this article, Section 1 provides the background and importance of inferring HTE, broadly discussing potential problems with classical estimation methods and explaining how causal forest and its extensions can address these issues to improve HTE inference.

Section 2 reviews the literature that forms the foundation for causal forest and DML, as well as related research on utilizing ML approaches to parameter estimation.

Section 3 is methodology. In Section 3.1, we introduce potential outcome framework, which is fundamental to causal inference, and discuss common measures for inferring HTE. We use simulations to demonstrate the limitations of classical OLS in estimation. Section 3.2 describes the methodology of the initial version of causal forest and uses simulations to show that, as an adaptive nearest neighbor method, it outperforms the benchmark K-NN in CATE estimation. Section 3.3 introduces the typical types of biases in estimation and discusses meta-learners, which are machine learning approaches for estimating CATE. It explains DML techniques and how they can be integrated with causal forest's framework to address these biases, particularly confounding bias, which is the R-learner causal forest. Additionally, it introduces the instrumental causal forest, which combines IV with DML in the causal forest framework to handle endogeneity. Simulations demonstrate that DML combined with causal forests provides robust CATE estimations that are resilient to confounding bias.

In Section 4, we apply R-learner causal forest to General Social Survey (GSS) dataset and instrumental causal forest to Oregon Health Insurance Experiment (OHIE) dataset. We compare the estimation results of classical OLS, logistic regression, and 2SLS regression with those of R-learner causal forest and instrumental causal forest, demonstrating the advantages of the latter two methods in providing more accurate estimation inference for HTE.

2 Literature Review

Machine learning (ML) is experiencing rapid growth and offers flexible, powerful prediction models. Numerous effective algorithms have been developed, including classification and regression trees (CART) (Breiman et al. 1984), random forests (Breiman 2001), LASSO (R. Tibshirani 1996), ridge regression (Hoerl and Kennard 1970), neural networks, as well as boosting techniques like AdaBoost (Freund and Schapire 1996) and XGBoost (Chen and Guestrin 2016). These models have been effectively applied across various industries. Despite their predictive success, many of these models operate as 'black boxes,' lacking interpretability (Wager 2016).

Traditionally, the OLS regression has been the classical method in econometric treatment effect estimation. However, the OLS is biased when its assumptions are violated, particularly in non-linear relationships (Baiardi and Naghi 2024). Additionally, OLS struggles in high-dimensional settings when data is sparse or corrupted (Singh 2023). Propensity score matching (PSM) (Abadie and Imbens 2006) estimates treatment effects using a nearest-neighbor approach but also faces challenges in high-dimensional settings (Hastie, R. Tibshirani, and Friedman 2009).

With advancement of ML and AI, Efron (2020) highlighted that ML provides statisticians with a powerful toolkit, though further development is needed for practical applicability. Athey (2019) noted that ML models offer flexible functional forms and improved prediction accuracy, enhancing estimation accuracy.

Under the Neyman-Rubin potential outcome framework (Imbens and D. B. Rubin 2015) and assuming unconfoundedness (Rosenbaum and D. B. Rubin 1983), many studies have utilized ML for parameter estimation. Laan and Rose (2011) employed ML techniques alongside maximum likelihood estimation to improve parameter estimation. Johansson, Shalit, and Sontag (2016) introduced deep neural networks for estimating treatment effects, demonstrating superior performance compared to traditional PSM and logistic regression. Chernozhukov, Chetverikov, et al. (2018) introduced double machine learning (DML) techniques, which combine ML prediction and parameter estimation using the "Neyman-Orthogonality" condition. DML bridges the gap between ML prediction and parametric estimation, efficiently handling high-dimensional nuisance parameters η_0 and ensuring $N^{-1/2}$ consistency in parameter estimation. DML can be integrated with partially linear models using Robinson's transformation (Robinson 1988) to address confounding bias and enhance treatment effect estimation (Chernozhukov, Chetverikov, et al. 2017). Additionally, it is doubly robust (Kennedy 2023). Shi, Blei, and Veitch (2019) expanded on DML by applying neural networks to treatment effect estimation, demonstrating improved performance over other neural-network-based and classical estimation methods.

Estimating heterogeneous treatment effects (HTE) is a key focus in causal inference. The conditional average treatment effect (CATE) is a vital metric for understanding HTE, leading to the development of several ML methods tailored for its estimation. Kunzel et al. (2019) proposed meta-learners for CATE estimation, introducing the X-learner, which addresses regularization bias but sensitive to confounding bias. Bayesian additive regression trees (BART) (Hill 2011; Chipman, George, and McCulloch 2010) are also effective for CATE estimation, as utilized by Green and Kern (2012) on the general social survey dataset. Hahn, Murray, and Carvalho (2020) combined BART with causal forests to address both confounding and regularization biases, though bayesian models require accurate prior estimation, which can be challenging. Causal forests (Wager and Athey 2018), which build upon the random forest method (Breiman 2001), partition the feature space into subgroups to build causal trees, using random feature selection and bagging (Breiman 1996) on multiple causal trees for CATE estimation. Wager (2016) provided an asymptotic analysis of causal forests, demonstrating its logarithmic scalability in high-dimensional cases. To address confounding bias, Athey, J. Tibshirani, and Wager (2019) enhanced causal forests by incorporating DML, leading to a substantial improvement in performance, especially confounding robustness. They also introduced instrumental causal forests, which apply instrumental variables (IV) within the DML framework to address endogeneity. Nie and Wager (2020) demonstrated that various ML models could be utilized in the DML process of causal forests. Beyond CATE, group average treatment effect (GATE) and local average treatment effect (LATE) (Imbens and Angrist 1994; Angrist, Imbens, and D. Rubin 1996) are also important metrics to infer HTE by measuring treatment effects within specific subgroups.

Causal forests have been applied across diverse fields to infer HTE, including epidemiology (Jawadekar et al. 2023), energy markets (O'Neill and Weeks 2018), education (Athey and Wager 2019), labor markets (Davis and Heller 2017). Its potential spans academic research, industry application, and government policymaking.

3 Methodology of Causal Forest

This section outlines the methodology employed in this research, focusing on causal forests framework and its enhancements through the integration of double machine learning (DML) and instrumental variables (IV).

Section 3.1 introduces the potential outcome framework and the general measures for heterogeneous treatment effects (HTE). We discuss the limitations of classical OLS regression in this context and use simulations to validate these points. Section 3.2 details the foundational aspects of the causal forest methodology, including its splitting criteria and how it extends from the random forest algorithm. In Section 3.3, we examine common types of bias in estimation, introduce the meta-learner approach to estimate CATE, and demonstrate how to combine causal forest framework with DML to achieve robust CATE estimation by addressing confounding bias.

The primary objective of this section is to demonstrate the advantages of incorporating ML techniques into causal inference, with a particular focus on the causal forest framework. We will illustrate how combining causal forests with DML enhances the accuracy and robustness of causal effect estimation, making it a powerful tool in the analysis of treatment effects.

3.1 Introduction to Machine Learning Causal Inference

Potential outcome framework is fundamental in ML causal inference, assuming that each sample has a set of outcomes for different treatment conditions. In treatment effect estimation, a key focus is on HTE, which examines how different subgroups of population react to a treatment. We introduce CATE and GATE as measures of HTE. Additionally, we introduce the limitations of traditional tool OLS, and how ML models offer more flexible approaches to handle complex data.

3.1.1 Potential Outcome Framework and General Measures for HTE

Given a set of observed i.i.d. samples $i = 1, 2, 3, \ldots, n$, potential outcome framework (Splawa-Neyman 1923; D. B. Rubin 1974) assumes each sample can be represented by a tuple (X_i, W_i, Y_i), comprising:

- A feature vector X_i of covariates in $\mathbb{R}^p$ (with p dimensions),
- A response variable Y_i in $\mathbb{R}$,
- A binary treatment assignment W_i.

This framework is also called Neyman-Rubin framework (Imbens and D. B. Rubin 2015), it assumes there are outcomes $Y_i(0)$ and $Y_i(1)$ correspond to treatment conditions $W_i = 0$ and 1, respectively. The treatment effect for sample i is **individual treatment effect (ITE)** τ_i, defined as difference in outcomes between two treatment states:

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

One important metric for evaluating treatment effect is **average treatment effect (ATE)** τ, which is defined as expected value of ITE:

$$\tau = \mathbb{E}[\tau_i] = \mathbb{E}[Y_i(1) - Y_i(0)]. \tag{2}$$

In practice, it's impossible to observe $Y_i(1)$ and $Y_i(0)$ simultaneously, so it's impossible to compute ITE τ_i directly from (1) and cannot directly estimate ATE using (2). However, ATE can be estimated through a **randomized controlled trial (RCT)**, which assumes treatment assignment W is entirely random across population:

$$\{Y_i(0), Y_i(1)\} \perp W_i. \tag{3}$$

Under RCT assumption (3), ATE can be calculated as follows:

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0] \\ &= \mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}[Y_i \mid W_i = 0].\end{aligned} \tag{4}$$

Although $Y_i(1)$ and $Y_i(0)$ cannot be observed simultaneously, we can compute ATE by $\tau = \mathbb{E}[Y_i \mid W_i = 1] - \mathbb{E}[Y_i \mid W_i = 0]$ under the RCT assumption (4).

In real-world scenarios, the RCT assumption is often difficult to satisfy. A more general assumption is that treatment assignment is random condition on same covariates X_i:

$$\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i. \tag{5}$$

This assumption is called **unconfoundedness** (Rosenbaum and D. B. Rubin 1983), can be made more tractable through **propensity score** $e(X_i)$ (6), which is the probability of a sample to be assigned treatment given its covariates X_i:

$$e(X_i) = \mathbb{P}(W_i = 1 \mid X_i). \tag{6}$$

Given the propensity score, unconfoundedness (5) is equivalent to:

$$\{Y_i(0), Y_i(1)\} \perp W_i \mid e(X_i). \tag{7}$$

Under unconfoundedness (5) or (7), ATE can be estimated by:

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1) \mid X_i, W_i] - \mathbb{E}[Y_i(0) \mid X_i, W_i]] \\ &= \mathbb{E}[\mathbb{E}[Y_i \mid W_i = 1, X_i] - \mathbb{E}[Y_i \mid W_i = 0, X_i]] \\ &= \mathbb{E}[\mu_1(X_i) - \mu_0(X_i)],\end{aligned} \tag{8}$$

$\mu_{(w)}(X_i)$ is the expected outcome given X and W: $\mu_{(w)}(X_i) = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$.

Heterogeneous treatment effect (HTE) is another critical aspect to infer, focusing on how different subgroups of a population respond to the same treatment. A key metric for HTE is **conditional average treatment effect**

(CATE), which measures expectation of τ given covariates X. Equation (8) can be modified to define CATE:

$$\text{CATE} = \tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]. \tag{9}$$

Beyond CATE, **group (generalized) average treatment effect (GATE)** is also important to infer HTE. Instead of focusing on a single point in the covariates' feature space, GATE represents the expected treatment effect across a subspace. Denoting feature space by X and a subspace by g_i, the feature space can be divided into p subspaces via partition Π:

$$\Pi = \{g_1, \ldots, g_p\}, \quad \text{with} \quad \bigcup_{i=1}^{p} g_i = \mathbb{X}. \tag{10}$$

GATE within subspace g_i, denoted as $\tau(g_i)$, is:

$$\text{GATE} = \tau(g_i) = \mathbb{E}[Y(1) - Y(0) \mid G = g_i]. \tag{11}$$

If GATE varies across different groups, it indicates the presence of HTE. Since GATE can be computed from CATE:

$$\tau(g_i) = \mathbb{E}[\tau(X_j) \mid X_j \in g_i], \tag{12}$$

CATE is crucial for estimating HTE.

In addition to GATE and CATE, metrics **local average treatment effect (LATE)**, and **conditional local average treatment effect (CLATE)**, and **group local average treatment effect (GLATE)** are also used to infer HTE. A more detailed introduction to these metrics is provided in section 3.3.4 on instrumental causal forests, as they are particularly relevant in the context of instrumental variable methods.

3.1.2 Limitation of Traditional Tools in Estimation

Ordinary least squares (OLS) method is classical in econometrics for estimating treatment effects. It assumes a linear relationship[1] and estimates ATE by running linear regression and interpreting the coefficient of W. In empirical studies, with p covariates, OLS assumes the true functional form is:

$$Y = \alpha_0 + \alpha_1 W + \sum_{i=2}^{p+1} \alpha_i x_{i-1} + \epsilon, \tag{13}$$

then run regression of Y on (W, X), estimated coefficient $\hat{\alpha}_1$ is taken as $\hat{\tau}$, the estimate of ATE. However, this approach may introduce model bias when the true functions $\mu_{(1)}(X_i)$ and $\mu_{(0)}(X_i)$ are not consistent, meaning $\mu_{(1)}(X_i) \neq \mu_{(0)}(X_i) + \tau$, details are as discussed in Appendix I.2. Considering (8), we can have a straightforward strategy for estimating τ without this type of model bias:

- Learn $\hat{\mu}_0(x)$ from samples where $W_i = 0$,
- Learn $\hat{\mu}_1(x)$ from samples where $W_i = 1$,

[1] Detailed assumptions are provided in Appendix I.1.

- Compute the estimated $\hat{\tau}$ by $\hat{\tau} = \mathbb{E}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{(1)}(X_i) - \sum_{i=1}^{n}\hat{\mu}_{(0)}(X_i)$.

In the OLS framework, including an intercept in X, and denoting coefficient vector as β_w, we have:

$$\hat{\mu}_{(w)}(X) = \beta_w X,$$

and $\hat{\tau}$ can be estimated as:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1 X_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_0 X_i = (\hat{\beta}_1 - \hat{\beta}_0)\bar{X}. \tag{14}$$

The traditional OLS method has its limitations, particularly when its assumptions do not hold and the true $\hat{\mu}_{(w)}(x)$ is not linear (Baiardi and Naghi 2024). To address this, ML models like random forests, can be employed using the same strategy to learn each $\hat{\mu}_{(w)}(x)$ and compute the average difference as $\hat{\tau}$. The following simulations, using linear and non-linear $\mu_{(w)}(x)$ settings, evaluate the performance of OLS versus ML[2] in estimating ATE.

Each simulation set sample size n to 500, 1000, 2000, and 3000. For each n, 100 trials are conducted using different random seeds, and the distribution of the estimates is plotted. In both cases, the true ATE is 0.05. All details on this simulation settings are in Appendix II.1:

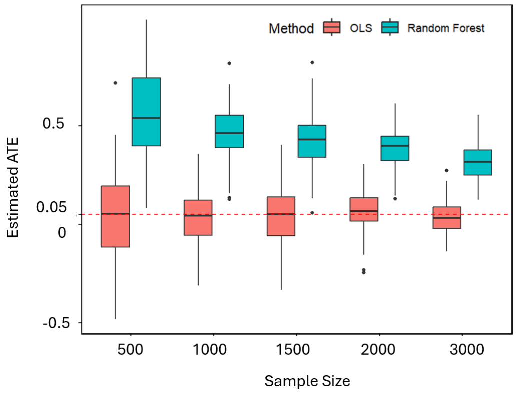


Figure 1: Linear setting: $\mu_{(0)}(X) = X_1 + X_2 + \epsilon$, $\mu_{(1)}(X) = X_1 + X_2 + 0.05 + \epsilon$

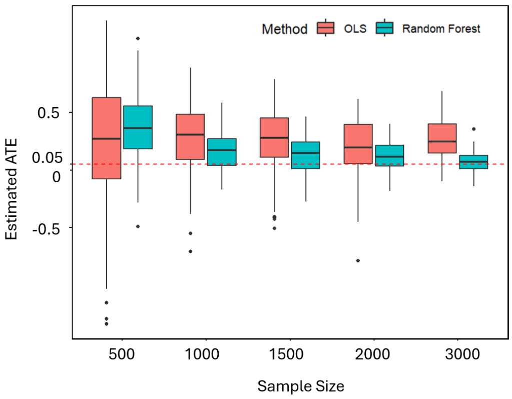


Figure 2: Non-linear setting: $\mu_{(0)}(X) = 4 \times \max(X_1, 0) + \frac{X_2^2}{2} + X_4^2 + \epsilon$, $\mu_{(1)}(X) = 4 \times \max(X_1, 0) + \frac{X_2^2}{2} + X_4^2 + 0.05 + \epsilon$

As the results show, OLS provides efficient and accurate estimates when the true relationship $\mu_{(w)}(x)$ is linear, and its parametric nature ensures a shorter computational time. However, in the non-linear setting, OLS introduces systematic bias. In contrast, regression forests provide estimates that converge to the true value as the sample size increases (Breiman 2004). However, due to their non-parametric nature, the computational time for regression forests is significantly longer than that for OLS, and it increases rapidly with sample size. Nevertheless,

[2] Regression forest is tuned via cross-validation, as with all simulations.

in empirical studies, the priority is often on obtaining unbiased estimates rather than minimizing computational time.

In addition to the systematic bias in non-linear settings, Singh (2023) highlight that OLS struggles with very high-dimensional data and fails when dealing with corrupted data.

Moreover, OLS is not well-suited for inferring HTE. Although it can measure GATE and CATE by adding interaction terms[3] $W * X$:

$$Y = \alpha_0 + \alpha_1 W + \sum_{i=2}^{p+1} \alpha_i x_{i-1} + \sum_{j=p+2}^{2p+1} \alpha_j W x_{j-1} + \epsilon, \tag{15}$$

this approach relies heavily on the linearity assumption and is prone to overfitting when p is large. The resulting functional form is also complex and difficult to manage.

Alternatively, OLS can estimate GATE by running separate regression for every subgroup g_i, yielding coefficient of W:

$$\hat{\tau}_{\text{OLS}}(g_i) = \text{lm}(Y \sim W + X, X \in g_i). \tag{16}$$

This approach also introduces bias when true underlying pattern is not linear and risks overfitting, as the sample size of each subgroup may be insufficient.

Compared with the limitation of classical approach, ML methods offer various alternative models that can be compared, incorporating the strengths of different ML algorithms into causal inference estimation and regularizing the model appropriately (Athey 2019).

3.2 Causal Forest

Random forest (Breiman 2001) is widely considered a highly successful ML prediction model, with applications across various fields. However, it operates as a "black box" model, often lacking interpretability. Causal forest, on the other hand, is a notable example of leveraging ML for treatment effect estimation, particularly for CATE. It extends the random forest framework by incorporating a new objective function based on expected mean squared error (EMSE) and employs honest splitting to prevent overfitting. In this section, we will illustrate algorithm behind causal trees and causal forests, and use simulations to demonstrate how causal forest, as an adaptive nearest neighbor method, can outperform the benchmark K-Nearest Neighbors (K-NN).

3.2.1 From CART to Honest Causal Tree

Classification and regression trees (CART) (Breiman et al. 1984) recursively partition feature space into subspaces (leaves ℓ) and provides a prediction for each leaf. The partition can be denoted as:

$$\Pi = \{\ell_1, \ldots, \ell_{\#(\Pi)}\}, \quad \text{with} \quad \bigcup_{j=1}^{\#(\Pi)} \ell_j = \mathbb{X}. \tag{17}$$

[3] Difference in difference (DiD) (Card and Krueger 1994) approach can be viewed as a special case of adding interaction term to OLS model.

For regression trees, the goal is to greedily (at each split) find the partition Π minimizes MSE

$$\text{MSE} = \sum_i \left[(Y_i - \hat{\mu}(X_i))^2 \right] \tag{18}$$

within each subspace ℓ. Equivalently, this can be viewed as maximizing the variance between different subspaces ℓ. For a split where the left and right leaves contain n_L and n_R samples, with predictions $\bar{y}_L$ and $\bar{y}_R$ respectively, the objective becomes:

$$n_L n_R (\bar{y}_L - \bar{y}_R)^2. \tag{19}$$

Regression trees uses a sample splitting method where the dataset S is divided into training set S^{tr} and test set S^{te}. Training set is for constructing partition Π and estimating model performance and tuning, typically through cross-validation, test set for evaluating final model. This sample splitting method is known as the **”adaptive split”**.

Athey and Imbens (2016) introduced causal trees for estimating CATE and provided valid confidence intervals. The causal tree has a similar structure to CART, as illustrated in Figure 3 below.[4]

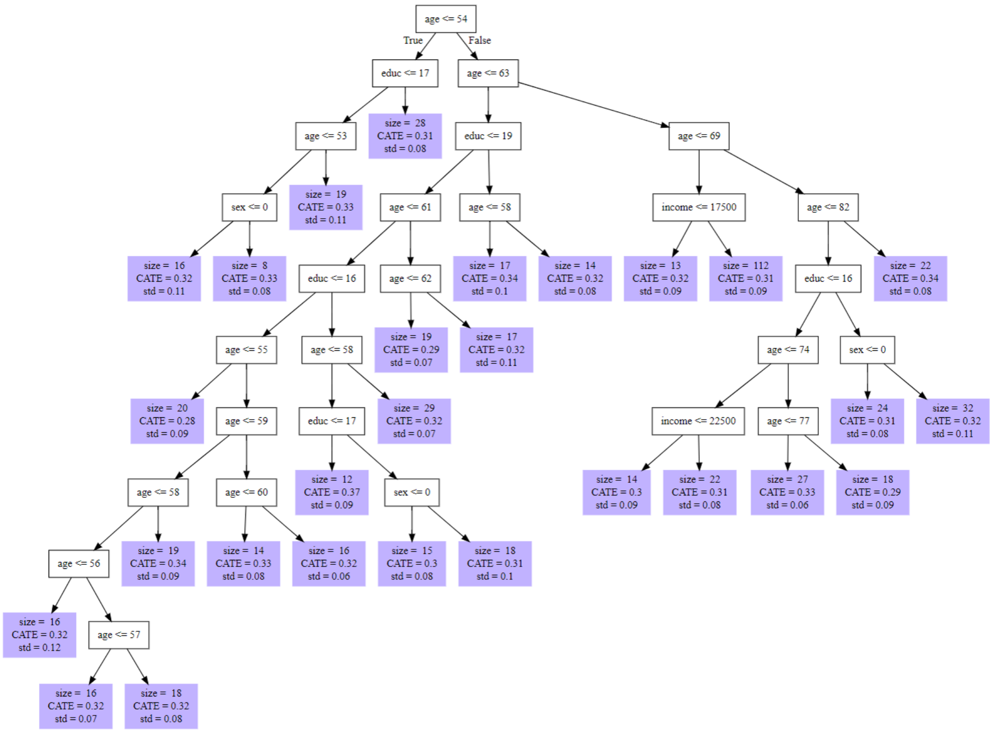

age <= 54
True
False
educ <= 17
age <= 63
age <= 53
size = 28 CATE = 0.31 std = 0.08
educ <= 19
age <= 69
sex <= 0
size = 19 CATE = 0.33 std = 0.11
age <= 61
age <= 58
income <= 17500
age <= 82
size = 16 CATE = 0.32 std = 0.11
size = 8 CATE = 0.33 std = 0.08
educ <= 16
age <= 62
size = 17 CATE = 0.34 std = 0.1
size = 14 CATE = 0.32 std = 0.08
size = 13 CATE = 0.32 std = 0.09
size = 112 CATE = 0.31 std = 0.09
educ <= 16
size = 22 CATE = 0.34 std = 0.08
age <= 55
age <= 58
size = 19 CATE = 0.29 std = 0.07
size = 17 CATE = 0.32 std = 0.11
age <= 74
sex <= 0
size = 20 CATE = 0.28 std = 0.09
age <= 59
educ <= 17
size = 29 CATE = 0.32 std = 0.07
income <= 22500
age <= 77
size = 24 CATE = 0.31 std = 0.08
size = 32 CATE = 0.32 std = 0.11
age <= 58
age <= 60
size = 12 CATE = 0.37 std = 0.09
sex <= 0
size = 14 CATE = 0.3 std = 0.09
size = 22 CATE = 0.31 std = 0.08
size = 27 CATE = 0.33 std = 0.06
size = 18 CATE = 0.29 std = 0.09
age <= 56
size = 19 CATE = 0.34 std = 0.09
size = 14 CATE = 0.33 std = 0.08
size = 16 CATE = 0.32 std = 0.06
size = 15 CATE = 0.3 std = 0.08
size = 18 CATE = 0.31 std = 0.1
size = 16 CATE = 0.32 std = 0.12
age <= 57
size = 16 CATE = 0.32 std = 0.07
size = 18 CATE = 0.32 std = 0.08

Figure 3: Example of causal tree structure, it partition the population into subgroups and provides CI estimation of CATE in each leaf.

