Causal Tree Estimation of the Conditional Average Treatment Effect of Education on Earning

Huaning Liu, Noah Simpson, Xue Wang, Wenqian Zhao

March 2023

1 Abstract

Over the past few decades, thousands of studies have been focusing on investigating the rate of return to education. Although a vast majority of them indicated a positive relationship between education level and income, social scientists who are using traditional methods such as classification and regression trees (CART) and OLS have to be meticulous to make inferences about their causal relationship and one of the big concerns is the omitted variables-It is hard to tell whether higher education attainment causes somebody to earn more or the higher income is caused by other factors such as the individual's ability. In the traditional use of CART, such methods are used to predict outcomes given a population and its covariates, typically with the ground truth for each observation already known. However, when using CART for estimating causal effects rather than prediction when we have no knowledge about the effect of the treatment, meaningful results become far more challenging to obtain, with bias and sparsity becoming increasingly arduous to maneuver. This paper intends to estimate the Conditional Average Treatment Effect (CATE) of education level on earning using the "honest" estimator proposed in the paper *Comprehension and Reproduction of Recursive Partitioning for Heterogeneous Causal Effects*, written by Susan Athey and Guido Imbens. The authors created and benchmarked an unbiased estimator of CATE across subsets of the population with different treatments, proposing an "honest" approach for estimation. An "honest" approach with CART involves using one sample to construct partitions in the tree and using a separate sample to estimate the treatment effects, which removes spurious correlations between the covariates and outcomes in the model. Therefore, we aim to apply this approach to get an unbiased estimate of the CATE of college education on earnings.

2 Introduction

Education is widely considered as the cornerstone of personal and societal progress. It has the power to shape individuals into well-rounded, informed citizens and provide them with the necessary skills to succeed in their personal and professional lives. Moreover, education is a crucial factor in the development of a thriving and prosperous society. From increasing economic growth to promoting civic engagement, education continues to be the key determinant of success in various aspects of life, especially of the career prospects and high-status income level[26]. Economists have started the investigation into the causal relationship between education attainments and income level in relatively early years. In a research of human capital conducted by Gary Becker in 1963, he identified the return to education as a central factor for labor market policy-making[6]. However, there are many challenges in terms of inferring the causal effect of education on earnings, such as the inaccessibility of demographic profiles, the infeasibility of the counterfactual potential outcomes, and the bias brought by the interference of other variables. To deal with those challenges, we query the IPUMS database to get adequate demographic data and used the "honest" estimator in the Causal Tree algorithm proposed by Susan Athey and Guido Imbens to get an unbiased estimation of the CATE.

2.1 Literature Review

Education vs. Income
The relationship between education attainment and income level has been a popular topic of discussion in academic literature. A wide range of studies have been conducted to examine this relationship and the findings have been illuminating. Albert E. Beaton's study in 1975 revealed that individuals with more years of education and similar aptitude, as measured by aptitude scores, tend to earn higher salaries on average within the sample population [5]. Sociologists and economists have reached a consensus regarding the connection between education and employment, which can be summarized in two aspects. Firstly, individuals who have received more education have developed

enhanced practical and interpersonal skills, which are directly linked to their promotion prospects and capacity to perform their job duties effectively [12]. Secondly, these individuals will possess a higher socio-economic status recognition by society, which provides them with greater access to prestigious employment opportunities[8].
However, a number of studies have also shown that other social factors can in some cases overshadow the effect of education on income. Donna Bobbitt-Zeher's research from 2007 uncovered a significant disparity in the annual income of women compared to men with comparable educational backgrounds, with women earning approximately $4,400 less per year[7]. This finding has motivated our investigation in this paper into the impact of education on income disparities between different genders.