Causal trees require the following assumptions:

- **Unconfoundedness**: $\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i$, as described in (5), given covariates, treatment assignment is random.

[4] This causal tree was constructed using the grf package in R, with data from the General Social Survey dataset.

- **Homogeneous Treatment Effect**: The treatment effect within each subgroup (leaf ℓ) is assumed to be the same.
- **Overlap (Positivity)**: $\eta < e(x) < 1 - \eta$ for $x \in \mathcal{X}$, where $0 < \eta < 0.5$. This assumption ensures that each subgroup has a sufficient treatment and control samples to provide accurate CATE estimation.

Causal trees differ from regression trees in two main aspects: (1) It uses **honest split** (also called cross-fitting) (Athey and Imbens 2016), which separates the samples used to construct the feature space partition Π from those used to estimate CATE; and (2) its outcome is CATE, which cannot be directly observed, unlike the outcomes in regression or classification tasks, where the true outcomes are observable. Thus, causal trees require a different objective function than the simple MSE in (19). We will focus on the honest split and the objective function of the honest causal tree.

Honest split divides population into three parts: S^{tr}, S^{est}, and S^{te}. The training set S^{tr} is for constructing causal trees (and cross-validation), estimation set S^{est} is for estimating treatment effects in each node, test set S^{te} is for evaluating estimation performance.

Primary reason for using honest split is **bias-variance trade-off**. The adaptive split uses the same data to determine both the tree structure and the estimates, which can lead to overfitting because tree is optimized to fit the peculiarities of data in training set. This overfitting can distort the true treatment effect. By separating the estimation and tree construction samples, the honest split reduces overfitting, providing more stable estimates and improving the model's generalization. Athey and Imbens (2016) demonstrated through simulations that the honest split outperforms the adaptive split in estimating τ in terms of both MSE and coverage.[5] We can also compare their performance through simulations, as discussed in Appendix II.2. In our simulations, the honest split performed significantly better than the adaptive split in estimating CATE under various settings.

Under the honest split, objective function for causal trees change from MSE to **expected mean squared error (EMSE)** of τ. For potential partitions Π, given the estimation set S^{est}, estimated CATE of sample i is denoted as $\hat{\tau}(X_i; \Pi, S^{\text{est}})$. The MSE and EMSE are defined as follows:

$$\begin{aligned}\text{MSE}(S^{\text{est}}, S^{\text{te}}) &= \frac{1}{n_{\text{te}}} \sum_{i \in S^{\text{te}}} \left(\tau_i - \hat{\tau}(X_i; S^{\text{est}}, \Pi)\right)^2 && (20)\\ &= \frac{1}{n_{\text{te}}} \sum_{i \in S^{\text{te}}} \left(\tau_i^2 - 2 \cdot \tau_i \cdot \hat{\tau}(X_i; S^{\text{est}}, \Pi) + \hat{\tau}^2(X_i; S^{\text{est}}, \Pi)\right), && (21)\end{aligned}$$

$$\begin{aligned}\text{EMSE} &= \mathbb{E}_{S^{\text{est}}, S^{\text{te}}} \left[\text{MSE}(S^{\text{est}}, S^{\text{te}})\right] && (22)\\ &= \mathbb{V}_{S^{\text{est}}, X_i} \left[\hat{\tau}(X_i; \Pi, S^{\text{est}})\right] - \mathbb{E}_{X_i} \left[\tau^2(X_i; \Pi)\right] + \mathbb{E}\left[\tau_i^2\right], && (23)\end{aligned}$$

[5] Coverage measures the probability that the confidence interval (CI) covers the true parameter, defined as Coverage = $\Pr(\theta \in \text{CI})$, where θ is the true parameter value and CI represents the confidence interval.

where $\mathbb{V}_{S^{\text{est}},X_i}\left[\hat{\tau}(X_i;\Pi,S^{\text{est}})\right]$ is the variance of $\hat{\tau}(X_i;\Pi,S^{\text{est}})$.[6] The objective function of honest causal tree is to minimize this EMSE.

Minimizing EMSE is same as maximizing variation of treatment effects between subgroups, derived from another objective function of regression trees in (19). But the prediction outcome is replaced with the estimated treatment effect:

$$n_L n_R(\hat{\tau}_L(X_i;\Pi,S) - \hat{\tau}_R(X_i;\Pi,S))^2. \tag{24}$$

The estimated treatment effect $\hat{\tau}(X_i;\Pi,S^{\text{est}})$ can be calculated by the average difference of outcome between treatment and control samples in each leaf ℓ in S^{est}. Denote the expected outcome of sample x under treatment w and partition Π as $\mu(w,x;\Pi)$:

$$\mu(w,x;\Pi) \equiv \mathbb{E}[Y_i(w) \mid X_i \in \ell(x;\Pi)], \tag{25}$$

ATE within each leaf ℓ is:

$$\tau(x;\Pi) \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i \in \ell(x;\Pi)] = \mu(1,x;\Pi) - \mu(0,x;\Pi). \tag{26}$$

Denote observed outcome be Y_i^{obs}. The estimated counterparts[7] are:

$$\hat{\mu}(w,x;\Pi,S) \equiv \frac{1}{\#(\{i \in S_w : X_i \in \ell(x;\Pi)\})} \sum_{i \in S_w : X_i \in \ell(x;\Pi)} Y_i^{\text{obs}}, \tag{27}$$

$$\begin{aligned} \hat{\tau}(x;\Pi,S) &\equiv \hat{\mu}(1,x;\Pi,S) - \hat{\mu}(0,x;\Pi,S) \\ &= \frac{1}{\#(\{i \in S_1 : X_i \in \ell(x;\Pi)\})} \sum_{i \in S_1 : X_i \in \ell(x;\Pi)} Y_i^{\text{obs}} \\ &- \frac{1}{\#(\{i \in S_0 : X_i \in \ell(x;\Pi)\})} \sum_{i \in S_0 : X_i \in \ell(x;\Pi)} Y_i^{\text{obs}}. \end{aligned} \tag{28}$$

The overall procedure for constructing a causal tree can be summarized as Algorithm 1, which is similar to regression tree (Breiman et al. 1984) but change objective function to EMSE and utilize honest split:

[6] Equation (23) is derived by substituting (20) into the EMSE definition and applying the relationship between expectation and variance: $\mathbb{V}[\tau_i] = \mathbb{E}[\tau_i^2] - \mathbb{E}^2[\tau_i]$.

[7] This is the initial estimation method by Athey and Imbens (2016). It is improved by Athey, J. Tibshirani, and Wager (2019) utilizing DML, which will be discussed in section 3.3.3

Algorithm 1 Procedure for Causal Tree Construction

```
    Input: Observational data $D$ comprising $\{(X_i, Y_i, W_i)\}_{i=1}^n$
    Output: Estimated CATE in each leaf of the pruned partition $\Pi^*$
1:  Divide data into $S^{\text{tr}}$, $S^{\text{est}}$, $S^{\text{te}}$
2:  In $S^{\text{tr}}$, use greedy algorithm to recursively partition covariate feature space $\mathbb{X}$ into partition $\Pi$
      a) Select the split minimizes estimate of EMSE (or equivalently maximizes the estimate of the difference between subgroups) among all possible binary splits at each node
      b) Retain minimum number of treated and control units in each child leaf
3:  Use cross-validation to select the depth $d^*$ to minimize the estimated MSE of CATE, utilizing remaining folds as the validation set
4:  Pruning $\Pi$ to depth $d^*$ and select partition $\Pi^*$, pruning leaves which can improve in fitness
5:  Estimate CATE in each leaf of $\Pi^*$ using $S^{\text{est}}$
```

Following this algorithm, we can estimate the CATE for each sample using a causal tree, denoted as T_{CT}:

$$\hat{\tau}(x) = T_{\text{CT}}\left(x; \{(X_i, Y_i, W_i)\}_{i=1}^n\right). \tag{29}$$

3.2.2 Honest Causal Forest

Breiman (2001) introduced the random forest, an ensemble learning model based on CART. Random forests utilize random feature selection and bagging (Breiman 1996) on multiple CARTs to construct more stable and accurate prediction. It is proven to be consistent to the true value (Breiman 2004).

Building on this idea, Wager and Athey (2018) combined bagging and random feature selection on multiple causal trees and introduced causal forests, estimate CATE $\hat{\tau}(x)$ for each sample. It is given by:

$$\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B} T^*_{\text{CT,b}}(x; \{(X_i, Y_i, W_i)\}_{i=1}^n), \tag{30}$$

where B is the causal trees number, and $T^*_{\text{CT,b}}$ represents estimation provided by each causal tree (as defined in (29)). Each causal tree is built using random feature selection and bootstrap sampling, with the causal forest aggregating these results to provide a more stable CATE estimation. When employing honest splitting, this method is known as **honest causal forest**.

For random forests, the prediction can be expressed as an adaptive neighborhood weighting method (Lin and Jeon 2006). Specifically, the random forest prediction can be expressed as:

$$\hat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} \frac{Y_i \cdot \mathbf{1}\{X_i \in L_b(x)\}}{|L_b(x)|}.$$

Reordering the summation yields

$$\begin{aligned}\hat{\mu}(x) &= \sum_{i=1}^{n} \frac{1}{B} \sum_{b=1}^{B} Y_i \cdot \frac{\mathbf{1}\{X_i \in L_b(x)\}}{|L_b(x)|} \\ &= \sum_{i=1}^{n} Y_i \alpha_i(x),\end{aligned}$$

indicating the random forest functions as a model providing prediction by the weighted average of similar samples. Every sample adaptively determines the weights for the other samples, making it an adaptive neighborhood estimation method. Similarly, causal forests can be reformulated from (30) to:

$$\tau(x) = \sum_{i=1}^{n} \tau_i \alpha_i(x). \tag{31}$$

This suggests that the causal forest is an **adaptive nearest neighbor** method for estimating CATE. Consequently, causal forests can be compared to a benchmark K-Nearest Neighbor (KNN) matching (Abadie and Imbens 2006), which is a non-adaptive nearest neighbor method. It estimates the average difference in outcomes between treatment and control group for each sample's neighborhood $S(x)$ as:

$$\hat{\tau}_{\text{KNN}}(x) = \frac{1}{k} \sum_{i \in S_1(x)} Y_i - \frac{1}{k} \sum_{i \in S_0(x)} Y_i. \tag{32}$$

We can use simulations to compare the estimation performance of K-NN matching[8] and causal forest.

We set the treatment effect τ to be related to both X_1 and X_2, with the feature space dimension d set to 6 and 20. The detailed simulation settings are in Appendix II.3. As shown in Figure 4, causal forest outperforms K-NN, providing estimates closer to the true effect due to its flexibility in choosing the weight of nearest neighbors. K-NN's performance deteriorates as d increases from 6 to 20, a manifestation of the "curse of dimensionality[9]" problem in non-parametric methods (Hastie, R. Tibshirani, and Friedman 2009).

Wager and Athey (2018) also demonstrated through simulations that causal forests performs better than K-NN in coverage and MSE, validated by setting different k, reinforcing the effectiveness of the adaptive nearest neighbor approach.

Wager (2016) and Wager and Athey (2018) also proved the statistical asymptotic properties for HTE inference and CATE estimation of causal forests. Wager (2016) showed that causal forests exhibit high signal strength[10], enabling them to accurately detect HTE even in high-dimensional settings. The signal strength scales logarithmically with the dimension, so despite being non-parametric, causal forests perform well in identifying and estimating CATE in high-dimensional contexts.

[8] We set k = 50 for K-NN matching to ensure there are enough treatment and control samples within each sample's nearest neighbors.

[9] As dimension increases, computational complexity, data sparsity, and model overfitting tend to increase, making non-parametric methods struggle.

[10] Signal strength refers to how effectively the model can identify and estimate the treatment effect from noisy data.

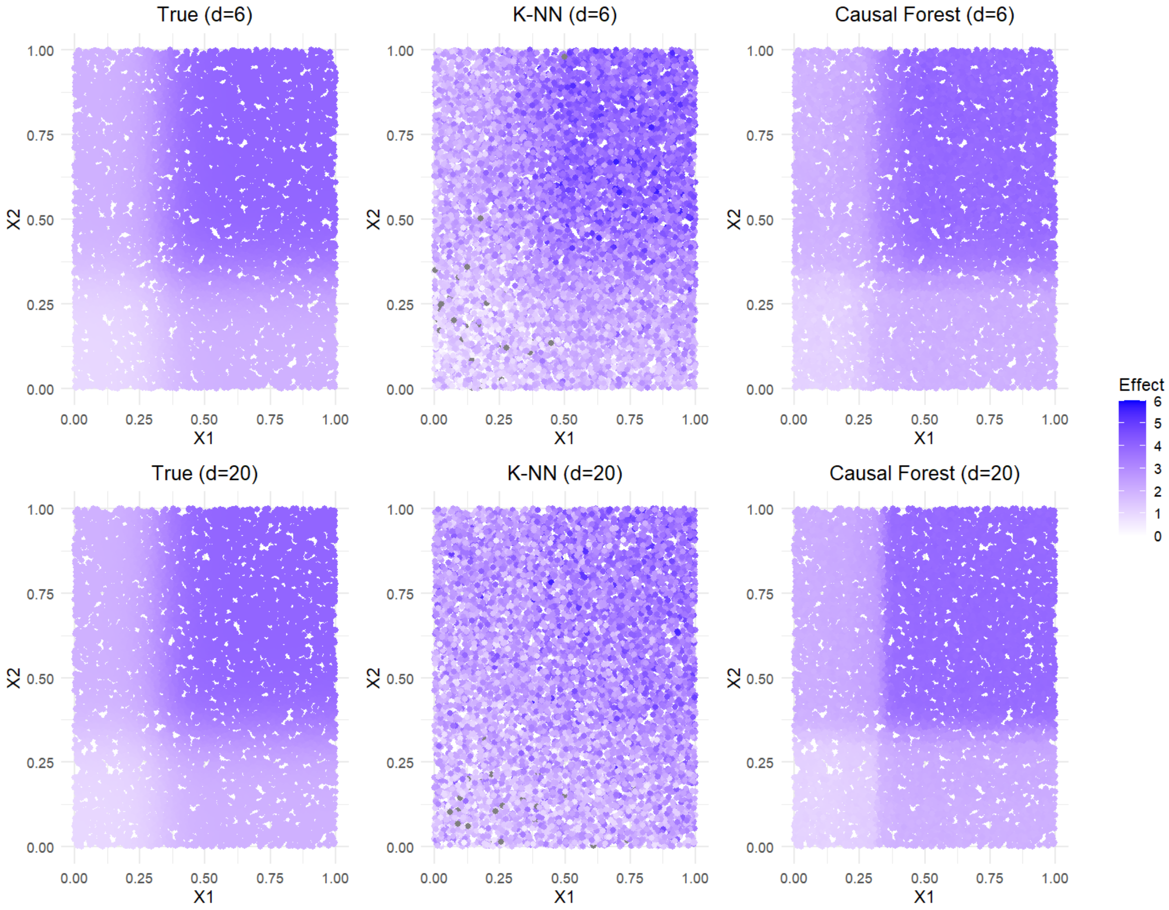


Figure 4: Causal forest vs. K-NN in estimating CATE under d = 6/20 and compare to the true τ. τ is related to both X_1 and X_2.

3.3 Causal Forest with DML

Previous section introduces the causal forest's framework. While they are effective at estimating CATE, they use simple averages to estimate treatment effects within each leaf, making them sensitive to bias (Athey, J. Tibshirani, and Wager 2019).

This section will firstly discuss the typical types of bias that can affect estimation. We will then explore techniques for estimating CATE, specifically focusing on meta-learners. Following this, we will delve into the methodology of DML and demonstrate how it can be integrated with causal forests to provide more robust, bias-resistant estimates, the R-learner causal forest. Finally, we will introduce instrumental causal forests, explaining how they combine instrumental variables (IV) with DML to handle endogeneity and further enhance estimation accuracy.

3.3.1 Typical Source of Bias in Estimation

When estimating treatment effects in the real world, there are several typical sources of bias: confounding bias, regularization bias, and model bias.

The first and most significant source of bias is **confounding bias**. Since in the real world it's difficult to achieve unconfoundedness described in (5). Confounding bias occurs when a confounder X influences outcome Y and treatment assignment W at the same time, making it hard to isolate the effect of W on Y. This leads to biased estimates of CATE. Confounding bias is especially prevalent in observational studies and can be illustrated with the following directed acyclic graph (DAG):

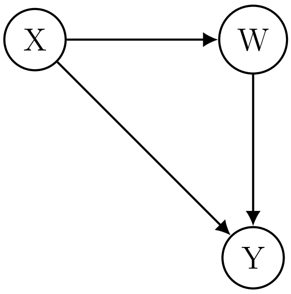


Figure 5: Confounding bias where X is confounder affecting both W and Y.

Another source of bias is **regularization bias**. This bias can arise due to sample size imbalances between treatment and control groups (also called imbalance bias in this case). Suppose we estimate CATE using $\hat{\tau}(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$, $\hat{\mu}_{(0)}(x)$ and $\hat{\mu}_{(0)}(x)$ are fitted on treatment and control group samples, respectively. Imbalance bias occurs because the model is more influenced by the group with the larger sample size. We illustrate this with the following simulation: with sample size $n = 100$, feature size $p = 1$, $X, \epsilon \sim N(0, 1)$, propensity score $e(x) = 0.1$, and outcome $Y = X + W \times \tau + \epsilon$. Here, we set treatment effect τ a constant 1, so there should be no HTE. We fit $\hat{\mu}_{(0)}(x)$ and $\hat{\mu}_{(1)}(x)$ by linear model:

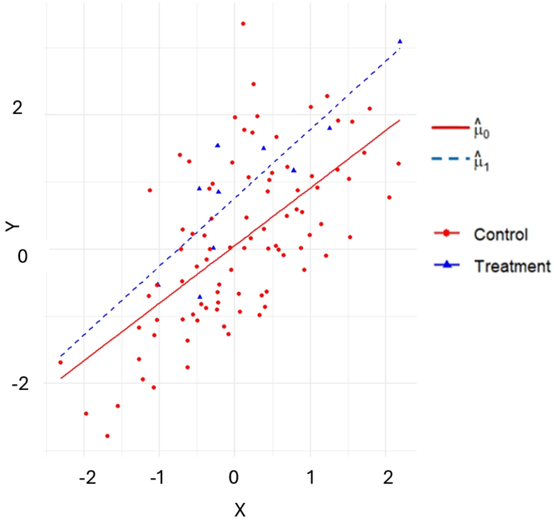


Figure 6: Imbalance bias illustration, there is no HTE in the setting but the estimation is showing HTE given imbalance sample size.

In this simulation, there should be no HTE since τ is constant at 1. However, Figure 6 shows an apparent increase in $\hat{\tau}(X)$ with X, visualized by the increasing distance between $\hat{\mu}_{(0)}(x)$ and $\hat{\mu}_{(1)}(x)$ as X increases. This bias occurs because the propensity score is 0.1, making a big difference in sample size, leading to differential fitting of $\hat{\mu}_{(0)}(x)$ and $\hat{\mu}_{(1)}(x)$.

Regularization bias also arises when machine learning models use regularization techniques. While regularization improves generalization and reduces variance, it introduces bias into parameter estimation. For instance, LASSO (R. Tibshirani 1996) uses L1 regularization:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{33}$$

In a linear setting, OLS provides unbiased estimates to the true parameter β, $E[\hat{\beta}_{\text{OLS}}] = \beta$ (Gauss-Markov Theorem). However, the LASSO regularization term in the objective function causes its estimates to be biased, such that $E[\hat{\beta}_{\text{LASSO}}] \neq \beta$.

Finally, **model bias** is another source of bias in treatment effect estimation. Model bias occurs when incorrect assumptions are made about the underlying data pattern in parametric model. For example, as shown in Figure 2 earlier in this section, when the true function is non-linear but a linear model is used (e.g., OLS), systematic bias occurs in the estimation. Similarly, using a single linear model to fit when $\mu_{(w)}(x)$ are not consistent (as discussed in Appendix I.2) will also lead to model bias.

3.3.2 Meta-learners in CATE Estimation

Estimation of CATE differs from that of ATE, as estimating ATE involves determining a constant τ, while estimating CATE requires estimating a function $\tau(x)$ to observe how the treatment effect varies across different samples. In this section, we will introduce meta-learners as some techniques in estimating CATE, which will also be benchmark to compare with causal forests in following section.

Kunzel et al. (2019) introduced the techniques of meta-learners to estimate CATE. Two straightforward methods are T-learners and S-learners.

T-learners (Two-Learners) are what we used to demonstrate imbalance bias in the previous section. It learns two separate functions from treatment and control groups, following this procedure:

- Learn $\hat{\mu}_{(0)}(x)$ from control samples.
- Learn $\hat{\mu}_{(1)}(x)$ from treatment samples.
- Estimate CATE by $\hat{\tau}(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$.

S-learners (Single-Learners) use all the data to fit one model. The procedure is as follows:

- Learn $\hat{\mu}(z)$ by predicting Y_i from $Z_i := (X_i, W_i)$ using all the data.
- Estimate $\hat{\tau}(x) = \hat{\mu}((x, 1)) - \hat{\mu}((x, 0))$.

Both of them have several weaknesses when it comes to estimating CATE. The first issue is imbalance bias, as shown in Figure 6, which arises when there is a non-negligible difference in sample sizes in treatment and control groups.

Additionally, these methods are susceptible to potential model bias. For instance, the S-learners may suffer from model bias if the true models for $\mu_{(w)}(x)$ are separate and cannot be combined into a single $\mu(z)$, as discussed in Appendix I.2.

This type of bias can also occur when using a parametric model but making incorrect assumptions about the underlying data pattern or guessing the functional structure incorrectly.

Furthermore, neither T-learners nor S-learners account for propensity score $e(x)$, which can lead to **confounding bias** if there are confounders X affect both treatment assignment W and outcome Y.

The **X-learner** (Cross-Learner) is a better technique introduced by Kunzel et al. (2019), it effectively deals with imbalance bias by weighting the estimated $\hat{\tau}_{(w)}(x)$ according to the estimated propensity score. The estimation procedure is as follows:

- Learn $\hat{\mu}_{(0)}(x)$ by predicting Y_i from X_i on the subset of observations where $W_i = 0$.
- Define $\Delta_i(1) = Y_i - \hat{\mu}_{(0)}(X_i)$, and learn $\hat{\tau}_{(1)}(x)$ by predicting $\Delta_i(1)$ from X_i on those observations where $W_i = 1$.
- Learn $\hat{\tau}_{(0)}(x)$ by swapping the roles of treated and control observations.
- Learn $\hat{e}(x)$ by predicting W_i from X_i.
- Report $\hat{\tau}(x) = \hat{e}(x)\hat{\tau}_{(0)}(x) + (1 - \hat{e}(x))\hat{\tau}_{(1)}(x)$.

X-learners perform well in RCT, as we will demonstrate in the simulations in the next section 3.3.3. However, they do not handle confounding bias. If confounders affect both W and Y, the estimation will still biased. In the next section, we will introduce DML to address confounding bias and explain how to apply DML techniques to causal forests (known as the R-learners, another type of meta-learners).

3.3.3 DML and its Application in Causal Forest: R-Learners

Although machine learning provides flexible models for estimating treatment effects, it can exhibit the typical biases discussed in section 3.3.1 when applied directly. Chernozhukov, Chetverikov, et al. (2018) introduced **double machine learning (DML)** to integrate modern ML methods into parameter estimation, addressing these biases. Athey, J. Tibshirani, and Wager (2019) applied DML within causal forests to provide confounding-robust CATE estimation.

In this section, we introduce the methodology of DML, focusing on its application in partially linear models. We will discuss its statistical properties and its use in CATE estimation (R-learners), then apply R-learners within causal forests framework, which we call R-learner causal forest and demonstrate its performance using simulations. We will also introduce an application of DML in estimating ATE, specifically the AIPW estimator $\hat{\tau}_{\text{AIPW}}$.