Causal Inference
The framework of Mincer's Model provided a foundation for most of the recent studies of education and income determination[24]:

$$\log w = \alpha + \rho s + \beta_1 x + \beta_2 x^2 + e,$$

where w stands for wage, s stands for years of schooling, and x stands for years of experience. Studies have demonstrated a positive and crucial role of education, especially high school and college, on one's enhancement of social economic status[2][20]. However, economists recognized that the correlation between years of education and income obtained by OLS might not be the same as their causal effect due to other factors such as ability bias, so when inferring causal relationships they have to be extra careful[10].

Angrist and Krueger used the quarter of birth in the U.S. as the instrumental variable for the OLS in their study of the return to education. Due to the fact that students cannot enroll in the first grade by the age of 6 and cannot drop out by the age of 17, they investigated the problem that whether differences in education determined by differences in birthdays translate into differences in earnings. In doing so, they estimated the return on earnings per year of education to be 0.102[1]. Harmon et al who similarly utilized the changes in the compulsory school attendance law as instrumental variables for educational attainment estimated the returns to education to be 0.15 [19]. To remove the bias brought by the correlation of other variables with education and income, Ashenfelter and Zimmerman utilized family backgrounds such as parent and sibling education as a control to estimate the causal effect of education on earning to be 0.08[3]. In their study of the causal effect of holding a GED(Graduate Equivalency Degree) on earnings in the labor market for high school dropouts, Tyler et al. utilized the difference-in-difference method to control for GED test scores and compare mean income for states with different standards for passing the GED. Their results indicate a 10 to 19 increase in earnings for holding a GED certificate [21].

Within the area of causal inference, causal decision trees have been discussed and extended since its first proposal, whose structures are continually being varied and optimized [23] [27]. Similar nonlinear tree-structured regression algorithms are broadly studied under existing works, such as Bayesian regression trees, probability trees, and factor modeling, etc [18] [14][13]. It is generally benefited from the advantages of tree algorithms, which could provide a compact graphic, coupled with a typically low run-time [23]. The work we are investigating and replicating in this paper examines such tree structure algorithms under heterogeneous causal effects cases, which is a large branch in the causal inference that is consistently and increasingly studied [25] [16]. The proposed unbiased splitting criterion and "honesty" accordingly fill in a blank of a population partitioning strategy in the decision tree algorithm and further generalizes the use of regression trees in causal works.

2.2 Dataset

We gained our inspiration from the IPUMS database, which contains U.S. census microdata from around 1980 to 2020, with nearly trillions of pieces of data and thousands of features. Census data has been widely applied to studies about the causal analysis of education on earnings- Dustmann and Schönberg used UK census data and Jäntti and Reinikka-Soininen used Vietnamese census data in their studies to estimate the causal effect of education on earnings, controlling for various factors such as ability, family background, and occupation[11] [22]. One of the advantages of using this kind of data is that the sample size will be large enough, so we don't need to worry about the insufficiency of data after cleaning and dropping null values. Upon checking, the data is fully available with an online database query; therefore we queried the database and did simple data processing to attain the desired features and variables.

In past analyses, researchers tend to use demographic, occupational, and geographic features such as age, gender, race, job type, and location as features for their causal study[1]. In Goldin and Katz's study, they used a regression analysis to estimate the returns to education, controlling for other factors that may affect earnings such as occupation, age, and experience. They also compare the returns to education for different groups of individuals based on race, gender, and birthplace to examine differences in the labor market outcomes of these groups[15]. Therefore, we would like to adopt similar feature-selection strategies to include features such as sex, state, and insurance status to estimate the causal effect. We also want to compare the CATE across demographic groups to see the difference like what Goldin and Katz did in their study.