The DML methodology begins by identifying a **moment condition** $\psi(Y, W, X; \tau, \eta)$, where (Y, W, X) are the inputs from potential outcome framework, τ is the parameter of interest, and η is the nuisance parameter.[11] A valid moment condition must satisfy the expectation that:

[11] A nuisance parameter is one that is not of direct interest but must be estimated to determine the parameter of direct interest. In DML treatment effect estimation, η typically represents the outcome and propensity score, while the parameter of direct interest is τ.

$$\mathbb{E}[\psi(Y, W, X; \tau, \eta)] = 0. \tag{34}$$

Moment condition (34) ensures that the relationship between the nuisance parameter and the observed data is correctly specified. To perform DML estimation, the moment condition also needs to satisfy **Neyman-Orthogonality**, meaning that its partial derivative with respect to η at the true η_0 is zero:

$$\left.\frac{\partial \mathbb{E}[\psi(Y, W, X; \tau, \eta)]}{\partial \eta}\right|_{\eta=\eta_0} = 0. \tag{35}$$

Given Neyman-Orthogonality, the DML estimator is insensitive to small errors in the ML prediction of η. For a moment condition that satisfies Neyman-Orthogonality, DML uses ML models to predict the nuisance parameter $\hat{\eta}$ through **cross-fitting**[12] and plugs it into equation (34) to obtain the estimated parameter of interest $\hat{\tau}$ by solving:

$$\mathbb{E}[\psi(Y, W, X; \tau, \hat{\eta})] = 0. \tag{36}$$

A typical application of DML is the partially linear model. Assuming a constant treatment effect τ, the model can be expressed as:

$$Y = g(X) + W\tau + \epsilon. \tag{37}$$

If we denote the conditional expectation of Y given X as $m(x)$, and the conditional expectation of W given X as $e(x)$ (the propensity score), we have:

$$e(x) = \mathbb{E}[W \mid X = x], \tag{38}$$

$$m(x) = \mathbb{E}[Y \mid X = x] = E[g(X) + W\tau + \epsilon \mid X = x] = g(x) + e(x)\tau. \tag{39}$$

By subtracting (39) from (37), the partially linear model can be rewritten to be:

$$Y - m(x) = (W - e(x))\tau + \epsilon, \tag{40}$$

which is known as **Robinson's Transformation** (Robinson 1988). Following Robinson's Transformation (40), we can find a moment condition $\psi(Y, W, X; \tau, \eta)$ that satisfies Neyman-Orthogonality:

$$\psi(Y, W, X; \tau, \eta) = (Y - m(X) - \tau(W - e(X))) \cdot (W - e(X)). \tag{41}$$

We can proof (41) is a valid moment condition and that it is Neyman-Orthogonal (see Appendix I.3). Since the moment condition is Neyman-Orthogonal, we can plug (41) and estimated nuisance parameter $\hat{\eta}$ into (34) and solve the equation to obtain the DML estimate $\hat{\tau}$:

$$\mathbb{E}[\psi(Y, W, X; \tau, \hat{\eta})] = \mathbb{E}[(Y - \hat{m}(X) - \tau(W - \hat{e}(X))) \cdot (W - \hat{e}(X))] = 0. \tag{42}$$

[12] Cross-fitting is similar to honesty, as introduced in section 3.2.1. It means that the data used to build the ML models for the nuisance parameter are not used in estimation, primarily to reduce overfitting, as discussed in the context of honesty.

We can change the form of (42) to:

$$\mathbb{E}[(Y - \hat{m}(X)) \cdot (W - \hat{e}(X)) - \tau(W - \hat{e}(X))^2] = 0, \tag{43}$$

and use the summation rule of expectation:

$$\mathbb{E}[(Y - \hat{m}(X)) \cdot (W - \hat{e}(X))] = \mathbb{E}[\tau(W - \hat{e}(X))^2], \tag{44}$$

then by pulling out τ from the right-hand side given it is constant, (44) becomes:

$$\tau = \frac{\mathbb{E}[(Y - \hat{m}(X)) \cdot (W - \hat{e}(X))]}{\mathbb{E}[(W - \hat{e}(X))^2]}. \tag{45}$$

Using averages to estimate the expectations, the estimated $\hat{\tau}$ should be:

$$\hat{\tau} = \frac{\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))(W_i - \hat{e}(X_i))}{\sum_{i=1}^{n}(W_i - \hat{e}(X_i))^2}. \tag{46}$$

It turns out that (46) is the OLS solution of residual-on-residual regression of Y on W. If we denote $\tilde{Y} = Y - \hat{m}(x)$ and $\tilde{W} = W - \hat{e}(x)$, $\tilde{Y}$ and $\tilde{W}$ are also called pseudo-outcomes.[13] Then (46) can also be written as:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \tilde{Y}_i \cdot \tilde{W}_i}{\sum_{i=1}^{n} \tilde{W}_i^2} = \frac{\tilde{Y}^T \cdot \tilde{W}}{\tilde{W}^T \cdot \tilde{W}}, \tag{47}$$

which is the OLS expression of regressing $\tilde{Y}$ on $\tilde{W}$: $\tilde{Y} = \text{intercept} + \tau \times \tilde{W}$.

DML estimation has multiple advantages. First, it leverages flexible ML models for τ estimation by predicting η, which can efficiently handle high-dimensional nuisance parameters and complex datasets. Second, its point estimator $\hat{\tau}$ is $N^{-1/2}$ consistent with the true value (Chernozhukov, Chetverikov, et al. 2017):

$$\sqrt{N}\,(\hat{\tau} - \hat{\tau}^*) \xrightarrow{p} 0, (\text{N is the sample size}) \tag{48}$$

this $N^{-1/2}$ consistency only requires $N^{-1/4}$ consistency rate of the nuisance parameters $\hat{m}(x)$ and $\hat{e}(x)$, which is not difficult to achieve for ML models (random forest has a consistency rate of $N^{-1/4}$ proofed by Scornet, Biau, and Vert 2015). This is proofed[14] in Appendix I.4. Additionally, DML can handle confounding bias through residual-on-residual regression and regularization bias through Neyman-Orthogonality and cross-fitting. Moreover, the DML estimator $\hat{\tau}$ is also doubly robust, meaning it is consistent with the true τ when either $\hat{m}(x)$ or $\hat{e}(x)$ is consistent with the true function (Kennedy 2023). It also achieves quasi-oracle consistency.[15]

If we relax the constant treatment effect constraint, the partially linear model can be modified to:

$$Y = g(X) + W\tau(x) + \epsilon. \tag{49}$$

[13] In the DML setting, a pseudo-outcome is a transformation of the original outcome variable that makes the moment condition satisfy Neyman-Orthogonality.

[14] It can also be broadly proofed, since $\hat{m}(x)$ and $\hat{e}(x)$ are $N^{-1/4}$ consistent, making $(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))$ and $(W - \hat{e}(X))$ $N^{-1/4}$ consistent, and thus their product $N^{-1/2}$ consistent.

[15] Quasi-oracle consistency refers to the estimator's performance approaching that of an oracle estimator as the sample size increases Nie and Wager 2020.

We can use (49) to estimate CATE. In memory of Robinson (and also for its residual nature), this method for estimating $\tau(x)$ is called the R-learner (Nie and Wager 2020).

R-learners can be utilized within the causal forest framework. Athey, J. Tibshirani, and Wager (2019) introduced the "generalized random forest" and pointed out that the initial version of the causal forest in section 3.2.2 is just a special case of the "generalized random forest." As mentioned in section 3.2.1, causal trees assume that samples within each leaf are homogeneous in treatment effect. Given this assumption, it satisfies the partially linear model (37) in each leaf, so the residual on residual regression can be used to estimate τ within each leaf, but the fundamental algorithms of the causal forest, such as recursive partitioning, maximizing variation between leaves, bagging, etc., remain unchanged. We can call this approach the R-learner causal forest and its estimation of $\tau(x)$ can still be considered an adaptive neighborhood estimation as in (31), but it involves running a residual-on-residual regression of Y on W within each sample's neighborhood and extracting the coefficient term τ:

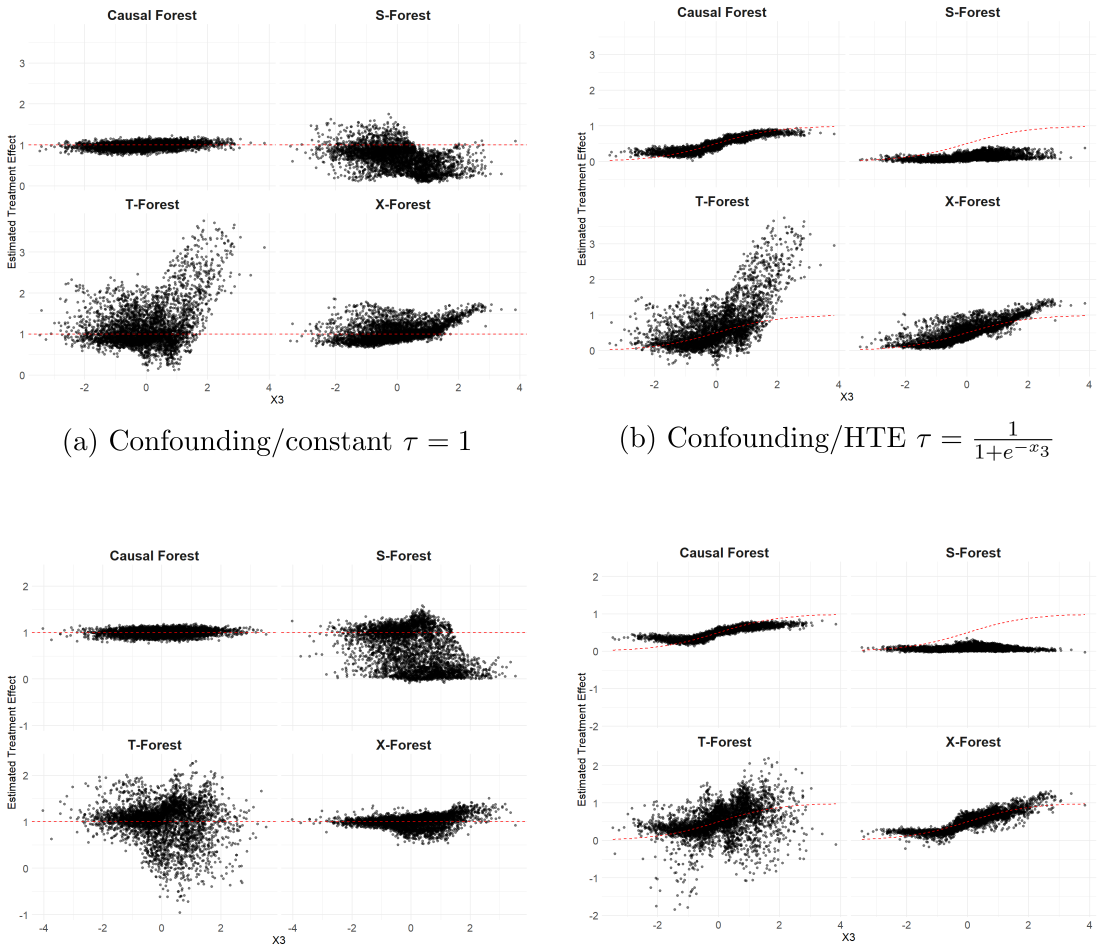
$$\hat{\tau}(x) = \operatorname{lm}\left(Y_i - \hat{m}^{(-i)}(X_i) \sim W_i - \hat{e}^{(-i)}(X_i), \text{ weights} = 1\{X_i \in \mathcal{N}(x)\}\right). \quad (50)$$

Here, $\mathcal{N}(x)$ is the leaf of each causal tree. The superscript (-i) denotes cross-fitting estimation. By utilizing residual-on-residual regression, R-learner causal forests address observed confounding bias. Athey, J. Tibshirani, and Wager (2019) used simulations and provided mathematical proof showing that it performs much better in estimating CATE than the initial causal forest in section 3.2.1, which utilizes simple averaging (28) to estimate $\tau(x)$. Its performance is superior in both capturing HTE and addressing confounding bias.

To further validate the performance of R-learner causal forests through simulations. We can compare it to other meta-learners introduced in section 3.3.2. To make these methods comparable, we use the same ML models (regression forest) to generate each meta-learner's ML predictions. The simulation settings are designed to include confounding/RCT, with HTE ($\tau(x)$ varies with x) / No HTE (constant $\tau(x)$), respectively. The details of the simulation are in Appendix II.4. The results are shown in Table 1 and Figure 7 below.

Confounding	**HTE**	**T-Forest**	**S-Forest**	**X-Forest**	**Causal Forest**
Y	N	0.32	0.18	0.05	0.01
Y	Y	0.32	0.18	0.03	0.01
N	N	0.17	0.29	0.02	0.004
N	Y	0.15	0.22	0.01	0.01

Table 1: MSE of different methods under different simulation settings



(a) Confounding/constant $\tau = 1$

(b) Confounding/HTE $\tau = \frac{1}{1+e^{-x3}}$

(c) RCT/constant $\tau = 1$

(d) RCT/HTE $\tau = \frac{1}{1+e^{-x3}}$

Figure 7: Comparison of treatment effect estimates of R-learner causal forest vs random forest with other meta learners. Red dash line is true $\tau(x)$, black dots are estimated $\hat{\tau}(x)$. Simulations are set to be confounding/RCT and constant treatment effect/HTE.

We can see that the R-learner causal forest outperforms all other meta-learners across various settings. It successfully addresses confounding bias through residual-on-residual regression, provides confounding-robust estimation, and accurately identifies and estimates HTE. X-forest also performs well when the RCT (unconfoundedness) assumption holds, but when there is confounding bias, its estimates are biased and unstable.

Apart from the partially linear model, another application of DML in estimating ATE is the Augmented Inverse Propensity Weighted (AIPW) estimator $\hat{\tau}_{\text{AIPW}}$ (Robins, Rotnitzky, and Zhao 1994):

$$\begin{aligned}\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\\ &+ \frac{W_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_{(1)}(X_i)) - \frac{1-W_i}{1-\hat{e}(X_i)}(Y_i - \hat{\mu}_{(0)}(X_i))\Bigg).\end{aligned} \tag{51}$$

The AIPW estimator $\hat{\tau}_{\text{AIPW}}$ improves from Inverse Propensity Weighted (IPW) estimator τ_{IPW}.[16] $\hat{\tau}_{\text{AIPW}}$ requires the overlap assumption as in section 3.2.1 to prevent the propensity score from being too close to 0 or 1, which would give extreme weight to the residual term in (51). $\hat{\tau}_{\text{AIPW}}$ is also doubly robust (proof in Appendix I.7) and provides $N^{-1/2}$ consistency when $N^{-1/4}$ consistent estimation of $\hat{\mu}_{(0)}(X_i)$, $\hat{\mu}_{(1)}(X_i)$, or $\hat{e}(X_i)$ is achieved, as proof in Appendix I.8. Its consistency rate is discussed in Appendix II.6. The AIPW estimator $\hat{\tau}_{\text{AIPW}}$ provides an efficient tool for estimating ATE τ.

3.3.4 IV: From 2SLS to Instrumental Causal Forest

Although R-learner causal forests discussed in the previous section provides confounding-robust estimation of CATE, it does not account for **endogeneity**, a common issue in observational studies. This section firstly introduces the problem of endogeneity, followed by a description of the classical method to address endogeneity—two-stage least squares (2SLS)—and its limitations. We then introduce the application of DML for instrumental variables (IV) in a partially linear model, and the application of IV within the DML framework in the form of instrumental causal forest. Finally, we utilize simulations to demonstrate its performance compared with R-learner causal forests described previously.

R-learner causal forests naturally assumes the treatment variable is exogenous, meaning W is uncorrelated with the error term ε in partially linear model (37):

$$Y = g(X) + W\tau + \varepsilon,$$

$$\mathrm{Cov}(W, \varepsilon) = 0. \tag{52}$$

However, in the real world, endogeneity is common in observational studies. Endogeneity implies that the explanatory variable (in treatment estimation, W) is correlated with the error term,[17] as shown below:

$$\mathrm{Cov}(W, \varepsilon) \neq 0, \text{ or } \mathbb{E}[W\varepsilon] \neq \mathbb{E}[W]\mathbb{E}[\varepsilon]. \tag{53}$$

In this case, both the treatment W and the error term ε affect the outcome Y. Since the error term ε affects both W and Y, the endogeneity problem can also be viewed as a form of **unobserved confounding bias**, where the error term ε acts as a confounder. This situation is depicted in Figure 8.

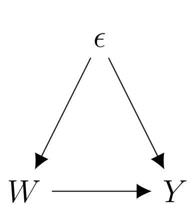


Figure 8: Endogeneity representation by DAG, ϵ affects both W and Y.

[16] $\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$. The estimation of the propensity score $\hat{e}(X_i)$ is crucial Hirano, Imbens, and Ridder 2003; more details on its properties can be found in Appendix I.5.

[17] Endogeneity can arise due to various reasons: omitted variable bias, measurement error, random shocks, simultaneous causality, reverse causality, etc.

While the R-learner causal forest in the previous section (46) uses residual-on-residual regression to eliminate observed confounding bias in the observed covariates X, it encounters challenges with unobserved confounding bias in the presence of endogeneity.

IV techniques are powerful tools for addressing endogeneity (Angrist and Pischke 2009). Denote the IV as Z and the endogenous variable as W. Z must satisfy the **relevance**, **exclusion restriction**, and **monotonicity** assumptions:

- **Relevance:** The IV Z must be correlated with endogenous explanatory variable W, satisfying $\text{Cov}(Z, W) \neq 0$.
- **Exclusion Restriction:** The IV Z affects the dependent variable Y only through the endogenous variable W, implying that Z is uncorrelated with the error term ϵ: $\text{Cov}(Z, \epsilon) = 0$.
- **Monotonicity:** There is no defier[18] in the population.

The graphical representation of an IV is shown in Figure 9.

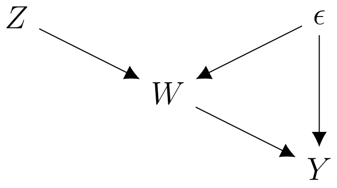


Figure 9: Instrumental variable Z affects Y only through W and addresses endogeneity.

The classical application of IV is the **Two-Stage Least Squares (2SLS)** approach (Angrist and Pischke 2014), which is widely applied (Angrist 1990). The methodology involves regressing W on Z in the first stage to obtain the predicted $\hat{W}$, and then regressing the outcome Y on the predicted $\hat{W}$ in the second stage:

- **Stage 1:** Use the model
$$W = \pi_0 + \pi_1 Z + \nu \tag{54}$$
to regress the endogenous variable W on the instrumental variable Z. We obtain the predicted values $\hat{W}$.
- **Stage 2:** Use the model
$$Y = \beta_0 + \beta_1 \hat{W} + \epsilon \tag{55}$$
to regress the outcome Y on predicted $\hat{W}$ from (55). We get:
$$Y = \hat{\beta}_0 + \hat{\beta}_{2\text{SLS}} \hat{W}. \tag{56}$$

[18] Defiers are samples take the treatment if not encouraged by the instrument and refuse the treatment if encouraged.

The 2SLS estimator $\hat{\beta}_{\text{2SLS}}$ can be expressed as:

$$\hat{\beta}_{\text{2SLS}} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(W_i - \bar{W})}, \text{ or } \hat{\beta}_{\text{2SLS}} = \frac{Z^T Y}{Z^T W}. \tag{57}$$

The mathematical derivation of $\hat{\beta}_{\text{2SLS}}$ in (57) is provided in Appendix I.9.

The treatment effect estimator $\hat{\beta}_{\text{2SLS}}$ is unbiased estimator of β under endogeneity, as Appendix I.10 proof. Imbens and Angrist (1994) and Angrist, Imbens, and D. Rubin (1996) showed $\hat{\beta}_{\text{2SLS}}$ estimates the **local average treatment effect** (LATE), which represents the treatment effect for compliers[19]. LATE can also be a measure for HTE since it reflects a subgroup's response to the treatment. If we denote the potential treatment under Z=1/0 as $D(1)/D(0)$, we have:

$$\beta_{\text{2SLS}} = \text{LATE} = \mathbb{E}[Y(1) - Y(0) \mid D(1) \neq D(0)]. \tag{58}$$

While 2SLS is widely used, it has limitations similar to those of classical models discussed in section 3.1.2. It inherently assumes that the true underlying relationships are linear in both stages, as in (54) and (55). However, in practice, these assumptions may be violated, with the true relationships being more complex or non-linear.

Chernozhukov, Chetverikov, et al. (2017) extended the DML methodology to IV settings, which can still be applied to the partially linear model with Robinson's Transformation (Robinson 1988). Following the DML methodology in section 3.3.3, we previously denoted the expectations of Y and W condition on covariates X as $m(x)$ and $e(x)$ in equations (39) and (38). Similarly, we can denote the expected value of Z condition on X to be $h(x)$:

$$h(x) = \mathbb{E}[Z \mid X = x]. \tag{59}$$

We can identify moment condition with a similar structure to Robinson's Transformation, where the potential outcome inputs change to $\{X, Y, W, Z\}$:

$$\psi(Y, W, Z, X; \tau, \eta) = (Y - m(X) - \tau(W - e(X)))(Z - h(X)). \tag{60}$$

It can be shown that (60) is a valid moment condition and Neyman-Orthogonal, as proofed in Appendix I.11. Therefore, we need to solve:

$$\mathbb{E}[\psi(Y, W, Z, X; \tau, \hat{\eta})] = 0, \tag{61}$$

expanding (60), we have:

$$\mathbb{E}[(Y - \hat{m}(X)) \cdot (Z - \hat{h}(X)) - \tau(W - \hat{e}(X)) \cdot (Z - \hat{h}(X))] = 0. \tag{62}$$

As in previous sections, we can use the summation rule of expectation and extract the constant τ, transforming (62) into:

$$\tau = \frac{\mathbb{E}[((Y - \hat{m}(X)) \cdot (Z - \hat{h}(X))]}{\mathbb{E}[((W - \hat{e}(X)) \cdot (Z - \hat{h}(X)))]}, \tag{63}$$

thus, estimated $\hat{\tau}$ can be calculated by:

[19] Compliers are individuals who take the treatment if encouraged by the instrument ($Z = 1$) and do not take the treatment if not encouraged by the instrument ($Z = 0$).

$$\hat{\tau} = \frac{\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))(Z_i - \hat{h}(X_i))}{\sum_{i=1}^{n}(W_i - \hat{e}(X_i))(Z_i - \hat{h}(X_i))}. \tag{64}$$

Letting $\tilde{Z} = Z - \hat{h}(x)$, (64) can also be written as:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \tilde{Y}_i \cdot \tilde{Z}_i}{\sum_{i=1}^{n} \tilde{W}_i \cdot \tilde{Z}_i} = \frac{\tilde{Z}^T \cdot \tilde{Y}}{\tilde{Z}^T \cdot \tilde{W}}. \tag{65}$$

It turns out that (65) is equivalent to (57), a 2SLS solution form when both stages involve residual-on-residual regression.

Athey, J. Tibshirani, and Wager (2019) introduced the **instrumental causal forest**. Similar to the R-learner causal forest in previous section, the fundamental algorithms of causal forests discussed in section 3.2 remain unchanged. However, it utilizes (65) to run a residual version of 2SLS within each leaf of the causal tree to estimate $\hat{\tau}$.

We conducted the following simulation with a strongly endogenous treatment variable W to validate the instrumental causal forest's ability to handle endogeneity. The detailed setting is provided in Appendix II.7. We used an IV $Z \sim \text{Bernoulli}(0.5)$ and an unobserved confounder $U \sim \text{Bernoulli}(0.5)$, with $W = U \times Z$ and true $\tau(x) = \frac{1}{1+e^{-X_1}}$. Table 2 and Figure 10 show the results, where the R-learner causal forest is systematically biased due to the unobserved confounder U.

Model	R-learner Causal Forest	Instrumental Causal Forest
MSE	0.49	0.03

Table 2: Comparison of MSE between Causal Forest and Instrumental Causal Forest

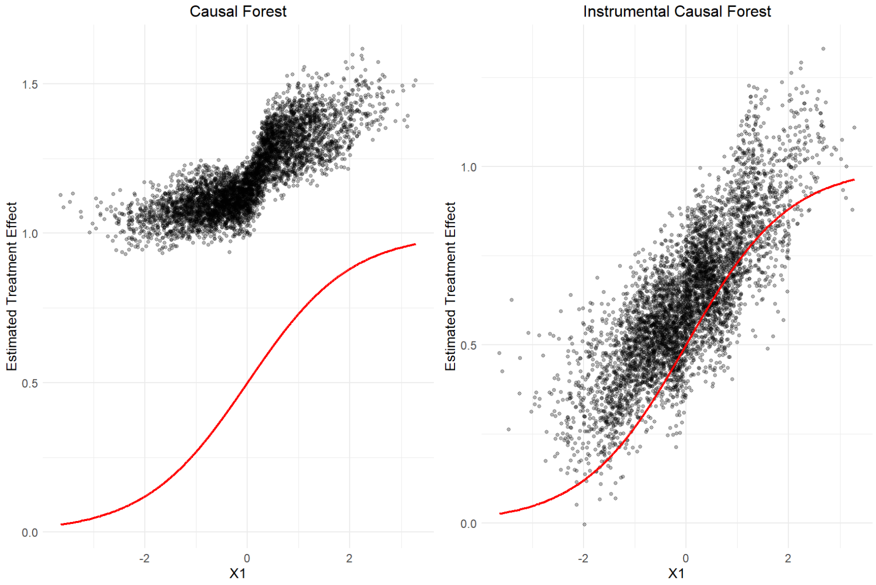


Figure 10: R-leaner causal forest vs. instrumental causal forest in strong endogeneity setting, red line is the true τ and black dots are estimations.

In combination with the concept of LATE, the instrumental causal forest effectively measures the **conditional local average treatment effect (CLATE)**, which shows how the treatment effect varies with X for different types of compliers:

$$\hat{\tau}_{\text{ICF}}(x) = \text{CLATE}(X) = \mathbb{E}[Y(1) - Y(0) \mid X; D(1) \neq D(0)]. \tag{66}$$

Similar to the definition of GATE in (11), we can define the expectation for a group of compliers as the **group local average treatment effect (GLATE)**:

$$\text{GLATE}(g) = \mathbb{E}[[Y(1) - Y(0) \mid X \in g; D(1) \neq D(0)]. \tag{67}$$

Both CLATE and GLATE are strong measures for identifying and estimating HTE.

It is also worth noting that both R-learner causal forests (47) and instrumental causal forests (65) ultimately rely on OLS/2SLS to estimate treatment effects. While machine learning provides flexible tools for parameter estimation, these advanced models still depend on classical statistical and econometric models for robust parameter estimation.

4 Application

In this section, we implement the R-learner causal forest on the GSS dataset and the instrumental causal forest on the OHIE dataset. The outcomes are compared with those derived from classical OLS, logistic regression, and 2SLS methods. Our analysis indicates that R-learner causal forests and instrumental causal forests offer improved certainty and precision in estimating HTE compared to classical approaches. But their estimates of ATE and LATE are similar to traditional methods.

4.1 General Social Survey Dataset

The General Social Survey (GSS) aims to understand how people react to welfare policies (Smith et al. 2018). It collects demographic features of respondents and their answers to various welfare-related questions to gauge their attitudes towards social welfare. As part of the survey, a randomized treatment was introduced, which altered the wording of a key question about government spending from question A to question B:

Question A: Do you think the government is spending too much on welfare?

Question B: Do you think the government is spending too much on assistance to the poor?

Respondents were required to choose Y/N in response to the question. This RCT allows for evaluating respondents' attitudes towards welfare based on the differences in their answers, and inferring heterogeneity by examining how different subgroups of the population react. Within the potential outcome framework, treatment W is whether the respondent was given question B rather than A, outcome Y is binary response, and covariates X are demographic features, described in Appendix III.1. We use survey data collected between 1986 and 2010, which contains 19,723 samples and 20 covariates after data cleaning.

Green and Kern (2012) used Bayesian additive regression tree (BART) to estimate both ATE and CATE for this dataset. The ATE estimated by BART was -0.364, with 95% confidence interval (CI) (-0.377, -0.351). Their analysis revealed significant heterogeneity in treatment effects, with CATEs ranging from -0.05 to -0.61. Conservatives (republicans) were more affected by the treatment than liberals (democrats), and younger people were more affected than elder.

Chernozhukov, Hansen, et al. (2024) applied Q-aggregation[20] ensemble on multiple ML models to estimate CATE of the change in wording. They estimated CATE range to be from -0.40 to -0.30, indicating heterogeneity, with political views and education being the most important features in determining respondents' attitudes towards welfare.

4.1.1 RCT Check

Firstly, we need to verify the validity of RCT. We utilized regression forest[21] to estimate $\hat{W}$ (or $\hat{e}(x)$). The distribution of estimated propensity scores is shown in Figure 11 below:

[20] Q-aggregation is a technique in parameter estimation that ensembles the estimations from multiple causal models and adaptively chooses the weight of each model based on a loss function.

[21] Tuned by cross-validation, similar to all the prediction models.

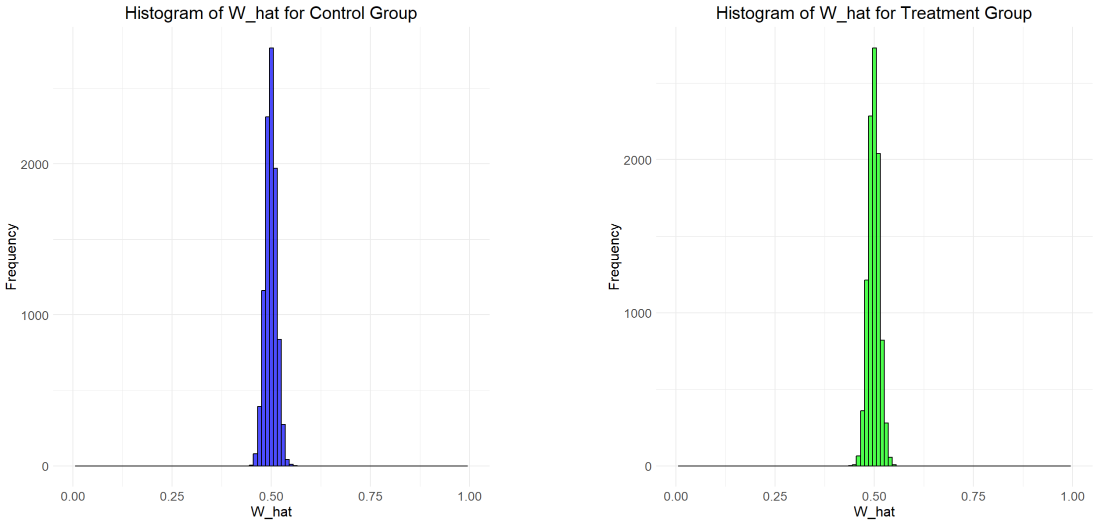


Figure 11: Estimated propensity score distribution by regression forest.

Since both histograms show a similar $\hat{W}$ distribution centered around 0.5, this provides evidence supporting the random assignment of treatment.

To further validate RCT, we conducted statistical testing (balance test) assuming mean of feature X_i is μ_1 in treatment group and μ_0 in control group, we tested the null hypothesis H_0: $\mu_0 = \mu_1$. Results are shown in Appendix III.2, indicate all covariates are statistically insignificant at a p-value threshold of 0.005, providing further evidence that RCT holds.

4.1.2 ATE Estimation with Classical Methods

Given the RCT nature of the dataset, there is no concern regarding endogeneity, which allows us to apply classical OLS and logistic regression to estimate ATE of the rewording treatment:

$$\begin{aligned}
Y = \beta_0 &+ \beta_1 \text{Treatment} + \beta_2 \text{Year} + \beta_3 \text{Weekly Hours Worked} \\
&+ \beta_4 \text{Number of Children} + \beta_5 \text{Age} + \beta_6 \text{Years of Education} \\
&+ \beta_7 \text{Gender} + \beta_8 \text{Race} + \beta_9 \text{Income} \\
&+ \beta_{10} \text{Political Views: Conservative} + \beta_{11} \text{Political Views: Extremely Conservative} \\
&+ \beta_{12} \text{Political Views: Extremely Liberal} + \beta_{13} \text{Political Views: Moderate} \\
&+ \beta_{14} \text{Political Views: Others} + \beta_{15} \text{Political Views: Slightly Conservative} \\
&+ \beta_{16} \text{Political Views: Slightly Liberal} + \beta_{17} \text{Marital Status: Never Married} \\
&+ \beta_{18} \text{Marital Status: Married} + \beta_{19} \text{Marital Status: Separated} \\
&+ \beta_{20} \text{Marital Status: Divorced} + \beta_{21} \text{Marital Status: Widowed} + \epsilon,
\end{aligned} \tag{68}$$

The estimated ATE from the OLS regression is -0.347 with 95% CI (-0.358, -0.335). Full regression results are shown in Appendix III.3.

Next, we employed logistic regression to estimate ATE of rewording, using the following model:

$$
\begin{aligned}
\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = {} & \alpha_0 + \alpha_1 \text{Treatment} + \alpha_2 \text{Year} + \alpha_3 \text{Weekly Hours Worked} \\
& + \alpha_4 \text{Number of Children} + \alpha_5 \text{Age} + \alpha_6 \text{Years of Education} \\
& + \alpha_7 \text{Gender} + \alpha_8 \text{Race} + \alpha_9 \text{Income} + \alpha_{10} \text{Political Views: Conservative} \\
& + \alpha_{11} \text{Political Views: Extremely Conservative} \\
& + \alpha_{12} \text{Political Views: Extremely Liberal} + \alpha_{13} \text{Political Views: Moderate} \\
& + \alpha_{14} \text{Political Views: Others} + \alpha_{15} \text{Political Views: Slightly Conservative} \\
& + \alpha_{16} \text{Political Views: Slightly Liberal} + \alpha_{17} \text{Marital Status: Never Married} \\
& + \alpha_{18} \text{Marital Status: Married} + \alpha_{19} \text{Marital Status: Separated} \\
& + \alpha_{20} \text{Marital Status: Divorced} + \alpha_{21} \text{Marital Status: Widowed} + \epsilon.
\end{aligned} \tag{69}
$$

The estimated $\hat{\alpha}_1$ from logistic regression is -2.023. Given logistic nature of model, we exponentiate $\hat{\alpha}_1$ and multiply by $P(Y = 0)$ to estimate ATE, resulting in an estimated ATE -0.393 with 95% CI (-0.398, -0.388). Full logistic regression results are available in Appendix III.3.

We applied R-learner causal forest to the dataset to estimate treatment effects. To determine the best model for predicting $\hat{m}(x)$ (or $\hat{Y}$), we compared different ML models. The RMSE of each model is summarized in table below:

Model	**RMSE**
OLS	0.44
LASSO	0.44
Ridge	0.44
XGBOOST	0.38
Regression Forest	0.42
Neural Network	0.45
SVR	0.50

Table 3: RMSE of Different Models in Estimating $\hat{m}(x)$

Based on the RMSE values, XGBoost was selected to predict $\hat{m}(x)$, and a R-learner causal forest was fitted. The AIPW estimator $\hat{\tau}_{\text{AIPW}}$ from it produced an ATE estimate of -0.346 with 95% CI (-0.357, -0.334). The comparison of three models' ATE estimations is illustrated in Figure 12 below, all estimates fall within range of -0.40 to -0.30.

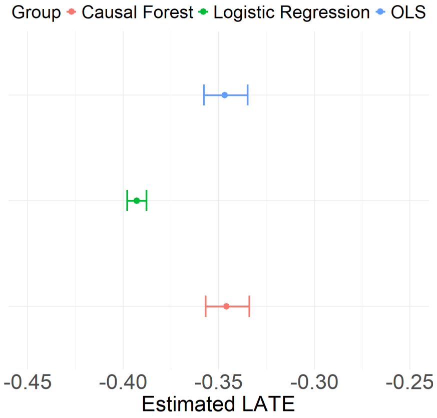


Figure 12: Comparison of estimating ATE by R-learner causal forest, OLS and logistic regression in GSS Dataset.

4.1.3 HTE Inference and GATE Estimation

To explore heterogeneity in treatment effects, we plot the out-of-bag[22] CATE distribution estimated by causal forest:

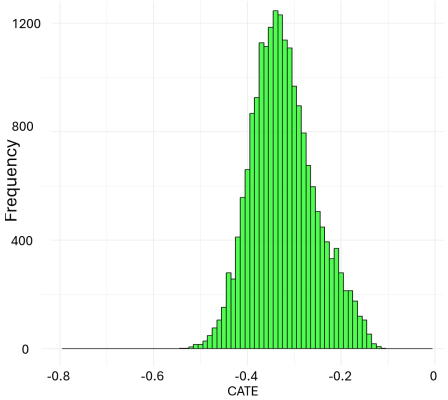


Figure 13: Distribution of estimated CATE by causal forest.

We can see CATE $\hat{\tau}(x)$ ranges from -0.5 to -0.1, suggesting variation in treatment effects across different respondents. We further assess heterogeneity using the treatment operating characteristic (TOC) curve and the QINI[23] curve, as discussed by Chernozhukov, Hansen, et al. (2024). The TOC curve measures the cumulative GATE over the ATE and is defined as:

$$\text{TOC}(q) = \text{GATE}(q) - \text{ATE}, \text{where GATE}(q) := \mathbb{E}[Y(1) - Y(0) \mid \hat{\tau}(X) \geq \mu(\hat{\tau}, q)].$$

QINI curve is calculated by multiplying the TOC curve by the probability of the condition:

$$\text{QINI}(q) := \text{TOC}(q) \times P(\hat{\tau}(X) \geq \mu(\hat{\tau}, q)).$$

More postive TOC and QINI curves provide stronger evidence of HTE. Their variances can be calculated based on group variance $V(q) = V(\hat{\tau} \mid \hat{\tau}(X) >$

[22] Out of bag prediction are prediction for those samples not selected by the bootstrap.

[23] Analogous to the Gini curve for classification models

$\mu(\hat{\tau}, q)$). The TOC and QINI curves with 95% confidence intervals are plotted below:

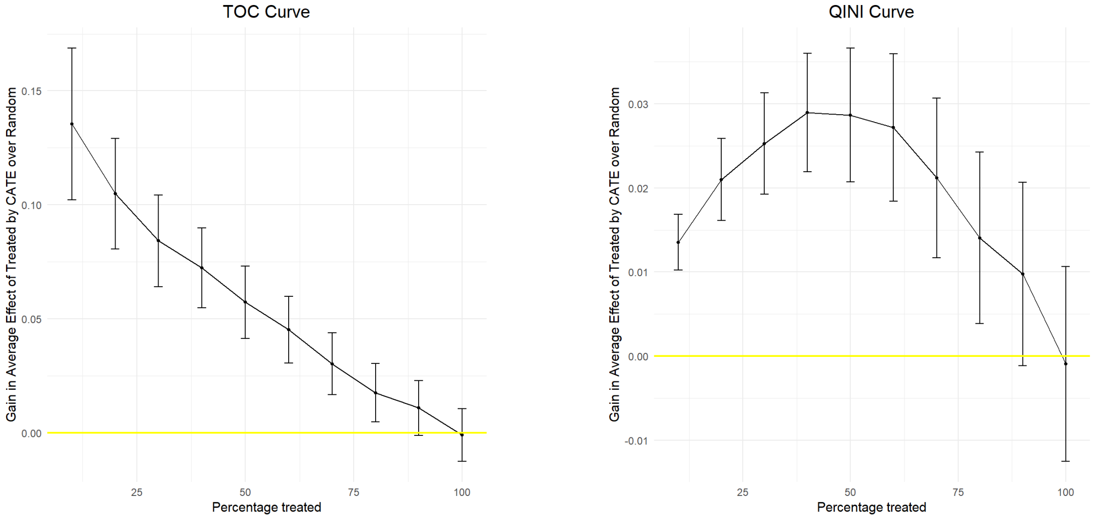


Figure 14: TOC and QINI curves, the more lower boundaries above 0, the more significant evidence for existing HTE.

Given that the lower bounds of TOC and QINI curves are mostly beyond zero, these curves confirm significant heterogeneity in the treatment effects. Further validation of heterogeneity is provided by the AUTOC and AUT QINI metrics (Chernozhukov, Hansen, et al. 2024), which are integrals of the TOC and QINI curves over the quantile q. The higher they beyond 0, the more significant for existing HTE. These are approximated as:

$$\widehat{\text{AUTOC}} = \sum_{\ell=1}^{10} \widehat{\text{TOC}}(q_\ell)\,(q_{\ell+1} - q_\ell)\,,$$

$$\widehat{\text{AUT QINI}} = \sum_{\ell=1}^{10} \text{QINI}(q_\ell)\,(q_{\ell+1} - q_\ell)\,.$$

The variances are computed as detailed in Appendix I.12. Their one-sided 95% confidence intervals are significantly greater than zero, as shown in Table 4, further supporting the presence of HTE.

	Estimation	**s.e.**	**One-Sided 95% CI**
AUTOC	0.0558	0.0000936	[0.0556, Infty]
AUT QINI	0.0189	0.0000207	[0.0189, Infty]

Table 4: AUTOC and AUT QINI Confidence Intervals

Given the existence of HTE, we proceed to estimate HTE using classical OLS and logistic regression models, and compare the results with those from the R-learner causal forest. We choose metric GATE rather than CATE since interaction term and grouping are commonly used in classical approaches (Angrist and Pischke 2009). We start by partitioning each feature into subgroups and using the fitted R-learner causal forest CATE estimates within each subgroup to estimate GATE:

$$\hat{\tau}_{\text{Causal Forest}}(g_i) = \mathbb{E}[\hat{\tau}_{\text{Causal Forest}}(X_j) \mid X_j \in g_i],$$

for OLS, we start by running following regression (70) on each subgroup g_i to obtain the treatment coefficient and its 95% confidence interval, X is covariates:

$$\hat{\tau}_{\text{OLS}}(g_i) = \text{lm}(Y \sim W + X, X \in g_i). \tag{70}$$

The comparison of GATE estimates and their confidence intervals obtained using group OLS and causal forest methods is shown in the following figures:

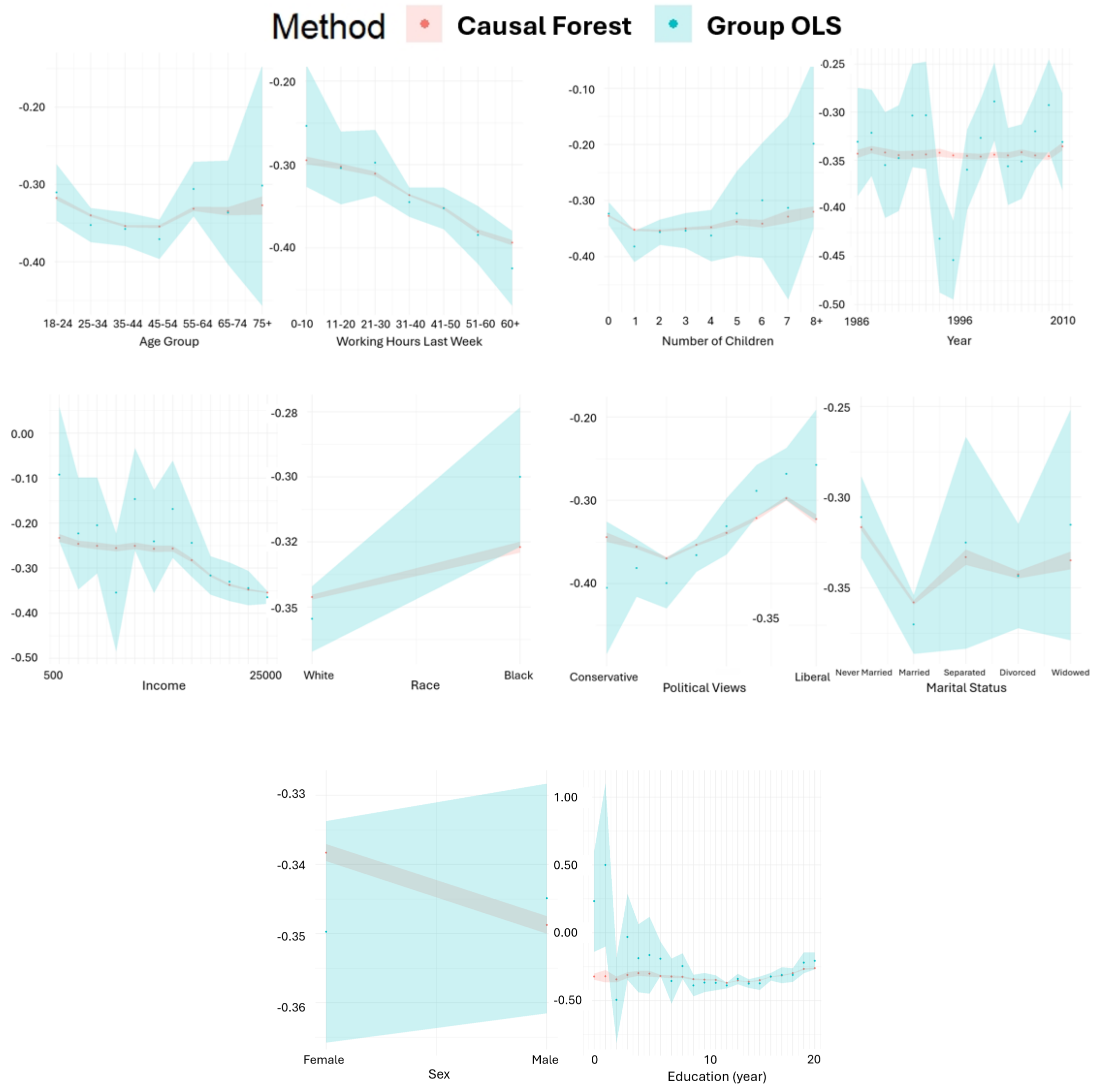


Figure 15: GATE estimation by grouping OLS and R-learner causal forest by different features.

The results demonstrate that GATE estimates obtained using group OLS have a much higher standard deviation compared to those obtained using the causal forest. This is because each OLS regression within a subgroup has a smaller sample size, whereas the causal forest leverages the entire dataset to fit the model and provides more certain and precise GATE estimates and confidence intervals.

In addition to grouping OLS, we can estimate GATE using classical models by incorporating interaction terms. By dividing features into subgroups, adding interaction terms to treatment and controlling other covariates, we can provide GATE estimation to infer HTE. Take the following OLS regression with interaction terms models the effect of age groups as example:

$$\begin{aligned}
Y = \beta_0 &+ \beta_1 \text{Treatment} + \beta_2 \text{Age Group: Under 40} + \beta_3 \text{Age Group: 40-60} \\
&+ \beta_4 \text{Treatment * Age Group: Under 40} + \beta_5 \text{Treatment * Age Group: 40-60} \\
&+ \beta_6 \text{Year} + \beta_7 \text{Weekly Hours Worked} + \beta_8 \text{Number of Children} \\
&+ \beta_9 \text{Years of Education} + \beta_{10} \text{Gender} + \beta_{11} \text{Race} + \beta_{12} \text{Income} \\
&+ \beta_{13} \text{Political Views: Conservative} + \beta_{14} \text{Political Views: Extremely Conservative} \\
&+ \beta_{15} \text{Political Views: Extremely Liberal} + \beta_{16} \text{Political Views: Moderate} \\
&+ \beta_{17} \text{Political Views: Others} + \beta_{18} \text{Political Views: Slightly Conservative} \\
&+ \beta_{19} \text{Political Views: Slightly Liberal} + \beta_{20} \text{Marital Status: Never Married} \\
&+ \beta_{21} \text{Marital Status: Married} + \beta_{22} \text{Marital Status: Separated} \\
&+ \beta_{23} \text{Marital Status: Divorced} + \beta_{24} \text{Marital Status: Widowed} + \epsilon.
\end{aligned} \tag{71}$$

The GATE for the above 60 age group is captured by β_1, for the under 40 age group it is $\beta_1 + \beta_4$, and for the 40-60 age group it is $\beta_1 + \beta_5$. Standard error of an interaction term (e.g., $\beta_1 + \beta_4$) is computed as:

$$\text{SE}(\beta_1 + \beta_4) = \sqrt{\text{Var}(\beta_1) + \text{Var}(\beta_4) + 2 \cdot \text{Cov}(\beta_1, \beta_4)}. \tag{72}$$

We apply this regression approach with interaction terms for each feature. The partition are as shown in Figure 16 first column, also in Appendix III.4. For logistic regression, we similarly estimate GATE by substituting the response variable Y with $\log\left(\frac{P(Y=1)}{P(Y=0)}\right)$ in (71). The GATE estimation CI from these three approaches are shown in Figure 16.