3 Setup

3.1 Causal Effect

For every unit i in N, where $i = 1, 2, ..., N$, there are two potential outcomes, denoted by:

$$(Y_i(0), Y_1(1))$$

τ_i is defined to be the unit-level difference in potential outcomes, denoted by:

$$\tau_i = Y_i(1) - Y_i(0)$$

$W_i \in \{0, 1\}$ is defined to be the indicator for whether unit i received the treatment or not. If $W_i = 0$: unit received control; If $W_i = 1$: unit received treatment. For every unit i we only have one of the previous values of W observed, therefore Y_i^{obs} is as follows:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0), & \text{if } W_i = 0 \\ Y_i(1), & \text{if } W_i = 1 \end{cases}$$

The population average outcome for all covariates in a given partition of a tree is defined to be

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

Let S_w denote the sample space for a specific treatment w, the unbiased estimator for the average outcome given a partition Π and a sample space S is defined as

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

Given the average outcome, the CATE is defined as

$$\tau(x; \Pi) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i \in l(x; \Pi)] = \mu(1, x; \Pi) - \mu(0, x; \Pi)$$

the difference of the average outcome between the treatment and control group. Therefore, the unbiased estimator for CATE is defined as

$$\hat{\tau}(s; S, \Pi) \equiv \hat{\mu}(1, x; S, \Pi) - \hat{\mu}(1, x; S, \Pi),$$

the difference of the estimated average outcome for all covariates given a sample space S and a partition Π between the treatment and control group.

3.2 Methodology

In this section, we briefly review the two tree-based method to be practiced in this work.

Honest Tree
The "honest" approach used for building and validating the Causal Tree is an extension and divergence from the classification and regression trees(CART) algorithm.

Due to the fact that we cannot observe both $Y_i(1)$ and $Y_i(0)$ for an individual, the true treatment effect τ is also not observable since we are missing half of the Y^{obs}. Thus, the "honest" version of $EMSE_\tau(\Pi)$:

$$EMSE_\tau(\Pi) \equiv \mathbb{E}_{S^{te}, S^{est}}[MSE_\tau(S^{te}, S^{est}, \Pi)]$$

is not feasible as we have no knowledge about τ_i. Therefore, the paper proposed an estimator for $EMSE_\tau(\Pi)$ by modifying the MSE_μ in CART to get an unbiased estimator $\widehat{MSE}_\tau$ for the treatment effect and the $EMSE_\tau(\Pi)$ in

”honest” algorithm to get an unbiased estimator $\widehat{EMSE}_\tau(S^{tr}, N^{est}, \Pi)$ for $EMSE_\tau(\Pi)$.

Let p denote the proportion of the treated individuals in a leaf, $S^{tr}_{control}$ denote the subsample of the control group in the training sample, and S^{tr}_{treat} denote the subsample of the treatment group in the training sample, the unbiased estimator for $EMSE_\tau(\Pi)$ for splitting is defined as

$$-\widehat{EMSE}_\tau(S^{tr}, N^{est}, \Pi) \equiv \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\tau}^2(X_i; S^{tr}, \Pi) - (\frac{1}{N^{tr}} + \frac{1}{N^{est}}) \cdot \sum_{l \in \Pi} (\frac{S^2_{S^{tr}_{treat}}(l)}{p} + \frac{S^2_{S^{tr}_{control}}(l)}{1-p}).$$

Using the same equation as the splitting criterion with cross-validation sample, the unbiased estimator of $EMSE_\tau(\Pi)$ for cross-validation is $-\widehat{EMSE}_\tau(S^{tr,cv}, N^{est}, \Pi)$.

Generalization to Random Forests

Breiman [9] first proposed the idea of growing random forest via bootstrap aggregation from decision trees on randomized subsamples. It is primarily defined on the setting of $(X_i, Y_i) \in \chi \times \mathbb{R}$ towards the estimation of $\mu(x) = \mathbb{E}[Y_i|X = x_i]$. Then for training set $\{X_i, W_i, Y_i\}_{i=1}^n$, a proportion $p \in (0, 1]$ fixed, we did the following two steps iteratively K times: for kth iteration, 1. draw a subset from the training set by proportion p with replacement, denote it to be Z_k 2. build an honest tree based on Z_k, denote it to be $h_k(\cdot)$. This process gives us a collection of honest trees, say $\{h_i\}_{i=1}^K$. Then considering an observation in the test set (X_i, W_i), we predict it on these K trees separately and take an average to derive the prediction result for treatment effect.