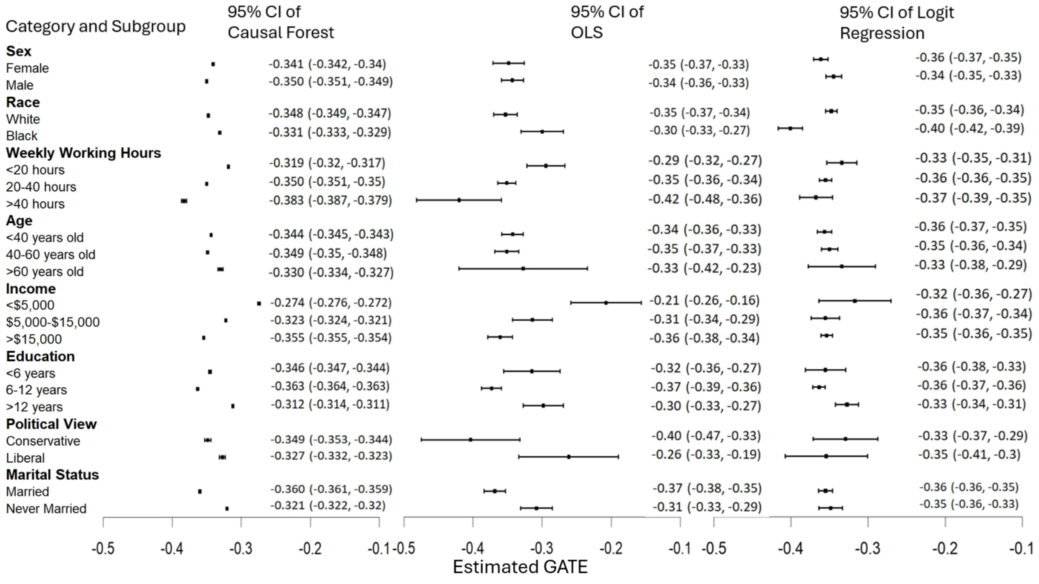


Figure 16: Comparison of causal forest vs classical OLS and logistic regression with interaction term in estimating GATE.

Figure 16 clearly demonstrates that GATE estimates obtained using the causal forest have much narrower CI compared to those from classical logistical regression and OLS with interaction terms. This further highlights the precision and certainty of causal forests estimations in HTE. Even classical model use all data to fit, its certainty and precision is still weaker than causal forest.

In terms of findings in this dataset, all models indicate similar trends in heterogeneity inference. Respondents with higher working hours and income exhibit more negative attitudes towards welfare, which aligns with the intuition that individuals who work hard may prioritize individual effort and be less inclined to support social welfare programs. Additionally, conservative respondents display more negative attitudes towards welfare compared to liberals.

We conducted statistical tests for each covariate to further validate this finding by comparing GATE estimates between the top 10th and bottom 10th percentile subgroups g_0 and g_1. The null hypothesis $H_0 : \tau(g_{\text{top}}) = \tau(g_{\text{bottom}})$ is tested, and the results are presented in Appendix III.5. The tests reveal that H_0 can be rejected for all covariates, with p-values threshold 0.0001, reinforcing the findings.

4.2 Oregon Health Insurance Experiment Dataset

Oregon Health Insurance Experiment (OHIE) was designed to assess impact of providing health insurance to low-income individuals. In this experiment, participants had the opportunity to enter a draft lottery. Winning the lottery granted them the option to enroll in the Medicaid program[24], thus gaining access to health insurance. After one year, participants were surveyed to answer whether they received all needed medical care.

The dataset[25] comprises 13,965 samples after data cleaning. It comprises 15 covariates, as detailed in Appendix III.6. Within Neyman-Rubin potential outcome framework, the outcome variable Y is binary, indicating whether a participant received all necessary medical care 12 months post-experiment. The treatment variable W represents whether the participant enrolled in Medicaid program. The instrumental variable Z is the draft lottery outcome, covariates X include demographic features.

Johnson, Cao, and Kang (2022) employed IV combined with matching techniques on the OHIE dataset. Their analysis revealed significant HTE of the Medicaid program on various health outcomes, with a stronger positive impact observed among older individuals and white participants.

Finkelstein et al. (2012) analyzed the OHIE dataset by dividing it into subgroups based on demographic characteristics. They utilized intention-to-treat (ITT) analysis and 2SLS regression to assess the HTE of Medicaid on health-related outcomes. They found overall benefits of the program were more pronounced among older participants and those with chronic conditions.

In this section, we will first explain why lottery outcome is a valid instrumental variable. Subsequently, we will apply classical 2SLS to estimate local average treatment effect (LATE) of Medicaid enrollment, comparing results with instrumental causal forest. Finally, we will evaluate the HTE by comparing the group

[24] Medicaid is a program that provides health insurance coverage.

[25] This dataset is publicly accessible at `https://www.nber.org/research/data/oregon-health-insurance-experiment-data`

local average treatment effect (GLATE) estimates from the instrumental causal forest with those derived from 2SLS with interaction terms and group 2SLS.

4.2.1 IV Check

As mentioned in section 3.3.4, a valid IV must satisfy the **monotonicity**, **relevance**, and **exclusion restriction** assumptions.

To assess monotonicity assumption, which posits there are no defiers in the population, we can examine the W/Z frequency table of the data:

		W (Medicaid Enrollment)	
		0	1
Z (Lottery Status)	0	6970	0
	1	4669	2326

Table 5: Frequency of Z and W

Since no participants who lost the lottery (Z=0) enrolled in the Medicaid program (W=1), this supports monotonicity, indicating the absence of defiers.

For the relevance assumption, we validate it by regressing enrollment (W) on lottery status (Z) while controlling for covariates X:

$$\begin{aligned}\text{Medicaid Enrollment} = {} & \beta_0 + \beta_1\text{Lottery Status} + \beta_2\text{Age} + \beta_3\text{Household Size} \\ & + \beta_4\text{Female} + \beta_5\text{First Day Signup} + \beta_6\text{Metropolitan Area} \\ & + \beta_7\text{Cigarettes per Day} + \beta_8\text{Average Weekly Working Hours} \\ & + \beta_9\text{Education Level} + \beta_{10}\text{Hispanic} + \beta_{11}\text{White} + \beta_{12}\text{Black} \\ & + \beta_{13}\text{Asian} + \beta_{14}\text{American Indian} \\ & + \beta_{15}\text{History of Disease} + \beta_{16}\text{Household Income} + \epsilon.\end{aligned} \tag{73}$$

The estimated $\hat{\beta}_1$ is 0.338 and significant at 0.001 level (full regression results in Appendix III.7). This indicates winning the lottery increases the probability of enrolling in Medicaid by 33.8%, providing strong evidence for relevance.

Regarding the exclusion restriction, because the lottery status is part of a RCT, it is independent of all other covariates. We validate this with statistical (balance) tests with null hypothesis H_0: $\mu_0 = \mu_1$ for the group win and lose lottery. Result is in Appendix III.8, it shows for nearly every covariate, H_0 cannot be rejected at $\alpha = 0.005$, reinforcing its RCT nature and exclusion restriction.

The causal relationships are represented in the following Figure 17:

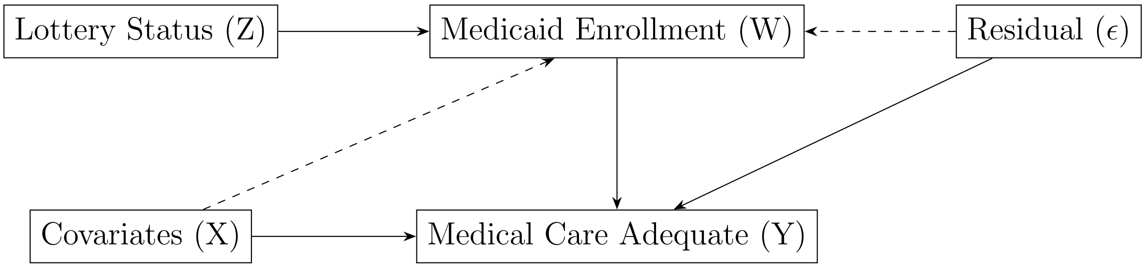


Figure 17: DAG for OHIE dataset, lottery status is an valid IV.

4.2.2 2SLS vs Instrumental Causal Forest in Estimating LATE

As discussed in section 3.3.4, when a valid IV is presented, we focus on estimating LATE, which captures treatment effect for the subgroup compliers. To begin, we apply classical 2SLS to estimate LATE. The first stage of 2SLS involves predicting Medicaid Enrollment using the following regression model fitted from (73):

$$\begin{aligned}\widehat{\text{Medicaid Enrollment}} = &\widehat{\beta}_0 + \widehat{\beta}_1\text{Lottery Status} + \widehat{\beta}_2\text{Age} + \widehat{\beta}_3\text{Household Size}\\ &+ \widehat{\beta}_4\text{Female} + \widehat{\beta}_5\text{First Day Signup} + \widehat{\beta}_6\text{Metropolitan Area}\\ &+ \widehat{\beta}_7\text{Cigarettes per Day} + \widehat{\beta}_8\text{Average Weekly Working Hours}\\ &+ \widehat{\beta}_9\text{Education Level} + \widehat{\beta}_{10}\text{Hispanic} + \widehat{\beta}_{11}\text{White} + \widehat{\beta}_{12}\text{Black}\\ &+ \widehat{\beta}_{13}\text{Asian} + \widehat{\beta}_{14}\text{American Indian}\\ &+ \widehat{\beta}_{15}\text{History of Disease} + \widehat{\beta}_{16}\text{Household Income},\end{aligned} \tag{74}$$

then plug the predicted $\widehat{\text{Medicaid Enrollment}}$ to the second stage regression:

$$\begin{aligned}\text{Medical Adequacy} = &\alpha_0 + \alpha_1\widehat{\text{Medicaid Enrollment}} + \alpha_2\text{Age} + \alpha_3\text{Household Size}\\ &+ \alpha_4\text{Female} + \alpha_5\text{First Day Signup} + \alpha_6\text{Metropolitan Area}\\ &+ \alpha_7\text{Cigarettes per Day} + \alpha_8\text{Average Weekly Working Hours}\\ &+ \alpha_9\text{Education Level} + \alpha_{10}\text{Hispanic} + \alpha_{11}\text{White} + \alpha_{12}\text{Black}\\ &+ \alpha_{13}\text{Asian} + \alpha_{14}\text{American Indian}\\ &+ \alpha_{15}\text{History of Disease} + \alpha_{16}\text{Household Income} + \epsilon.\end{aligned} \tag{75}$$

The estimated LATE ($\hat{\alpha}_1$) is 0.223, with 95% CI (0.199, 0.247). Full regression results for the second stage can be found in Appendix III.9.

Similarly to the previous GSS dataset, we chose ML models by its RMSE in predicting Y from X, as the table below we choose regression forest to predict nuisance parameters and fitted an instrumental causal forest.

Model	**RMSE**
OLS	0.48
LASSO	0.48
Ridge	0.48
XGBOOST	0.39
Regression Forest	0.33
Neural Network	0.47
SVR	0.54

Table 6: RMSE of different models

The AIPW estimator of τ can be generalized to LATE in IV setting (Athey, J. Tibshirani, and Wager 2019), we apply AIPW estimator on instrumental causal forest and the estimated LATE is 0.205, with 95% CI (0.169, 0.237). The comparison between two methods is shown in Figure 18 below, where we observe minimal differences in both the estimated values and the associated uncertainty:

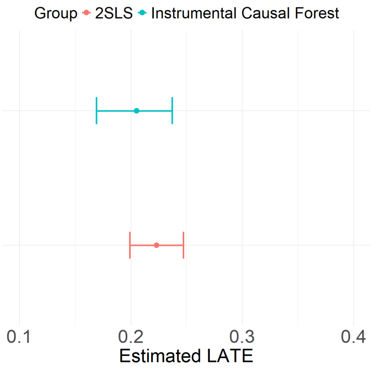


Figure 18: Comparison of 2SLS and instrumental causal forest in estimating LATE.

4.2.3 HTE Inference and GLATE Estimation Comparison

Estimated conditional local average treatment effect (CLATE) from instrumental causal forest is distributed as below:

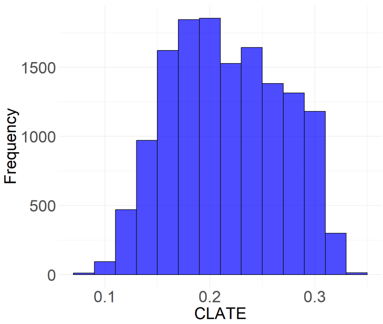


Figure 19: Estimated CLATE distribution by instrumental causal forest.

It is shown in Figure 19 that CLATE ranges from 0.1 to 0.3, indicating different subgroups of population react differently. To further validate existence of HTE, we used TOC and QINI curves as introduced in section 4.1.3.

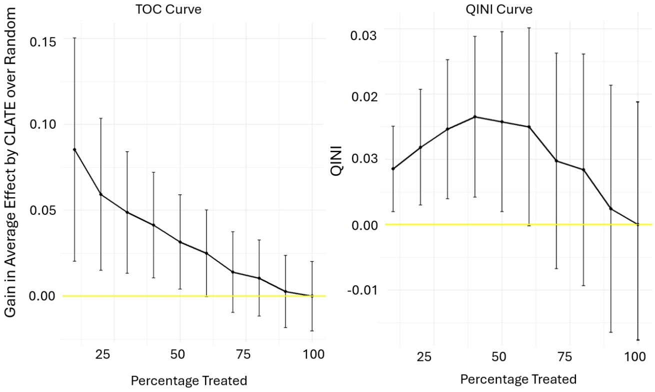


Figure 20: TOC and QINI curves in OHIE dataset, indicating significant HTE.

We can see the TOC and QINI curve have lower boundary above 0, indicating the significant HTE. In terms of $\widehat{AUTOC}$ and $\widehat{AUT\,QINI}$, the lower bounds of the 95% CI are also above 0, further providing significant evidence to existence of HTE.

	Estimation	s.e.	One-Sided 95% CI
AUTOC	0.0318	0.00005019	[0.0317, Infty]
AUT QINI	0.0102	0.00002527	[0.0102, Infty]

Table 7: AUTOC and AUT QINI Confidence Intervals

After validating existence of HTE, we compare estimation of instrumental causal forest and classical 2SLS with interaction term to treatment and grouping 2SLS. For each feature we divided data into different subgroups g_i to estimate GLATE. The splitting criteria can be found in the first column of Figure 21 or as defined in Appendix III.10.

For instrumental causal forest GLATE is computed from CLATE within each subgroup[26]:

$$\hat{\tau}_{\text{Instrumental Causal Forest}}(g_i) = \mathbb{E}[\hat{\tau}_{\text{Instrumental Causal Forest}}(X_j) \mid X_j \in g_i],$$

for classical approaches we apply grouping 2SLS and 2SLS with interaction terms to W in both stages to estimate GLATE. For grouping 2SLS, we have:

$$\text{Stage 1: } \hat{W}_{\text{2SLS}}(X) = \text{Predicted value from } \text{lm}(W \sim X + Z, X \in g_i).$$
$$\text{Stage 2: } \hat{\tau}_{\text{2SLS}}(g_i) = \text{lm}(Y \sim \hat{W}_{\text{2SLS}}(X) + X, X \in g_i).$$

For 2SLS with interaction term, we create dummy variables from subgroups and adding interaction terms to Z and W in both stages. Take age as an example, divided into under 35, 36-50 and above 50, the first stage predicted W is:

$$\begin{aligned}\widehat{\text{Medicaid Enrollment}} = \beta_0 &+ \beta_1\text{Lottery Status} + \beta_2(\text{Lottery Status} \times \text{Age 36-50})\\ &+ \beta_3(\text{Lottery Status} \times \text{Age above 50})\\ &+ \beta_4\text{Age 36-50} + \beta_5\text{Age above 50}\\ &+ \beta_6\text{Household Size} + \beta_7\text{Female} + \beta_8\text{First Day Signup}\\ &+ \beta_9\text{Metropolitan Area} + \beta_{10}\text{Cigarettes per Day}\\ &+ \beta_{11}\text{Average Weekly Working Hours} + \beta_{12}\text{Education Level}\\ &+ \beta_{13}\text{Hispanic} + \beta_{14}\text{White} + \beta_{15}\text{Black}\\ &+ \beta_{16}\text{Asian} + \beta_{17}\text{American Indian}\\ &+ \beta_{18}\text{History of Disease} + \beta_{19}\text{Household Income},\end{aligned} \tag{76}$$

then plugging this into the second stage gives:

[26] Standard error is the standard error of CLATE within each subgroup

$$
\begin{aligned}
\text{Medical Adequacy} = {} & \alpha_0 + \alpha_1 \widehat{\text{Medicaid Enrollment}} + \alpha_2(\widehat{\text{Medicaid Enrollment}} \times \text{Age 36-50}) \\
& + \alpha_3(\widehat{\text{Medicaid Enrollment}} \times \text{Age above 50}) + \alpha_4 \text{Age 36-50} \\
& + \alpha_5 \text{Age above 50} + \alpha_6 \text{Household Size} + \alpha_7 \text{Female} \\
& + \alpha_8 \text{First Day Signup} + \alpha_9 \text{Metropolitan Area} + \alpha_{10} \text{Cigarettes per Day} \\
& + \alpha_{11} \text{Average Weekly Working Hours} + \alpha_{12} \text{Education Level} \\
& + \alpha_{13} \text{Hispanic} + \alpha_{14} \text{White} + \alpha_{15} \text{Black} \\
& + \alpha_{16} \text{Asian} + \alpha_{17} \text{American Indian} \\
& + \alpha_{18} \text{History of Disease} + \alpha_{19} \text{Household Income} + \epsilon.
\end{aligned}
\tag{77}
$$

Here, α_1 represents the baseline GLATE for the age group under 36, $\alpha_1 + \alpha_2$ represents GLATE for the age group 36-50, and $\alpha_1 + \alpha_3$ represents GLATE for the age group above 50. The standard errors are computed as described in (72), and the results are shown in Figure 21:

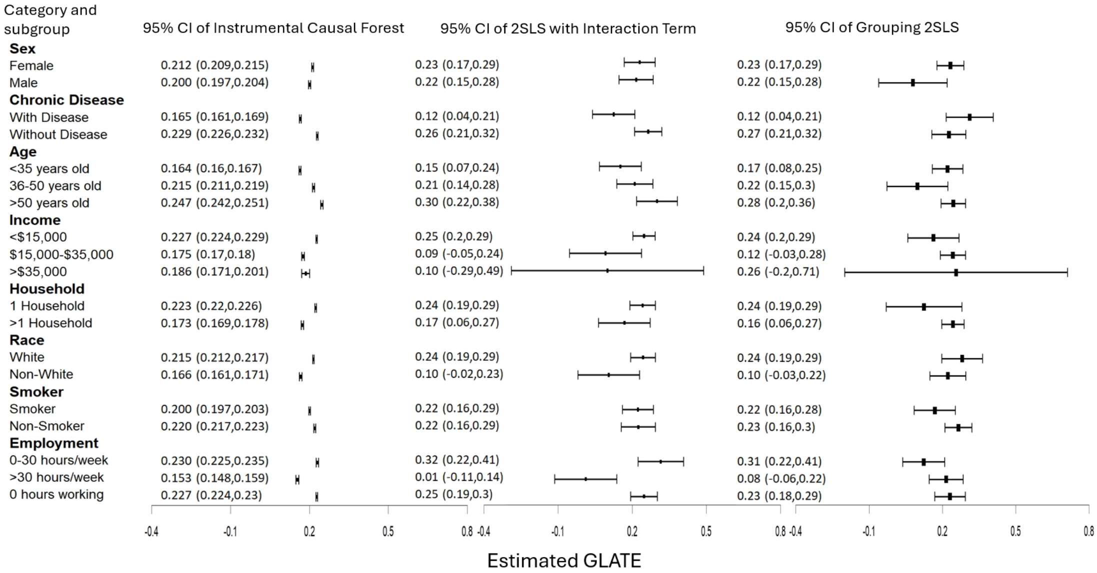


Figure 21: GLATE 95% CI estimation comparison of classical grouping 2SLS and 2SLS with interaction term vs instrumental causal forest, instrumental causal forest provides a narrower CI with more certainty and precision.

Compared with classical approaches, the instrumental causal forest provides a narrower CI for GLATE, offering more certain and precise estimates. This conclusion is in accordance with the model comparison we did in the GSS dataset of R-learner causal forest estiamtion of GATE with OLS and logistic regression.

In terms of findings in the dataset, elder individuals, people with chronic diseases, lower-income individuals, and those with fewer employment hours tend to benefit more from Medicaid, consistent with previous studies. To reinforce these findings, we compute GLATE for subgroups within the top/bottom 10th percentiles of each feature by the estiamted CLATE of instrumental causal forest and test the null hypothesis $\tau(g_{\text{top}}) = \tau(g_{\text{bottom}})$ for GLATE. The results are shown in Appendix III.11, and we found that all null hypotheses are rejected with $p = 0.0001$. This reinforce our findings.

We can further plot the distribution of CLATE across different covariates and it coincide with our finding:

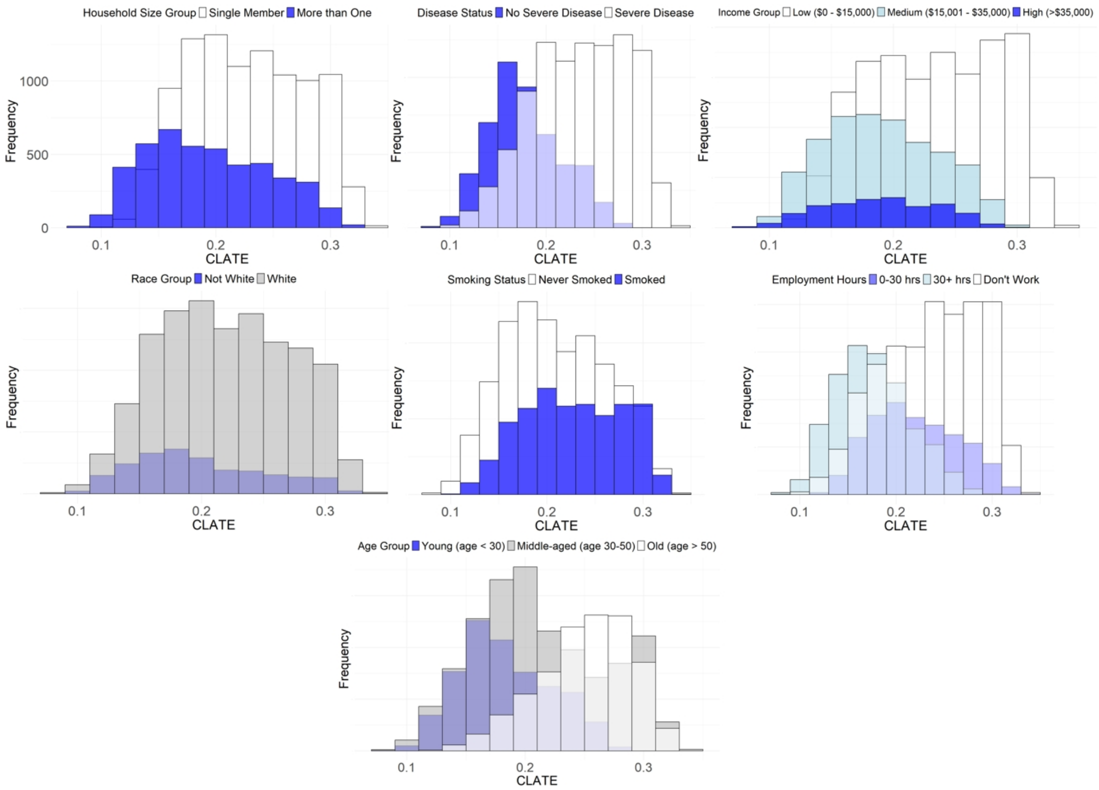


Figure 22: CLATE distribution by different covariates and divide into subgroups, coincide with our finding in previous figure.

5 Discussion and Conclusion

In this report, we explored the limitations of classical econometric tools in parameter estimation and examined the application flexible ML models in parameter estimation. Our primary focus was on the causal forest methodology, particularly its integration with DML and IV to provide confounding-robust estimation. At each stage, we employed simulations to validate the performance of these models.

We applied R-learner causal forest and instrumental causal forest models on the GSS and OHIE datasets, comparing their estimation results with those from traditional methods, including OLS, logistic regression, and 2SLS. We found causal forest's extensions provides more certain and precise estimation of HTE than classical approaches.

We conclude that the causal forest model is an effective tool for identifying and estimating HTE. Besides, our findings indicate that while these advanced ML techniques offer significant improvements, they still rely on classical models to ensure robust estimation. The DML extension of causal forest returns to classical OLS and 2SLS as its element, underscores the importance of traditional econometric theory in achieving confounding-robust estimates. This suggests that the application of ML in causal inference should complement, rather than replace, established econometric and statistical theory and techniques.

As we advance into the era of big data, the identification and estimation of HTE in real-world applications will become increasingly prevalent. By combining traditional econometric theory with modern ML models, we anticipate a flourishing of the field of ML causal inference, offering powerful tools for robust and insightful analysis.

I Appendix: Model Assumption and Mathematical Proof

I.1 Section 3.1.2 OLS Assumptions

In OLS regression we have the following model

$$Y = f(\mathbf{X}, \boldsymbol{\beta}) + \epsilon$$

Here X is the independent variable, Y is dependent variable, β is unknown parameters and ϵ is the error term. We have the following four main assumptions:

1. Errors have **zero mean**: $\mathbb{E}(\epsilon_i) = 0$
2. Errors have constant finite **variance** (homoscedastic): $\mathrm{Var}(\epsilon_i) = \sigma^2 < \infty$
3. Errors are **uncorrelated**: $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, $\forall i \neq j$
4. Independent variables are **linearly independent**

I.2 Section 3.1.2 Proof of Single OLS Potential Model Bias

If utilize all data to fit function $\mu(X, W)$ and estimate τ by running regression and get the coefficient of W, which is $\mu(X, 1) - \mu(X, 0)$, it can only work when $\mu_{(1)}(X_i) = \mu_{(0)}(X_i)+\tau$, which means the true function are consistent for treatment and control and the only difference between $\mu_{(0)}(X_i)$ and $\mu_{(1)}(X_i)$ is the treatment effect term τ. Because only in this case we can plug into (8) and have:

$$E[\mu(X_i, 1) - \mu(X_i, 0)] = E[\mu(X_i, 0) + \tau - \mu(X_i, 0)] = \tau.$$

But when $\mu_{(1)}(X_i) \neq \mu_{(0)}(X_i) + \tau$, the above equation doesn't hold and a simple OLS will have model bias and provide systematically biased estimation for τ.

I.3 Section 3.3.3 Proof of Partially Linear Model is Neyman-Orthogonal

Consider the partially linear model:

$$Y = g(X) + \tau W + \epsilon,$$

where $\hat{m}(X)$ and $\hat{e}(X)$ are the estimated nuisance parameters. We have:

$$\mathbb{E}[\hat{m}(X)] = \mathbb{E}(Y), \quad \mathbb{E}[\hat{e}(X)] = \mathbb{E}(W)$$

The moment condition $\psi(Y, W, X; \tau, \eta)$ is:

$$\psi(Y, W, X; \tau, \eta) = [Y - \hat{m}(X) - \tau(W - \hat{e}(X))] \cdot [W - \hat{e}(X)] .$$

It is a moment condition since it satisfies $\mathbb{E}[\psi(Y, W, X; \tau, \eta)] = 0$:

$$\begin{aligned}
\mathbb{E}[\psi(Y, W, X; \tau, \eta)] &= \mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))\,(W - \hat{e}(X))\right] \\
&= \mathbb{E}\left[YW - Y\hat{e}(X) - \tau W^2 + \tau W\hat{e}(X)\right] + \mathbb{E}\left[-\hat{m}(X)W + \hat{m}(X)\hat{e}(X)\right] \\
&= \mathbb{E}\left[YW - Y\hat{e}(X) - \tau W^2 + \tau W\hat{e}(X)\right] + \mathbb{E}\left[-\hat{m}(X)W + \hat{m}(X)\hat{e}(X)\right] \\
&= \mathbb{E}\left[YW - Y\hat{e}(X)\right] + \mathbb{E}\left[-\tau W^2 + \tau W\hat{e}(X)\right] + \mathbb{E}\left[-\hat{m}(X)W + \hat{m}(X)\hat{e}(X)\right] \\
&= 0
\end{aligned}$$

To proof the orthogonality, we need to proof:

$$\left.\frac{\partial \mathbb{E}[\psi(Y, W, X; \theta, \eta)]}{\partial \eta}\right|_{\eta=\eta_0} = 0.$$

Here η_0 is $\hat{e}(X)$ and $\hat{m}(X)$. Taking the partial derivative with respect to $\hat{e}(X)$:

$$\frac{\partial \psi(Y, W, X; \tau, \eta)}{\partial \hat{e}(X)} = \tau W - \tau\hat{e}(X) + \hat{m}(X) + \tau W - 2\tau\hat{e}(X),$$

$$\mathbb{E}\left[\frac{\partial \psi(Y, W, X; \tau, \eta)}{\partial \hat{e}(X)}\right] = \mathbb{E}\left[\tau W - \tau\hat{e}(X) + \hat{m}(X) + \tau W - 2\tau\hat{e}(X)\right] = 0.$$

Taking the partial derivative with respect to $\hat{m}(X)$:

$$\frac{\partial \psi(Y, W, X; \tau, \eta)}{\partial \hat{m}(X)} = -W + \hat{e}(X),$$

$$\mathbb{E}\left[\frac{\partial \psi(Y, W, X; \tau, \eta)}{\partial \hat{m}(X)}\right] = \mathbb{E}\left[\hat{e}(X) - W\right] = 0.$$

So partially linear model is Neyman-Orthogonal.

I.4 Section 3.3.3 Derivation of $\sqrt{n}$-Consistency in Double Machine Learning

Consider the problem of estimating a parameter θ using moment conditions that satisfy Neyman-orthogonality. The moment condition is given by:

$$\mathbb{E}[\psi(Y, W, X; \theta, \eta)] = 0,$$

where θ_0 is the true parameter of interest and η_0 represents the true nuisance parameters. The moment condition satisfies Neyman-orthogonality:

$$\left.\frac{\partial \mathbb{E}[\psi(Y, W, X; \theta, \eta)]}{\partial \eta}\right|_{\eta=\eta_0} = 0.$$

Define the residuals based on the nuisance parameter estimates:

$$\tilde{Y} = Y - \hat{g}(X),$$

$$\tilde{W} = W - \hat{e}(X).$$

The moment condition using residuals is given by:

$$\psi(Y, W, X; \theta, \hat{\eta}) = (Y - \hat{g}(X) - \theta(W - \hat{e}(X))) \cdot (W - \hat{e}(X)).$$

Consider the sample analog of the moment condition:

$$\hat{\psi}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, W_i, X_i; \theta, \hat{\eta}).$$

Perform a first-order Taylor expansion of $\hat{\psi}(\theta)$ around θ_0:

$$\hat{\psi}(\theta) \approx \hat{\psi}(\theta_0) + \left.\frac{\partial \hat{\psi}(\theta)}{\partial \theta}\right|_{\theta=\theta_0} (\theta - \theta_0).$$

By the Law of Large Numbers and Central Limit Theorem, we have:

$$\hat{\psi}(\theta_0) \approx \mathbb{E}[\psi(Y, W, X; \theta_0, \hat{\eta})] + O_p(n^{-1/2}).$$

Because of Neyman-orthogonality, the first order term of Tylar expansion is 0, we have:

$$\mathbb{E}[\psi(Y, W, X; \theta_0, \hat{\eta})] \approx \mathbb{E}[\psi(Y, W, X; \theta_0, \eta_0)] + O_p(\|\hat{\eta} - \eta_0\|^2).$$

Since $\mathbb{E}[\psi(Y, W, X; \theta_0, \eta_0)] = 0$ and $\|\hat{\eta} - \eta_0\| = O_p(n^{-1/4})$, we get:

$$\mathbb{E}[\psi(Y, W, X; \theta_0, \hat{\eta})] = O_p(n^{-1/2}).$$

Combining these results, we have:

$$\hat{\psi}(\theta_0) = O_p(n^{-1/2}).$$

The derivative term:

$$\left.\frac{\partial \hat{\psi}(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \approx \mathbb{E}\left[\frac{\partial \psi(Y, W, X; \theta_0, \eta_0)}{\partial \theta}\right].$$

Set $\hat{\psi}(\theta) = 0$ and solve for θ:

$$0 = \hat{\psi}(\theta_0) + \left.\frac{\partial \hat{\psi}(\theta)}{\partial \theta}\right|_{\theta=\theta_0} (\hat{\theta} - \theta_0).$$

Rearrange to find:

$$\hat{\theta} - \theta_0 = -\left(\left.\frac{\partial \hat{\psi}(\theta)}{\partial \theta}\right|_{\theta=\theta_0}\right)^{-1} \hat{\psi}(\theta_0).$$

Given $\hat{\psi}(\theta_0) = O_p(n^{-1/2})$ and $\left.\frac{\partial \hat{\psi}(\theta)}{\partial \theta}\right|_{\theta=\theta_0} \neq 0$, we have:

$$\hat{\theta} - \theta_0 = O_p(n^{-1/2}).$$

I.5 Section 3.3.3 Proof of $\hat{\tau}_{\mathbf{IPW}}$ Property

Apart from estimating $\mu_{(w)}(x)$ in (8) and estimate CATE, we can also estimate CATE by Inversed Propensity Weighted (IPW) method as follows, suppose we know the true propensity score function to be $e(x_i)$, the **oracle IPW estimator** of ATE $\hat{\tau}^*_{IPW}$ is:

$$\hat{\tau}^*_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right). \tag{78}$$

We can easily proof that oracle IPW estimator of ATE $\hat{\tau}^*_{IPW}$ is unbiased to the true τ:

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}^*_{IPW}\right] &= \mathbb{E}\left[\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right] \\
&= \mathbb{E}\left[\frac{W_iY_i(1)}{e(X_i)} - \frac{(1-W_i)Y_i(0)}{1-e(X_i)}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{W_iY_i(1)}{e(X_i)} \mid e(X_i)\right] - \mathbb{E}\left[\frac{(1-W_i)Y_i(0)}{1-e(X_i)} \mid e(X_i)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{W_iY_i(1)}{e(X_i)} \mid e(X_i)\right] - \mathbb{E}\left[\frac{(1-W_i)Y_i(0)}{1-e(X_i)} \mid e(X_i)\right]\right] \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \tau.
\end{aligned} \tag{79}$$

But in reality since we don't know the true $e(x)$, we need to estimate by $\hat{e}(x)$, which introduces the **IPW estimator** $\hat{\tau}_{IPW}$:

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right). \tag{80}$$

It turns out to have non-negligible regularization bias in (80) as proofed below. It doesn't satisfy the $\sqrt{N}$ **- consistent property** or orthogonality, which is:

$$\sqrt{N}\left(\hat{\tau} - \hat{\tau}^*\right) \xrightarrow{p} 0.\text{(N is the sample size)} \tag{81}$$

I.6 Section 3.3.3 IPW has Non-negligible Bias

When we are estimating τ by $\hat{\tau}_{\text{IPW}}$, it can be decomposed to the oracle IPW estimator and the model bias due to the incorrect estimation of $\hat{\tau}_{\text{IPW}}$ to the $\hat{\tau}^*_{\text{IPW}}$:

$$\hat{\tau}_{\text{IPW}} = \underbrace{\hat{\tau}^*_{\text{IPW}}}_{\text{a good estimator}} + \underbrace{(\hat{\tau}_{\text{IPW}} - \hat{\tau}^*_{\text{IPW}})}_{\text{due to errors in } \hat{e}(\cdot)}$$

As we defined in (78) and prove in (79), oracle estimator of $\hat{\tau}^*_{\text{IPW}}$ is the unbiased estimator of τ, according to central limit theorem, since it has finite variance and each sample is iid, it will converge to the true τ at the rate of $\sqrt{n}$. (Casella and Berger 2002) We have:

$$\sqrt{n}(\hat{\tau}^*_{\text{IPW}} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \text{Var}\left[\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right]\right)$$

For the $\hat{\tau}_{\text{IPW}} - \hat{\tau}^*_{\text{IPW}}$ error term, we need to see if it is lower order to $1/\sqrt{n}$ so can be neglected, we can apply the Cauchy-Schwarz inequality:

$$\begin{aligned}
\hat{\tau}_{\text{IPW}} - \hat{\tau}^*_{\text{IPW}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\hat{e}(X_i)} - \frac{1-W_i}{1-\hat{e}(X_i)} - \frac{W_i}{e(X_i)} + \frac{1-W_i}{1-e(X_i)}\right)Y_i \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\hat{e}(X_i)} - \frac{1-W_i}{1-\hat{e}(X_i)} - \frac{W_i}{e(X_i)} + \frac{1-W_i}{1-e(X_i)}\right)^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2} \\
&\approx \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{e}(X_i) - e(X_i)\right)^2}
\end{aligned}$$

So we can see that the error term is on the same scale as the RMSE of $\hat{e}(x)$. Since if in the parametric problem, we have RMSE $\sim \frac{1}{\sqrt{n}}$, if it is non-parametric problem, we have RMSE $>> \frac{1}{\sqrt{n}}$, so we can never achieve RMSE $<< \frac{1}{\sqrt{n}}$ and the error term is not lower order to $1/\sqrt{n}$ and can not be neglected.

I.7 Section 3.3.3 Double Robustness of AIPW

AIPW estimator $\hat{\tau}_{AIPW}$ is "Double Robustness" since AIPW is unbiased when any one of $\hat{\mu}_{w(x)}(x)$, $\hat{e}(x)$ is consistent to the true function, the proof is as follows:

When $\hat{\mu}_{w(x)}(x)$ is consistent to $\mu_{w(x)}(x)$, we can change the form of (51) to:

$$\begin{aligned}
\hat{\tau}_{AIPW} = &\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right)}_{\text{a consistent unbiased treatment effect estimator}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{1-W_i}{1-\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\right)}_{\approx\text{mean-zero noise}}.
\end{aligned}$$

Since $\hat{\mu}_{w(x)}(x)$ is consistent to $\mu_{w(x)}(x)$, the first part is the unbiased estimator of τ as we show in (8), the second part have $Y_i - \hat{\mu}_{(1)}(X_i)$ to have mean 0, so (51) is unbiased to τ.

When $\hat{e}(x)$ is consistent to the true propensity score $e(x)$, we can change the form of (51) to:

$$\begin{aligned}
\hat{\tau}_{AIPW} = &\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)}_{\text{the oracle IPW estimator}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i)\left(1 - \frac{W_i}{\hat{e}(X_i)}\right) - \hat{\mu}_{(0)}(X_i)\left(1 - \frac{1-W_i}{1-\hat{e}(X_i)}\right)\right)}_{\approx\text{mean-zero noise}}.
\end{aligned}$$

The first part is the oracle IPW estimator of τ and it is unbiased estimator of τ as we proof in (79). The second part has $(1 - \frac{W_i}{\hat{e}(X_i)})$ to be zero mean noise, so in this case (51) is still unbiased to τ.

So AIPW estimator provides a "Double Robustness" estimator of τ, which means we only need to ensure any one of $\hat{e}(x)$ or $\hat{\mu}_{(w)}(x)$ to be consistent to the true model and $\hat{\tau}_{AIPW}$ will be consistent to τ.

I.8 Section 3.3.3 $\sqrt{n}-$ Consistent Property of $\hat{\tau}_{AIPW}$

Here we use the **cross fitting** approach (Wager 2023). The idea of cross fitting is train the model by half of data estimate in the other half. Here at first we split the data randomly into two halves: $\mathcal{I}_1$ and $\mathcal{I}_2$, then $\hat{\tau}_{AIPW}$ can be written as:

$$\begin{aligned}\hat{\tau}_{AIPW} &= \frac{|\mathcal{I}_1|}{n}\hat{\tau}_{\mathcal{I}_1} + \frac{|\mathcal{I}_2|}{n}\hat{\tau}_{\mathcal{I}_2}, \\ \hat{\tau}_{\mathcal{I}_1} &= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i)\right) \\ &+ \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(W_i\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i)}{1-\hat{e}^{\mathcal{I}_2}(X_i)}\right),\end{aligned} \tag{82}$$

Apart from the 3.2.1 overlap assumption, we also need the following two assumptions:

- **Consistency:** All machine learning adjustments are sup-norm consistent, which means they will converge to the true function.

$$\sup_{x\in\mathcal{X}}\left|\hat{\mu}^{\mathcal{I}_2}_{(w)}(x) - \mu_{(w)}(x)\right|, \quad \sup_{x\in\mathcal{X}}\left|\hat{e}^{\mathcal{I}_2}(x) - e(x)\right| \to_p 0.$$

- **Risk decay:** The product of the errors for the outcome and propensity models decays as

$$\mathbb{E}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(w)}(X_i) - \mu_{(w)}(X_i)\right)^2\right]\mathbb{E}\left[\left(\hat{e}^{\mathcal{I}_2}(X_i) - e(X_i)\right)^2\right] = o\left(\frac{1}{n}\right)$$

  where the randomness above is taken over both the training of $\hat{\mu}_{(w)}$ and $\hat{e}$ and the test example X. A simple way to satisfy this condition is to have all regression adjustments be $o(n^{-1/4})$ consistent in root-mean squared error (RMSE). For random forest it can achieve this $o(n^{-1/4})$ consistent rate as proven by Scornet, Biau, and Vert (2015)

What we need to prove is:

$$\sqrt{n}\left(\hat{\tau}_{AIPW} - \hat{\tau}^*\right) \xrightarrow{p} 0$$

We can write the oracle estimator of τ to be:

$$\hat{\tau}^* = \frac{|\mathcal{I}_1|}{n}\hat{\tau}^{\mathcal{I}_1,*} + \frac{|\mathcal{I}_2|}{n}\hat{\tau}^{\mathcal{I}_2,*}$$

Then we can decompose $\hat{\tau}^{\mathcal{I}_1}$ as:

$$\hat{\tau}_{\mathcal{I}_1} = \hat{\mu}^{\mathcal{I}_1}_{(1)} - \hat{\mu}^{\mathcal{I}_1}_{(0)}$$

Here for $\hat{\mu}_{(1)}^{\mathcal{I}_1}$ (similar for $\hat{\mu}_{(0)}^{\mathcal{I}_1}$, $\hat{\mu}_{(1)}^{\mathcal{I}_1,*}$, $\hat{\mu}_{(0)}^{\mathcal{I}_1,*}$) can be written as:

$$\hat{\mu}_{(1)}^{\mathcal{I}_1} = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} \right)$$

So we just need to proof:

$$\sqrt{n} \left(\hat{\mu}_{(1)}^{\mathcal{I}_1} - \hat{\mu}_{(1)}^{\mathcal{I}_1,*} \right) \xrightarrow{p} 0$$

We can decompose the term as follows:

$$\begin{aligned}
\hat{\mu}_{(1)}^{\mathcal{I}_1} - \hat{\mu}_{(1)}^{\mathcal{I}_1,*} &= \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} - \mu_{(1)}(X_i) - W_i \frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} \right) \\
&= \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(1 - \frac{W_i}{e(X_i)} \right) \right) \\
&+ \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} W_i \left(Y_i - \mu_{(1)}(X_i) \right) \left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)} \right) \\
&- \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} W_i \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)} \right)
\end{aligned}$$

For the first term, given the double machine learning approach, we have:

$$\begin{aligned}
&\mathbb{E} \left[\left(\frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(1 - \frac{W_i}{e(X_i)} \right) \right)^2 \right] \\
&= \mathbb{E} \left[\mathbb{E} \left[\left(\frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(1 - \frac{W_i}{e(X_i)} \right) \right)^2 \mid \mathcal{I}_2 \right] \right] \\
&= \mathbb{E} \left[\mathrm{Var} \left(\frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(1 - \frac{W_i}{e(X_i)} \right) \mid \mathcal{I}_2 \right) \right] \\
&= \frac{1}{|\mathcal{I}_1|} \mathbb{E} \left[\mathrm{Var} \left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(1 - \frac{W_i}{e(X_i)} \right) \mid \mathcal{I}_2 \right) \right] \\
&= \frac{1}{|\mathcal{I}_1|} \mathbb{E} \left[\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right)^2 \left(\frac{1}{e(X_i)} - 1 \right)^2 \mid \mathcal{I}_2 \right] \\
&\leq \frac{1}{\eta |\mathcal{I}_1|} \mathbb{E} \left[\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right)^2 \mid \mathcal{I}_2 \right] = o_P \left(\frac{1}{n} \right)
\end{aligned}$$

Which means the first term satisfies the condition of $o_P \left(\frac{1}{\sqrt{n}} \right)$ converge rate. And for the second term it is similar because of overlap, the third term we can also use Cauchy-Schwarz inequality:

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_1|}\sum_{i:i\in\mathcal{I}_1,W_i=1}\left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)\right)\\
&\leq\sqrt{\frac{1}{|\mathcal{I}_1|}\sum_{i:i\in\mathcal{I}_1,W_i=1}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\mu_{(1)}(X_i)\right)^2}\\
&\quad\times\sqrt{\frac{1}{|\mathcal{I}_1|}\sum_{i:i\in\mathcal{I}_1,W_i=1}\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)^2}\\
&=o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

By the cross fitting approach, we can prove that for a given $o(n^{-1/4})$ consistent machine learning model to approximate $\hat{\mu}_{w(x)}(x)$ and $\hat{e}(x)$, we can provide $\sqrt{n}$-consistent ATE estimator by AIPW.

For the Robinson's transformation we can think intuitively it is a semiparametric approach so should have better efficiency than non-parametric method, so it is also $\sqrt{n}$-consistent to the oracle estimator.

I.9 Section 3.3.4 Derivation of the 2SLS Estimator

Consider the structural equation:

$$Y=\beta_0+\beta_1W+\varepsilon$$

where W is an endogenous variable correlated with the error term ε.

We have an instrumental variable Z which is correlated with W but uncorrelated with ε.