Note that the algorithm above takes a deviation from Athey et al. work on generalizing random forest [4], that we simplified the authors' setting to be a simple bootstrap. Still, in the empirical application part, we employed the ”grf” library they offered for lower computation cost.

3.3 Dataset Description

Query & Covariates Extraction

We will utilize the US census data to investigate the impact of college completion on yearly income. Specifically, we have transformed the education level variable (EDUC), which ranges from 0 to 11, into a binary variable with a value of 1 if the level is equal to 10, indicating college completion, and 0 for individuals with level of 6. We also discarded observations with age less than 30 to make sure most individuals has completed their terminal degree by the time taking this survey, in order to consolidate the binarity of treatments. This transformed variable is denoted as the treatment variable W_i. Furthermore, we have logarithmically transformed the income level variable (INCWAGE) to serve as our outcome variable Y. This transformation is necessary due to the typically skewed nature of income distributions and the tendency for income variance to increase with income level, leading to heteroscedasticity [17]. By taking the logarithm of income, the data is more amenable to our analysis as the skewness is reduced, and variance is stabilized. We also consider other demographic variables such as gender (SEX), age (AGE), and race (RACE) as candidates of covariates.

Further Processing & Missing Data

The census data being used was very clean; there were almost no missing values, only some being found in the INCWAGE/AGE categories. We looked only at ages from 30-65, eliminating outliers, and also to be more confident that those in the census who have a high school or college education were done with schooling. We also deem retired wages above 65 to be unimportant and potentially biased in our treatment effect, as there are many different financial mechanisms that occur in retirement that influence yearly wage. Missing values in INCWAGE were also removed. Consider the variable race, sex and state to be categorical in the raw tabular data, we use one-hot encoder to encode them into matrix with binary entries to erase the ordinality within data that might affect the regression and cause bias. Therefore, we process and extract such covariates, denoted to be $\{X_i\}$ following the setting above, with a column dimension of around 40 since some states are not included. In light of computational limitations, we constrained our data analysis to a subset of 5 million raw observations drawn from the IPUMS database, out of which we ultimately retained a sample of 1 million processed data points. To mitigate any potential sampling bias, we executed a purely random query process to ensure that our sub-sample constituted an unbiased representation of the population data, thus facilitating our analysis.

Problem of Interest

To further our study into specific research problems, several datasets are created, but all datasets more or less utilize the same covariates. First, we investigated the effect of college on income for several different years, being 2010, 2000, and 1990. All three datasets utilize the same variables, and none of the sets contain any missing values. Second, we investigated the effect of college on income for males and females. We split these datasets up, so the sex covariate was dropped from both. Neither contain any missing values, and both groups attend college at approximately the same rate. Further, only data from 2010 is used. Lastly, we investigated the effect of college on yearly income for different age groups, being people in their 30s, 40s, and 50s. Again, except the problem that compares the decades, the other two datasets were both taken from 2010 real data.

The variables being used are as follows in Table [1]:

Variable	Type	Description
AGE	Covariate	Age of individual, ranging from 30 to 65.
SEX	Covariate	Sex of individual, encoded as 1 for male and 0 for female.
RACE	Covariate	Race of individual, one hot encoded from a range of 1 to 9. Legend in appendix.
STATEFIP	Covariate	State FIPS code, one hot encoded from 1 to 56. Legend in appendix.
HCOVANY	Covariate	Whether an individual has any form of healthcare, encoded as 1 for yes and 0 for no.
EDUC	Treatment	Whether an individual has completed college education or completed high school education, encoded as 1 for college and 0 for high school. This is the observed treatment.
INCWAGE	Outcome	Log of the yearly wage of an individual.

Table 1: Variable(raw Covariate) Description

3.4 Exploratory Data Analysis

In order to get a more general sense, we plotted the distribution of variables that are relevant to our research problems.

The analysis of Figure 1 reveals that there are no discernible variations in the logged mean annual income between different age cohorts in 2010. It is noteworthy, however, that the elderly group exhibits marginally higher mean income, although this difference is inconspicuous on account of the utilization of a logarithmic scale. Additionally, the 3D scatter plot highlights that individuals in their thirties and seventies tend to have the lowest income, which corroborates the veracity of the notion that individuals who are entering or exiting the labor market typically earn a lower income.

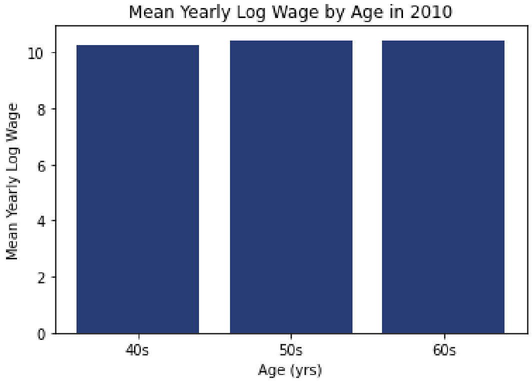

Figure 1: (log)Wages in Different Ages

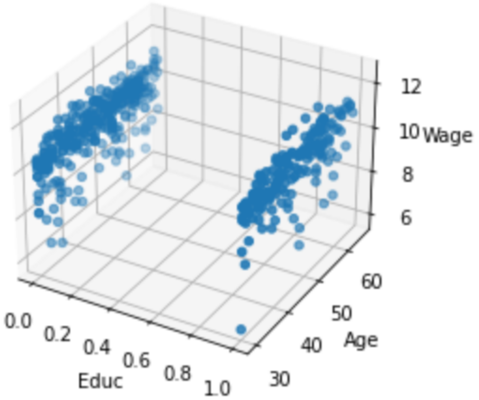

Figure 2: 3D Scatterplot of Education Level v.s. (log)Wage vs. Age

Figure 3 presents a conspicuous distinction in the proportion of individuals who received a college education across various age cohorts in 2010. Notably, the older age groups exhibit lower rates of college attendance. As a result, we seek to ascertain whether a difference exists in the conditional average treatment effect of education on earnings between individuals belonging to different age categories.

Based on the findings presented in Figure 4, it is apparent that there exists a variation in the logged mean annual income throughout different decades, indicating a tendency towards increased income in later years. In light of this observation, we are inclined to investigate whether there are any dissimilarities in the CATE of education on earnings across the various decades.

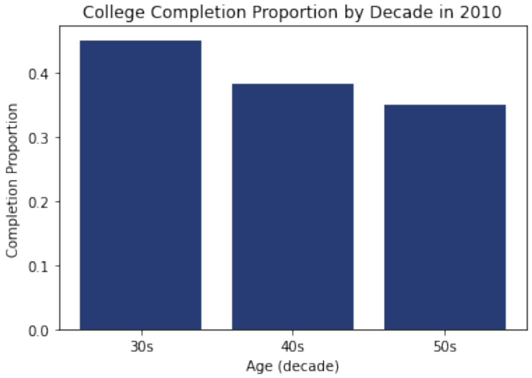

Figure 3: College Proportion in Different Ages in 2010

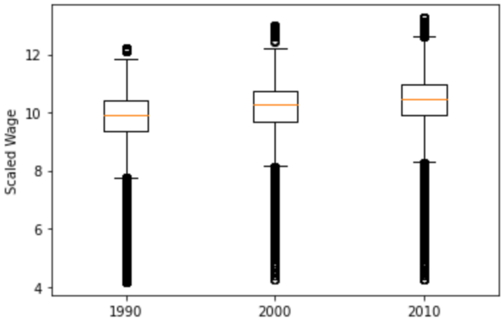

Figure 4: (log)Wage in Different Decades

4 Honest Tree Results

4.1 Training and Testing Details

Focusing on the three problems of interest above, we extend the practical application to tree-based method discussed in 3.2, and present the CATE estimation on each leaf. In this case, we are interested in the visualization of the trees with respect to each of the subgroups proposed by the three questions of interest above. Related visualizations and tables will be presented below. We are sampling a training set of size $50,000$ *with no replacement* and another testing set of size $5,000$ from the database (processed and encoded).