First, regress W on Z to obtain the predicted values of W:

$$W=\pi_0+\pi_1Z+\eta$$

Using Ordinary Least Squares (OLS), the coefficients π_0 and π_1 are estimated as:

$$\hat{\pi}_1=\frac{\sum_{i=1}^n(Z_i-\bar{Z})(W_i-\bar{W})}{\sum_{i=1}^n(Z_i-\bar{Z})^2}$$

$$\hat{\pi}_0=\bar{W}-\hat{\pi}_1\bar{Z}$$

The predicted values of W are:

$$\hat{W}=\hat{\pi}_0+\hat{\pi}_1Z$$

Next, use the predicted values $\hat{W}$ in place of W in the original regression:

$$Y=\beta_0+\beta_1\hat{W}+u$$

Using OLS again, the coefficient β_1 is estimated as:

$$\hat{\beta}_{2\text{SLS}}=\frac{\sum_{i=1}^n(\hat{W}_i-\bar{\hat{W}})(Y_i-\bar{Y})}{\sum_{i=1}^n(\hat{W}_i-\bar{\hat{W}})^2}$$

Since $\hat{W} = \hat{\pi}_0 + \hat{\pi}_1 Z$, we substitute this into the above expression:

$$\hat{\beta}_{\text{2SLS}} = \frac{\sum_{i=1}^{n}(\hat{\pi}_0 + \hat{\pi}_1 Z_i - \bar{\hat{W}})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(\hat{\pi}_0 + \hat{\pi}_1 Z_i - \bar{\hat{W}})^2}$$

Simplifying this expression, we note that $\hat{\pi}_0$ and $\bar{\hat{W}}$ are constants and can be factored out, we substitute $\bar{\hat{W}}$ by $\bar{\hat{W}} = \hat{\pi}_0 + \hat{\pi}_1 \bar{Z}$:

$$\hat{\beta}_{\text{2SLS}} = \frac{\hat{\pi}_1 \sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\hat{\pi}_1^2 \sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

Canceling out $\hat{\pi}_1$:

$$\hat{\beta}_{\text{2SLS}} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\hat{\pi}_1 \sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

Since $\hat{\pi}_1 = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(W_i - \bar{W})}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$, we substitute this back in:

$$\hat{\beta}_{\text{2SLS}} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(W_i - \bar{W})}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2} \sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

Simplifying, we get:

$$\hat{\beta}_{\text{2SLS}} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(W_i - \bar{W})}$$

In matrix form, if Z, W and Y are column vector, we have the 2SLS estimator:

$$\hat{\beta}_{\text{2SLS}} = \frac{Z^T Y}{Z^T W}$$

I.10 Section 3.3.4 Derivation of 2SLS Unbiased Estimator

Let's denote $\hat{X} = Z(Z'Z)^{-1}Z'X$. The 2SLS estimator of β is given by the second stage OLS is:

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y.$$

Substituting $\hat{X}$ into the estimator:

$$\hat{\beta}_{2SLS} = \left[(Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X)\right]^{-1}(Z(Z'Z)^{-1}Z'X)'y.$$

Simplify the expression inside the inverse:

$$\hat{\beta}_{2SLS} = \left[X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X\right]^{-1}X'Z(Z'Z)^{-1}Z'y,$$

$$\hat{\beta}_{2SLS} = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1}X'Z(Z'Z)^{-1}Z'y.$$

Now, using the model $y = X\beta + \epsilon$, multiply Z' on both side of equation:

$$Z'y = Z'X\beta + Z'\epsilon.$$

Substitute this into the 2SLS estimator:

$$\hat{\beta}_{2SLS} = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}(Z'X\beta + Z'\epsilon),$$

$$\hat{\beta}_{2SLS} = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'X\beta + \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'\epsilon.$$

Notice that the first term simplifies to β:

$$\hat{\beta}_{2SLS} = \beta + \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'\epsilon.$$

To show that the estimator is unbiased, we need to show that the expectation of the second term is zero:

$$E\left[\left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'\epsilon\right] = 0.$$

Since $E[\epsilon|Z] = 0$ (by exogeneous assumption of IV, Z is uncorrelated with ϵ),

$$E\left[Z'\epsilon\right] = 0.$$

Hence,

$$E\left[\left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'\epsilon\right] = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}E\left[Z'\epsilon\right] = 0.$$

Thus,

$$E[\hat{\beta}_{2SLS}] = \beta.$$

Therefore, the 2SLS estimator is unbiased.

I.11 Section 3.3.4 Proof of IV Neyman-Orthogonal

The moment condition is given by:

$$\psi(Y, W, Z, X; \beta, \eta) = (Y - \hat{m}(X) - \beta(W - \hat{e}(X)))(Z - \hat{h}(X)).$$

Proof of Valid moment condition

A valid moment condition satisfies the requirement that the expectation of the moment condition is zero when evaluated at the true parameter values:

$$\mathbb{E}[\psi(Y, W, Z, X; \tau, \eta_0)] = 0.$$

Expanding the moment condition, we have:

$$\psi(Y, W, Z, X; \tau, \eta) = (Y - \hat{m}(X) - \tau(W - \hat{e}(X)))(Z - \hat{h}(X)).$$

We need to show that:

$$\mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))(Z - \hat{h}(X))\right] = 0.$$

Breaking it down:

$$\mathbb{E}\left[(Y - \hat{m}(X))(Z - \hat{h}(X))\right] - \tau\mathbb{E}\left[(W - \hat{e}(X))(Z - \hat{h}(X))\right].$$

Given the exogeneous assumption of Z to W, we know it should be uncorrelated with the error term, so The residuals of Z and W should be orthogonal, so we have $\mathbb{E}[(W - \hat{e}(X))(Z - \hat{h}(X))] = 0$. Also by the exclusion assumption Z is uncorrelated with the error term of Y, so $\mathbb{E}\left[(Y - \hat{m}(X))(Z - \hat{h}(X))\right] = 0$.

Since both terms are zero, the moment condition is valid:

$$\mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))(Z - \hat{h}(X))\right] = 0.$$

Neyman Orthogonality

To prove Neyman orthogonality, we need to show that the partial derivative of the moment condition with respect to the nuisance parameters η is zero at the true values of these parameters η_0:

$$\frac{\partial}{\partial \eta} \mathbb{E}[\psi(Y, W, Z, X; \tau, \eta)]\bigg|_{\eta = \eta_0} = 0$$

Expanding the expectation of the moment condition, we have:

$$\mathbb{E}[\psi(Y, W, Z, X; \tau, \eta)] = \mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))\,(Z - \hat{h}(X))\right]$$

Taking the partial derivative with respect to η and evaluating at $\eta = \eta_0$:

$$\frac{\partial}{\partial \eta} \mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))\,(Z - \hat{h}(X))\right]\bigg|_{\eta = \eta_0}$$

Consider each component of the partial derivative:

$$\frac{\partial}{\partial \hat{m}(X)} \mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))\,(Z - \hat{h}(X))\right]$$

Since $\mathbb{E}[Y|X] = \hat{m}(X)$, the partial derivative with respect to $\hat{m}(X)$ is:

$$-\mathbb{E}[(Z - \hat{h}(X))] = 0$$

Similarly,

$$\frac{\partial}{\partial \hat{e}(X)} \mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))\,(Z - \hat{h}(X))\right]$$

Since $\mathbb{E}[W|X] = \hat{e}(X)$, the partial derivative with respect to $\hat{e}(X)$ is:

$$-\tau\mathbb{E}[(Z - \hat{h}(X))] = 0$$

Lastly,

$$\frac{\partial}{\partial \hat{h}(X)} \mathbb{E}\left[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))\,(Z - \hat{h}(X))\right]$$

Since $\mathbb{E}[Z|X] = \hat{h}(X)$, the partial derivative with respect to $\hat{h}(X)$ is:

$$\mathbb{E}[(Y - \hat{m}(X) - \tau(W - \hat{e}(X)))] = 0$$

Thus, the moment condition is Neyman orthogonal.

I.12 Section 4.1.3 Variance of AUTOC/AUT QINI

Let $q \in Q := \{q_1, \ldots, q_p\}$ denote a grid of quantiles. Let $\alpha = (\alpha_1, \ldots, \alpha_p)$ denote the p-dimensional vector. Following the Neyman-Orthogonality DML framework, we have our estimated parameter to be:

$$\hat{\theta}(q; \nu) = \mathbb{E}_n \left[\psi(W; \nu) \right].$$

If we denote

$$\mathbb{I}(q) := 1\{\tau_*(X) \geq \mu(\tau_*, q)\},$$

and

$$\psi_\ell(W) = (Y(\eta_0) - \theta_0) \left(\frac{\mathbb{I}(q_\ell)}{\pi_0(q_\ell)} - 1 \right) - \alpha_\ell,$$

with

$$\theta_0 = \mathbb{E}[Y(\eta_0)] \quad \text{and} \quad \pi_0(q) = \mathbb{E}[\mathbb{I}(q)].$$

We have

$$\widehat{V} = \mathbb{E}_n \hat{\psi}(W)^2, \hat{\psi}(W) = \sum_{\ell=1}^{p} \hat{\psi}_\ell(W)(q_{\ell+1} - q_\ell).$$

Here $\hat{V}$ is the estiamted variance of $\widehat{\text{AUTOC}}$, and for $\widehat{\text{AUT QINI}}$ we can estimate the variance similarly.

II Appendix: Simulation Setting

II.1 Section 3.1.2 Simulations in Linear/Non-linear Settings

Simulation 1: Linear Setting: We generate data from the following distribution, denote noise to be ϵ, d (number of features) = 20

$$X_i \sim \mathcal{N}(0,1)$$

$$\epsilon \sim \mathcal{N}(0,4)$$

$$\mu_0(X) = X_1 + X_2 + \epsilon$$

$$\mu_1(X) = X_1 + X_3 + 0.05 + \epsilon$$

$$e(X) = \frac{1}{1 + e^{(-X_1)}}$$

Simulation 2: Non-Linear Setting: We generate data from the same distribution but change $\mu_{(w)}(x)$ to be quadratic.

$$X_i \sim \mathcal{N}(0,1)$$

$$\epsilon \sim \mathcal{N}(0,4)$$

$$\mu_0(X) = 4 \times max(X_1, 0) + \frac{X_2^2}{2} + X_4^2 + \epsilon$$

$$\mu_1(X) = 4 \times max(X_1, 0) + \frac{X_2^2}{2} + X_3^2 + 0.05 + \epsilon$$

$$e(X) = \frac{1 + sin(X_1)}{2}$$

II.2 Section 3.2.1 Adaptive VS Honest Estimation

We use the following simulation to generate data and compare the performance of adaptive causal forest and honest causal forest in constant treatment effect and non-linear treatment effect cases, we generate 2000 samples and 10 features (n=2000, d=10):

$$\epsilon \sim \mathcal{N}(0,1)$$
$$\mathrm{e}(x) = 0.3$$
$$X_i \sim \mathcal{N}(0,1)$$
$$W \sim \mathrm{B}(\mathrm{e}(x))$$

$$\tau(X) = \frac{1}{1 + \exp(-X_3)} \quad \text{(Non-linear setting)}$$

$$\tau(X) = 2 \quad \text{(Constant setting)}$$

$$Y = \max(X_1 + X_2, 0) + \frac{1}{1 + \exp(-X_1)} + 2X_3 + W\tau + \epsilon$$

The performance is as Figure 23 and Table 8 below, we can see honest causal forest performs much better than adaptive causal forest in both linear and non-linear settings, which validate its reduction in over-fitting and reduction in bias by a separation of tree construction and estimation samples.

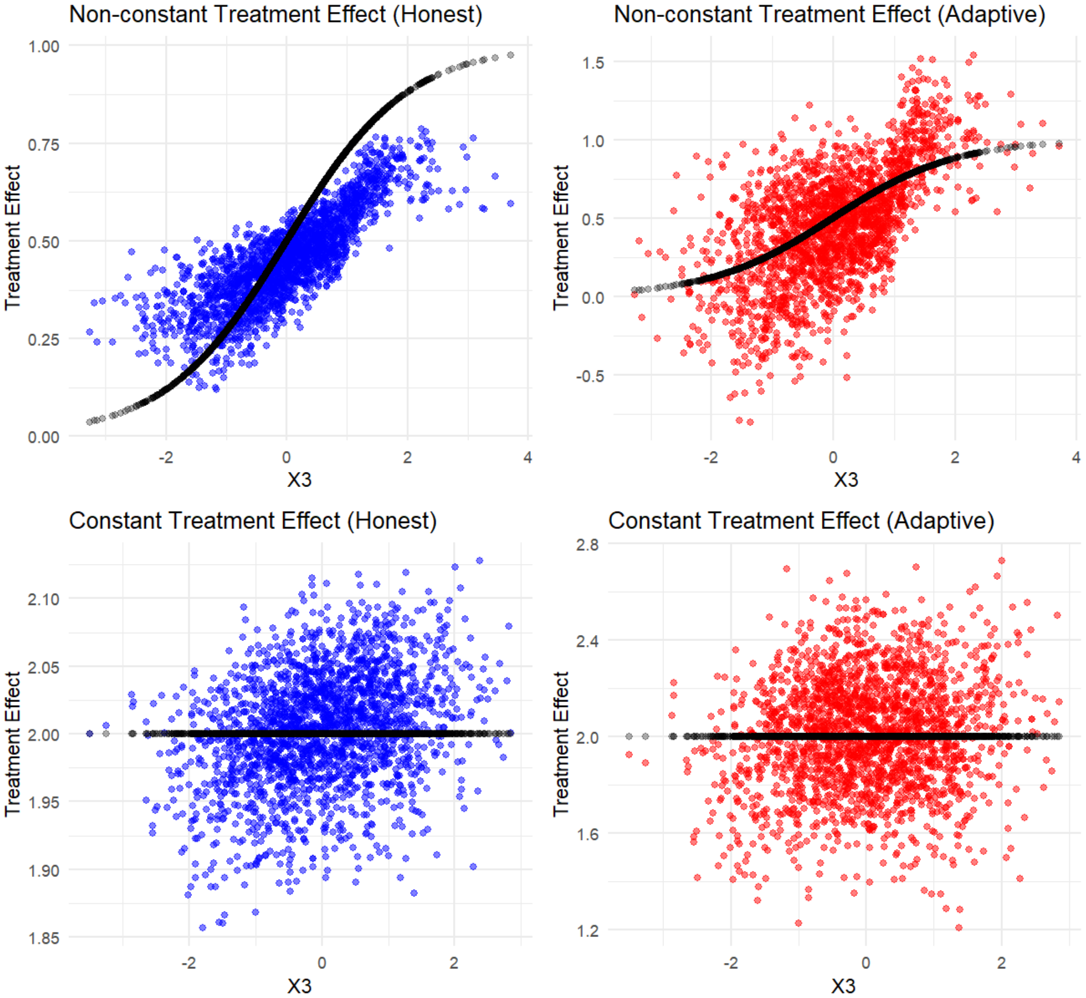


Figure 23: Comparison of Adaptive Causal Forest and Honest Causal Forest

Method	RMSE
Honest (Non-Constant)	0.14
Adaptive (Non-Constant)	0.29
Honest (Constant)	0.04
Adaptive (Constant)	0.24

Table 8: RMSE for Different Methods and Treatment Effects

II.3 Section 3.2.2 K-NN vs Causal Forest Simulation Setting

Wager and Athey (2018) used the following simulation and we replicate it. The sample size n is 10000 and dimension d is 6 and 20. τ is set to related to X_1 and X_2 to make it visualized better in a two dimensional case.

$$\epsilon \sim \mathcal{N}(0,1)$$

$$e(X_1) = \tfrac{1}{4}\left(1 + \beta_{2,4}(X_1)\right)$$

$$X_i \sim \mathcal{N}(0,1)$$

$$W \sim \text{Bernoulli}(e(X_1))$$

$$S(x) = 1 + \tfrac{1}{1+\exp(-20(x-\frac{1}{3}))}$$

$$\tau(X) = S(X_1)S(X_2)$$

$$Y = \tau(X)W + 2X_1 - 1 + \epsilon$$

II.4 Section 3.3.3 Simulation Setting in DML with Causal Forest

We use the following simulation, X_3 is the confounder. We set the treatment effect $\tau(x)$ to be constant and varies with X_3

$$n = 4000, \; p = 10$$

$$\tau(x) = \tfrac{1}{1+e^{x_3}} \text{ (HTE case)}$$

$$\tau(x) = 1 \text{ (constant treatment effect case)}$$

$$e(x) = \tfrac{1}{1+e^{-x_3}} \text{ (confounding case)}$$

$$e(x) = 0.3 \text{ (RCT case)}$$

$$Y = 2 \times \max\{X_1 + X_2 + X_3, 0\} + \tau \times W + \epsilon$$

$$\epsilon \sim \mathcal{N}(0,1)$$

II.5 Section 3.3.3 Violation of Overlap

If we use the same non-linear simulation 2 in II.1, where the propensity score is $e(X) = \frac{1+sin(X_1)}{2}$, it violates the overlap assumption since it can have propensity score very close to 0 or 1. If we simulate data from this setting and estimate the ATE by simple method in (8), IPW method in (80) and AIPW in (51), it provides the result as figure 24 below:

We can see that the true treatment effect is 0.05, the AIPW and IPW can have extreme estimation to the ATE since the data doesn't satisfy the overlap assumption 3.2.1.

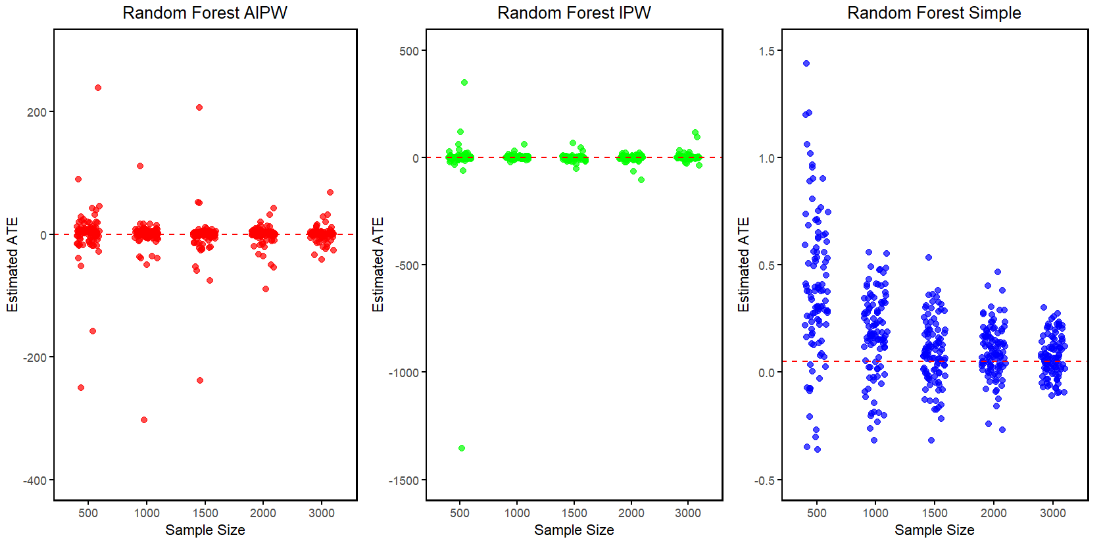


Figure 24: Estimation without overlapping assumption

II.6 Section 3.3.3 AIPW Asymptotic Property Simulation

To test the performance of these three ATE estimator $\hat{\tau}_{Simple}$ (8), $\hat{\tau}_{IPW}$ (80) and $\hat{\tau}_{AIPW}$ (51). We generate data from the simulation in II.1 the non linear setting. In order to satisfy the overlap assumption 3.2.1, we change the propensity score to $e(x) = 0.3 + 0.4 \times \Phi^{-1}(X_1)$, which make $e(x)$ to be between 0.3 and 0.7 and don't provide extreme values. To be consistent and comparable, we all use non-parametric approach random forest to estimate $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$, result is as in the following Figure 25:

We can see the mean of AIPW estimator $\hat{\tau}_{AIPW}$ approximates the true ATE slightly faster than simple estimation $\hat{\tau}_{Simple}$ and much better than IPW estimator $\hat{\tau}_{IPW}$, which is due to the double robustness which can fix the misspecification of $\hat{\mu}_{(w)}(x)$ or $\hat{e}(x)$ and DML technique's orthogonality property ensures the $\frac{1}{\sqrt{n}}$ consistency which makes it converge to the true value faster.

II.7 Section 3.3.4 Simulation Setting for IV Forest in Endogeneity

We set U to be unobserved confounder and X are the observed covariates. Sample size n = 5000 and dimension d = 6.

$$n = 5000$$
$$d = 6$$

$$Z \sim \text{Bernoulli}(0.5)$$
$$U \sim \text{Bernoulli}(0.5)$$
$$W = U \cdot Z$$

$$\tau(x) = \frac{1}{1+e^{-X_1}}$$

$$\epsilon \sim N(0, 1)$$

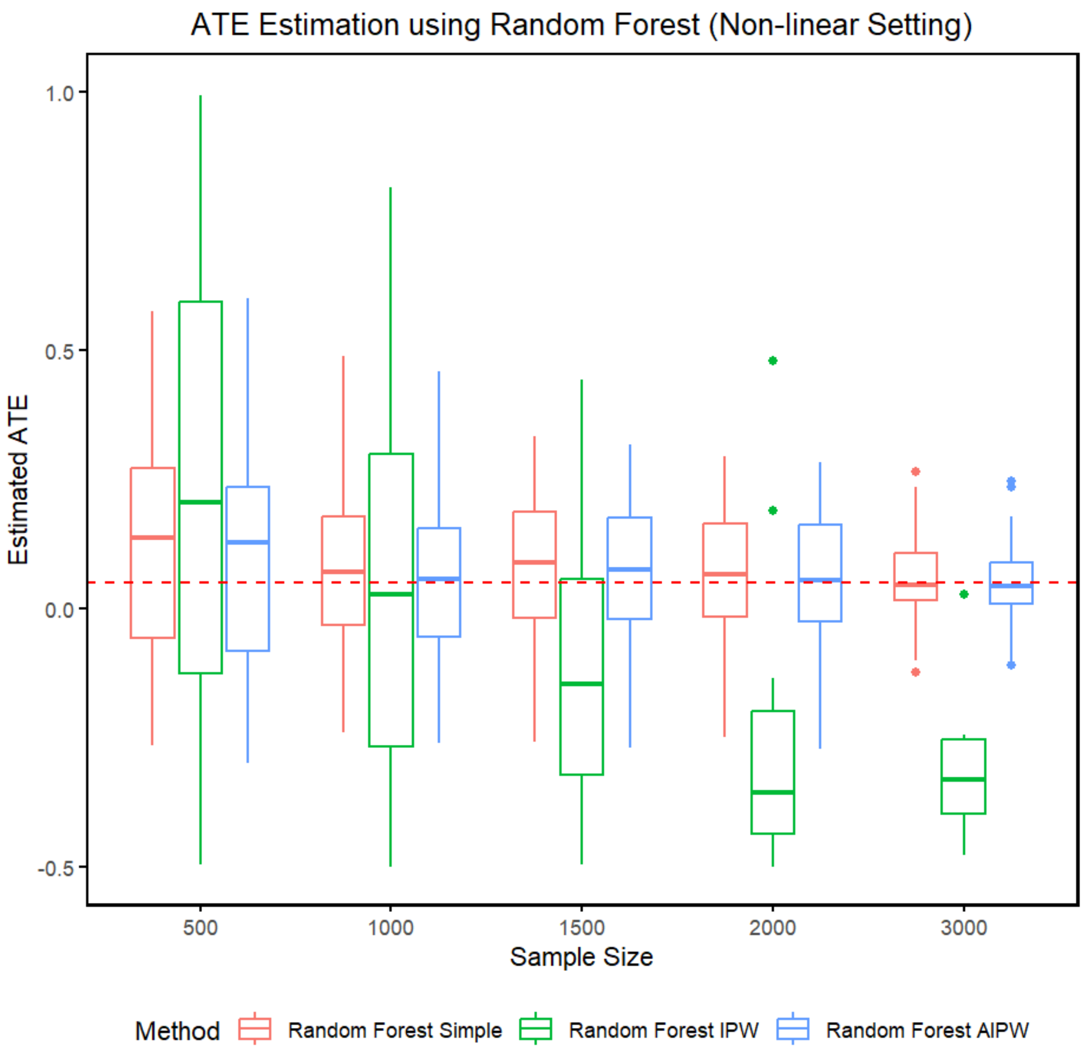


Figure 25: Comparison of different estimation of ATE

$$Y = \epsilon + X_1 + X_2 + U + \tau \cdot W$$

III Appendix: Case Study Results

III.1 GSS: Covariates Table

We choose 20 covariates from the GSS webpage as the table below:

Variable Name	**Description**
Number of Hours Worked Last Week	Number of hours worked last week
Gender	Gender of respondent, 1 is male
Number of Children	Number of children of respondent
Race	Race of respondent, 1 is black
Income	Labor force status of respondent
Age	Age of respondent
Education	Highest year of school completed
Political Views: Extremely Conservative	Binary variable of political views
Political Views: Conservative	Binary variable of political views
Political Views: Slightly Conservative	Binary variable of political views
Political Views: Moderate	Binary variable of political views
Political Views: Slightly Liberal	Binary variable of political views
Political Views: Liberal	Binary variable of political views
Political Views: Extremely Liberal	Binary variable of political views
Political Views: Others	Binary variable of political views
Marital Status: Never Married	Binary variable of marital status
Marital Status: Married	Binary variable of marital status
Marital Status: Separated	Binary variable of marital status
Marital Status: Divorced	Binary variable of marital status
Marital Status: Widowed	Binary variable of marital status

Table 9: Variables of GSS Data

III.2 GSS: Statistical Test Result of RCT

Variable	Treatment	Control	p
Year	1997.42 (7.2)	1997.69 (6.97)	0.01
Numbers of Hours Worked Last Week	41.85 (14.39)	41.89 (14.12)	0.86
Number of Children	1.67 (1.56)	1.6 (1.53)	0.01
Age	40.35 (12.53)	40.5 (12.47)	0.39
Years of Education	13.76 (2.82)	13.73 (2.81)	0.55
Gender	0.49 (0.5)	0.51 (0.5)	0.01
Race	0.14 (0.35)	0.13 (0.34)	0.24
Income	22071.37 (5739)	22157.93 (5622.51)	0.29
Political Views: Conservative	0.13 (0.34)	0.13 (0.34)	0.73
Political Views: Extremely Conservative	0.03 (0.16)	0.03 (0.16)	0.86
Political Views: Extremely Liberal	0.02 (0.15)	0.03 (0.16)	0.09
Political Views: Liberal	0.11 (0.32)	0.11 (0.31)	0.35
Political Views: Moderate	0.33 (0.47)	0.33 (0.47)	0.82
Political Views: Others	0.11 (0.31)	0.11 (0.31)	0.44
Political Views: Slightly Conservative	0.15 (0.36)	0.15 (0.36)	0.58
Political Views: Slightly Liberal	0.12 (0.33)	0.12 (0.33)	0.07
Marital Status: Never Married	0.25 (0.43)	0.25 (0.43)	0.56
Marital Status: Married	0.52 (0.5)	0.52 (0.5)	0.78
Marital Status: Separated	0.04 (0.19)	0.04 (0.19)	0.91
Marital Status: Divorced	0.16 (0.37)	0.17 (0.37)	0.34
Marital Status: Widowed	0.03 (0.18)	0.03 (0.17)	0.75
Propensity Score Estimated	0.5 (0)	0.5 (0)	0.56

Table 10: Descriptive Statistics and T-Test Results

III.3 GSS: OLS and Logistic Regression Full Result

Table 11: Linear Model Coefficients

	Dependent variable:
	Treatment Effect
Year	-0.001^{***}
	(0.0004)
Weekly Hours Worked	0.001^{***}
	(0.0002)
Number of Children	0.0003
	(0.002)
Age	-0.0001
	(0.0003)

Continued on next page

Table 11 – continued from previous page

	Dependent variable:
	Treatment Effect
Years of Education	-0.004^{***}
	(0.001)
Gender (1 = Male)	-0.004
	(0.006)
Race (1 = Black)	-0.091^{***}
	(0.009)
Income	0.00000^{***}
	(0.00000)
Political Views: Conservative	0.171^{***}
	(0.012)
Political Views: Extremely Conservative	0.255^{***}
	(0.020)
Political Views: Extremely Liberal	-0.007
	(0.020)
Political Views: Moderate	0.061^{***}
	(0.010)
Political Views: Others	0.042^{***}
	(0.013)
Political Views: Slightly Conservative	0.102^{***}
	(0.011)
Political Views: Slightly Liberal	0.014
	(0.012)
Treatment	-0.347^{***}
	(0.006)
Marital Status: Never Married	0.006
	(0.019)
Marital Status: Married	0.011
	(0.018)

Continued on next page

Table 11 – continued from previous page

	Dependent variable:
	Treatment Effect
Marital Status: Separated	-0.0001
	(0.023)
Marital Status: Divorced	0.018
	(0.018)
Constant	3.181^{***}
	(0.847)
Observations	19,723
R^2	0.182
Adjusted R^2	0.181
Note:	*p<0.05; **p<0.01; ***p<0.001

Table 12

	Dependent variable:
	Respond Too Much
Year	-0.009^{***}
	(0.003)
Weekly Hours Worked	0.007^{***}
	(0.001)
Number of Children	0.002
	(0.014)
Age	-0.0004
	(0.002)
Years of Education	-0.028^{***}
	(0.007)
Gender (1 = Male)	-0.028
	(0.037)
Race (1 = Black)	-0.615^{***}
	(0.059)
Income	0.00003^{***}
	(0.00000)

Continued on next page

Table 12 – continued from previous page

	Dependent variable:
	Respond Too Much
Political Views: Conservative	0.922***
	(0.071)
Political Views: Extremely Conservative	1.372***
	(0.116)
Political Views: Extremely Liberal	−0.123
	(0.132)
Political Views: Liberal	−0.095
	(0.079)
Political Views: Moderate	0.304***
	(0.062)
Political Views: Others	0.184**
	(0.078)
Political Views: Slightly Conservative	0.548***
	(0.069)
Marital Status: Never Married	0.044
	(0.119)
Marital Status: Married	0.076
	(0.110)
Marital Status: Separated	0.008
	(0.144)
Marital Status: Divorced	0.126
	(0.114)
Treatment	−2.023***
	(0.040)
Constant	15.995***
	(5.213)
Observations	19,723
Log Likelihood	−9,738.640

Continued on next page

Table 12 – continued from previous page

	Dependent variable:
	Respond Too Much
Akaike Inf. Crit.	19,519.280
Note:	*p<0.1; **p<0.05; ***p<0.01

III.4 GSS: Group Definition in Interaction Term

Category	**Group**	**Definition**
AGE	Young	<40 years old
	Middle	40-60 years old
	Old	>60 years old
Education	High	>12 years
	Middle	6-12 years
	Low	<6 years
Working Hours	High	>40 hours
	Middle	20-40 hours
	Low	<20 hours
Income	High	>15,000
	Middle	\$5,000-\$15,000
	Low	<\$5,000

Table 13: Grouping criteria for age, education, working hours, and income

III.5 GSS: Hypothesis Testing for GATE Difference

We choose g_1 and g_0 of each feature as the table below:

Variable Name	g_0	g_1
Political Views	Most Conservative	Most Liberal
Marital Status	Never Married	Married
Race	White	Black
Gender	Female	Male
Years of Education / Age / Number of Children / Income / Hours Worked Last Week	Top 10 Percentile of Feature	Bottom 10 Percentile of Feature

Table 14: Groups by Different Features

The statistical testing result is as table below, we can see for all covariates H_0 can be rejected and p-value are 0 if saved to 3 digits. The result reinforce the evidence of HTE across each feature and we can observe marital status and hours worked last week have the most significant t statistics. Intuitively more hard working people more value their labor force so holding more negative attitude towards welfare which comes from their hard work's tax. For married repondents since they have family and need more money, so they also have more negative attitude towards welfare which increases their tax burden.

Feature	g_0	g_1	t-statistic (p-value)
Political Views	-0.344 (0.003)	-0.322 (0.003)	-5.330 (1.21e-07)
Marital Status	-0.317 (0.001)	-0.358 (0.001)	39.068 (1.67e-310)
Race	-0.346 (0.001)	-0.327 (0.001)	-16.099 (1.80e-57)
Gender	-0.338 (0.001)	-0.349 (0.001)	11.530 (1.17e-30)
Years of Education	-0.281 (0.001)	-0.339 (0.001)	31.627 (2.69e-158)
Income	-0.354 (0.005)	-0.288 (0.002)	-6.794 (2.43e-11)
Age	-0.334 (0.002)	-0.321 (0.001)	-10.108 (9.92e-24)
Number of Children	-0.326 (0.001)	-0.328 (0.001)	-7.152 (0.000)
Hours Worked Last Week	-0.388 (0.001)	-0.301 (0.001)	47.152 (0.000)

Table 15: T-test Results for GATE across Different Features

III.6 OHIE: Covariates Table

Variable	Description
Positive Medical Care Attitude	Binary, 1 if the respondent has a positive attitude towards medical care in the past 6 months
Medicaid Enrollment	Binary, 1 if the respondent enrolls in the Medicaid program
Lottery Winner	Binary, 1 if the respondent wins the lottery
Age	Age of the respondent
Household Size	Number of people in the respondent's household
Female	Binary, 1 if the respondent is female
First Day Signup	Binary, 1 if the respondent signs up for the lottery on the first day
Metropolitan Area	Binary, 1 if the respondent is from a metropolitan statistical area
Cigarettes per Day	Number of cigarettes smoked per day by the respondent
Average Weekly Working Hours	Average number of working hours per week
Education Level	Education level of the respondent
Hispanic	Binary, 1 if the respondent is Hispanic
White	Binary, 1 if the respondent is White
Black	Binary, 1 if the respondent is Black
Asian	Binary, 1 if the respondent is Asian
American Indian	Binary, 1 if the respondent is American Indian
History of Disease	Binary, 1 if the respondent has ever been diagnosed with Diabetes, Asthma, High Blood Pressure, COPD, Heart Disease, Congestive Heart Failure, High Cholesterol, or Kidney Problems
Household Income	Income of the respondent's household

Table 16: Description of Variables

III.7 OHIE: First Stage Regression Result

Table 17: Regression Results

	Dependent variable:
	Medicaid Enrollment (W)
First Day Listed	0.014
	(0.009)
Zip Metropolitan Area	−0.002
	(0.006)
Number of Household Members	-0.023^{***}
	(0.006)
Age	−0.0003

Continued on next page

Regression Results (Continued)

	Dependent variable:
	Medicaid Enrollment (W)
	(0.0002)
Female	-0.017^{***}
	(0.006)
Education (Years)	-0.002
	(0.003)
Race: Pacific Islander	-0.040
	(0.030)
Race: Asian	0.008
	(0.017)
Race: American Indian	-0.015
	(0.012)
Race: Black	0.002
	(0.018)
Race: White	0.019^{*}
	(0.011)
Race: Hispanic	-0.006
	(0.011)
Household Income Category	-0.013^{***}
	(0.001)
Employment Hours Per Week	-0.011^{***}
	(0.002)
Smoking Average (Moderate)	0.0002
	(0.0004)
Combined Diagnosis	0.016^{***}
	(0.006)
Lottery Status (Z)	0.338^{***}
	(0.005)
Constant	0.142^{***}
	(0.020)
Observations	14,219
R^2	0.242
Adjusted R^2	0.241
Residual Std. Error	0.325 (df = 14201)
F Statistic	266.075^{***} (df = 17; 14201)
Note:	*p<0.1; **p<0.05; ***p<0.01

III.8 OHIE: Balance Test

Table 18: Covariate Balance by Lottery Status (Mean, SD, and p-value)

Covariate	Lottery Status = 0	Lottery Status = 1	p-value
First Day Listed	0.1 (0.3)	0.1 (0.31)	0.8490
Zip Metropolitan Area	0.75 (0.43)	0.74 (0.44)	0.1970
Number of Household Members	1.27 (0.45)	1.37 (0.49)	0.0000
Age	41.69 (12.19)	41.67 (12.21)	0.9281
Female	0.61 (0.49)	0.59 (0.49)	0.0375
Education (Years)	2.35 (0.83)	2.35 (0.84)	0.6940
Race: Pacific Islander	0.01 (0.09)	0.01 (0.09)	0.6878
Race: Asian	0.03 (0.18)	0.04 (0.19)	0.2992
Race: American Indian	0.06 (0.24)	0.05 (0.23)	0.0840
Race: Black	0.03 (0.16)	0.04 (0.2)	0.0242
Race: White	0.85 (0.36)	0.84 (0.36)	0.6054
Race: Hispanic	0.1 (0.3)	0.1 (0.3)	0.7540
Household Income Category	6.87 (4.86)	7.1 (4.96)	0.0063
Employment Hours Per Week	2.13 (1.77)	2.16 (1.32)	0.0777
Smoking Average (Moderate)	4.68 (7.88)	4.73 (7.93)	0.7066
Combined Diagnosis	0.67 (0.47)	0.65 (0.48)	0.0346

III.9 OHIE: Second Stage Regression Result

Table 19: Second Stage Regression Results

	Dependent variable:
	Medical Adequate (Y)
Fitted Enrollment Medicaid (W)	0.223^{***}
	(0.024)
Age	-0.013
	(0.013)
Number of Household Members (Numhh)	0.002
	(0.009)
Gender	0.032^{***}
	(0.009)
First Day	0.001^{***}
	(0.0003)
Zip Metropolitan	-0.031^{***}
	(0.008)

Continued on next page

Table 19 – continued from previous page

	Dependent variable:
	Medical Adequate (Y)
Cigarette per Day	0.001 (0.005)
Average Working Hours	0.037 (0.044)
Education	0.051^{**} (0.025)
Race Hispanic	-0.070^{***} (0.018)
Race White	0.007 (0.026)
Race Black	-0.066^{***} (0.016)
Race Asian	0.053^{***} (0.016)
Race Amerindian	0.006^{***} (0.001)
Has Ever Disease	0.006^{*} (0.003)
Household Income	-0.003^{***} (0.001)
Xcombined_diagnosis	-0.113^{***} (0.009)
Constant	0.586^{***} (0.030)
Observations	14,219
R^2	0.039
Adjusted R^2	0.038
Residual Std. Error	0.475 (df = 14201)
F Statistic	33.948^{***} (df = 17; 14201)
Note:	*p<0.1; **p<0.05; ***p<0.01

III.10 OHIE: Group Definition in Interaction Term

Category	**Group**	**Definition**
Age	Young	<35 years old
	Middle	36-50 years old
	Old	>50 years old
Income	Low	<15,000
	Middle	15,000-35,000
	High	>35,000
Household Size	1 Person	= 1
	>1 Person	>1
Smoker	Non-smoker	0
	Smoker	>0
Employment Hours	Part-time	0-30 hours
	Full-time	>30 hours

Table 20: Grouping criteria for age, income, household size, smoking status, and employment hours

III.11 Statistical Testing of GLATE by Instrumental Causal Forest

Table 21: Statistical Test Results for Top 10% vs Bottom 10% CLATE

Feature	**Top_Mean_CLATE**	**Bottom_Mean_CLATE**	**p_value**
Household Size	0.194 (0.052)	0.227 (0.050)	0.00
Age	0.243 (0.033)	0.194 (0.040)	0.00
Weekly Working Hours	0.176 (0.039)	0.236 (0.050)	0.00
Income	0.187 (0.045)	0.246 (0.050)	0.00
Smoking	0.239 (0.049)	0.210 (0.053)	0.00
White	0.219 (0.052)	0.199 (0.053)	0.00
Severe Disease	0.236 (0.049)	0.178 (0.038)	0.00

Bibliography

Abadie, Alberto and Guido Imbens (2006). “Large sample properties of matching estimators for average treatment effects”. In: *Econometrica* 74.1, pp. 235–267.

Angrist, Joshua (1990). “Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records”. In: *The American Economic Review* 80.3, pp. 313–336.

Angrist, Joshua, Guido Imbens, and Donald Rubin (1996). “Identification of causal effects using instrumental variables”. In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Angrist, Joshua and Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion.* Princeton, NJ: Princeton University Press. ISBN: 978-0691120355.

— (2014). *Mastering ’Metrics: The Path from Cause to Effect.* Princeton, NJ: Princeton University Press. ISBN: 978-0691152844.

Athey, Susan (2019). “The Impact of Machine Learning on Economics”. In: *The Economics of Artificial Intelligence: An Agenda.* Ed. by Agrawal, Ajay, Gans, Joshua, and Goldfarb, Avi. National Bureau of Economic Research, NBER. Chicago: University of Chicago Press, pp. 507–547. ISBN: 978-0-226-61333-8.

Athey, Susan and Guido Imbens (2016). “Recursive Partitioning for Heterogeneous Causal Effects”. In: *Proceedings of the National Academy of Sciences* 113.27. Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 20, 2016 (received for review June 25, 2015), pp. 7353–7360. DOI: `10.1073/pnas.1510489113`. URL: `http://www.pnas.org/content/113/27/7353`.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). “Generalized random forests”. In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Athey, Susan and Stefan Wager (2019). “Estimating Treatment Effects with Causal Forests: An Application”. In: *Observational Studies* 5, pp. 37–51. URL: `https://obsstudies.org/estimation-treatment-effects-with-causal-forests/`.

Baiardi, Anna and Andrea A. Naghi (2024). “The value added of machine learning to causal inference: evidence from revisited studies”. In: *The Econometrics Journal* 27.2. First version received: 15 September 2022; final version accepted: 13 December 2022, pp. 213–234. DOI: `10.1093/ectj/utae004`. URL: `https://doi.org/10.1093/ectj/utae004`.

Breiman, Leo (1996). “Bagging Predictors”. In: *Machine Learning* 24.2, pp. 123–140.

— (2001). “Random forests”. In: *Machine learning* 45.1, pp. 5–32.

— (2004). “Consistency for a Simple Model of Random Forests”. In: *arXiv preprint arXiv:1405.2881.* URL: `https://arxiv.org/abs/1405.2881`.

Breiman, Leo et al. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth International Group.

Card, David and Alan B. Krueger (1994). “Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania”. In: *The American Economic Review* 84.4, pp. 772–793.

Casella, George and Roger L Berger (2002). *Statistical Inference.* Duxbury.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794.

Chernozhukov, Victor, Denis Chetverikov, et al. (2017). *Double/Debiased Machine Learning for Treatment and Causal Parameters*. arXiv: `1608.00060 [stat.ML]`. URL: `https://arxiv.org/abs/1608.00060`.

— (2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68.

Chernozhukov, Victor, Christian Hansen, et al. (Mar. 2024). *Applied Causal Inference Powered by ML and AI*. Online.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch (2010). "BART: Bayesian Additive Regression Trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298. DOI: `10.1214/09-AOAS285`.

Davis, Jonathan M.V. and Sara B. Heller (2017). "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs". In: *American Economic Review: Papers & Proceedings* 107.5, pp. 546–550. DOI: `10.1257/aer.p20171000`. URL: `https://doi.org/10.1257/aer.p20171000`.

Efron, Bradley (2020). "Prediction, Estimation, and Attribution". In: *Journal of the American Statistical Association* 115.530, pp. 636–655.

Finkelstein, Amy et al. (Aug. 2012). "The Oregon Health Insurance Experiment: Evidence from the First Year". In: *The Quarterly Journal of Economics* 127.3, pp. 1057–1106. DOI: `10.1093/qje/qjs020`. URL: `https://doi.org/10.1093/qje/qjs020`.

Freund, Yoav and Robert E. Schapire (1996). "Experiments with a new boosting algorithm". In: *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 148–156.

Green, Donald P. and Holger L. Kern (2012). "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees". In: *The Public Opinion Quarterly* 76.3, pp. 491–511. URL: `https://www.jstor.org/stable/41684581`.

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho (2020). "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects". In: *Bayesian Analysis* 15.3, pp. 965–1056. DOI: `10.1214/19-BA1195`. URL: `https://projecteuclid.org/journals/bayesian-analysis/volume-15/issue-3/Bayesian-Regression-Tree-Models-for-Causal-Inference--Regularization-Confounding/10.1214/19-BA1195.full`.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York, NY: Springer. ISBN: 978-0387848570. URL: `https://web.stanford.edu/~hastie/ElemStatLearn/`.

Hill, Jennifer L. (2011). "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240. DOI: `10.1198/jcgs.2010.08162`.

Hirano, Keisuke, Guido Imbens, and Geert Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score". In: *Econo-*

metrica 71.4, pp. 1161–1189. DOI: 10.1111/1468-0262.00442. URL: https://www.jstor.org/stable/30038086.

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1, pp. 55–67. DOI: 10.1080/00401706.1970.10488634. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

Imbens, Guido and Joshua Angrist (1994). "Identification and estimation of local average treatment effects". In: *Econometrica* 62.2, pp. 467–475.

Imbens, Guido and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York, NY: Cambridge University Press. ISBN: 9780521885881. DOI: 10.1017/CBO9781139025751.

Jawadekar, Neal et al. (2023). "Practical Guide to Honest Causal Forests for Identifying Heterogeneous Treatment Effects". In: *American Journal of Epidemiology* 192.7, pp. 962–974. DOI: 10.1093/aje/kwad043. URL: https://academic.oup.com/aje/article/192/7/1154/7059288.

Johansson, Fredrik, Uri Shalit, and David Sontag (2016). "Deep Learning for Causal Inference". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1025–1034. DOI: 10.1145/2939672.2939786.

Johnson, Michael, Jiongyi Cao, and Hyunseung Kang (2022). "Detecting heterogeneous treatment effects with instrumental variables and application to the Oregon health insurance experiment". In: *The Annals of Applied Statistics* 16.2, pp. 1111–1129. DOI: 10.1214/21-AOAS1535. URL: https://doi.org/10.1214/21-AOAS1535.

Kennedy, Edward H. (2023). *Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects.* arXiv: 2004.14497 [math.ST].

Kunzel, Sören R. et al. (2019). "Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning". In: *Proceedings of the National Academy of Sciences* 116.10, pp. 4156–4165.

Laan, Mark J. van der and Sherri Rose (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4419-9781-6. DOI: 10.1007/978-1-4419-9782-3.

Lin, Yi and Yongho Jeon (2006). "Random forests and adaptive nearest neighbors". In: *Journal of the American Statistical Association* 101.474, pp. 578–590. DOI: 10.1198/016214505000001230.

Nie, Xinkun and Stefan Wager (2020). "Quasi-Oracle Estimation of Heterogeneous Treatment Effects". In: *arXiv preprint arXiv:1712.04912.*

O'Neill, Eoghan and Melvyn Weeks (Oct. 2018). *Causal Tree Estimation of Heterogeneous Household Response to Time-of-Use Electricity Pricing Schemes.* Cambridge Working Papers in Economics 1865. University of Cambridge, Faculty of Economics. URL: https://www.econ.cam.ac.uk/research-files/repec/cam/pdf/cwpe1865.pdf.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of Regression Coefficients When Some Regressors are not Always Observed". In: *Journal of the American Statistical Association* 89.427, pp. 846–866.

Robinson, Peter M. (1988). "Root-N-consistent semiparametric regression". In: *Econometrica* 56.4, pp. 931–954.

Rosenbaum, Paul R. and Donald B. Rubin (1983). “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1, pp. 41–55.

Rubin, Donald B. (1974). “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies”. In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Scornet, Erwan, Gérard Biau, and Jean-Philippe Vert (2015). “Consistency of random forests”. In: *Annals of Statistics* 43.4, pp. 1716–1741.

Shi, Claudia, David M Blei, and Victor Veitch (2019). “Adapting Neural Networks for the Estimation of Treatment Effects”. In: *arXiv preprint arXiv:1906.02120*.

Singh, Rahul (June 2023). “Essays on Econometrics, Causal Inference, and Machine Learning”. Submitted to the Department of Economics and the Statistics and Data Science Center in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics and Statistics. PhD thesis. Cambridge, MA: Massachusetts Institute of Technology. URL: `https://hdl.handle.net/1721.1/129456`.

Smith, Tom W. et al. (2018). *General Social Surveys, 1972–2016 [machine-readable data file]*. Ed. by NORC. Principal Investigator, Smith, Tom W.; Co-Principal Investigators, Peter V. Marsden and Michael Hout; Sponsored by National Science Foundation, ed. NORC, Chicago: NORC: NORC at the University of Chicago [producer and distributor], Data accessed from the GSS Data Explorer website at gssdataexplorer.norc.org.

Splawa-Neyman, Jerzy (1923). “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9”. In: *Statistical Science* 5.4. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923) 1-51 (Annals of Agricultural Sciences), pp. 465–480.

Tibshirani, Robert (1996). “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. ISSN: 1369-7412. URL: `https://academic.oup.com/jrsssb/article/58/1/267/1729929`.

Wager, Stefan (May 2016). “Causal Inference with Random Forests”. PhD dissertation. Stanford, CA: Stanford University. URL: `https://purl.stanford.edu/kz350cg6301`.

— (2023). *Machine Learning and Causal Inference.* Lecture notes. Department of Statistics.

Wager, Stefan and Susan Athey (2018). “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.