4.2 CATE across decades

The first problem of interest focuses on the conditional average treatment effect of college education on yearly wages in the years 1990, 2000, and 2010. Our objective is to examine the variation in CATE estimates over time (decades). Our analysis reveals that deeper trees and larger CATE estimates are produced in the later decades. Specifically, the CATE estimation across leaves for the 1990s, 2000s, and 2010s are 0.7723, 0.547, and 0.481, respectively. This finding suggests a decreasing trend in the impact of college education on income over time. The results of this study contribute to the literature on the relationship between education and income and provide insights for policymakers and educators.

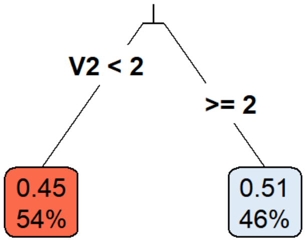

Figure 5: Tree by data in 1990

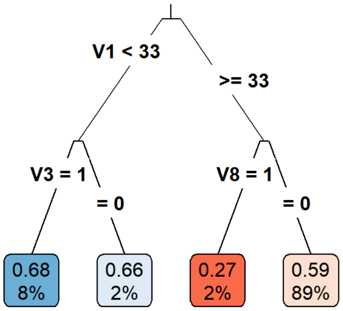

Figure 6: Tree by data in 2000

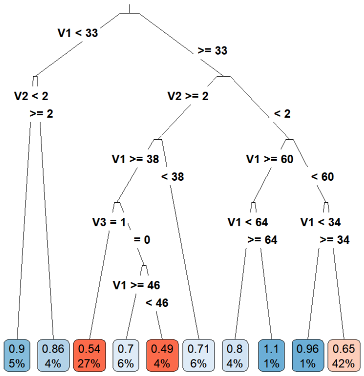

Figure 7: Tree by data in 2010

Leafs	2010		2000		1990	
	CATE Est.	Std. Error	CATE Est.	Std. Error	CATE Est.	Std. Error
Leaf1	0.493	0.158	0.265	0.206	0.448	0.036
Leaf2	0.541	0.054	0.585	0.033	0.514	0.054
Leaf3	0.651	0.042	0.661	0.197	-	-
Leaf4	0.702	0.143	0.677	0.099	-	-
Leaf5	0.708	0.117	-	-	-	-
Leaf6	0.798	0.157	-	-	-	-
Leaf7	0.860	0.163	-	-	-	-
Leaf8	0.90	0.121	-	-	-	-
Leaf9	0.96	0.221	-	-	-	-
Leaf10	1.11	0.282	-	-	-	-

Table 2: CATE Estimations for different years (across decades)

4.3 CATE across ages

This second problem of interest aims to investigate the conditional average treatment effect of college education on yearly wages for individuals in different age intervals. Specifically, we divided our sample into three groups: individuals aged between 30 to 40, 40 to 50, and 50 to 60. We utilized data from 2010 to ensure the latest possible data availability. Our analysis shows that younger age groups exhibit higher tree depth and higher estimated CATE than the older ones. Nonetheless, the differences in CATE estimates across age groups are less pronounced than those observed across decades. The average CATE estimate for individuals in their 30s, 40s, and 50s is 0.74, 0.679, and 0.645, respectively. These results indicate a decreasing trend in the effect of college education on income as individuals get older. These findings may provide valuable insights for policymakers and educators concerning the importance of college education in terms of its impact on income among different age groups.

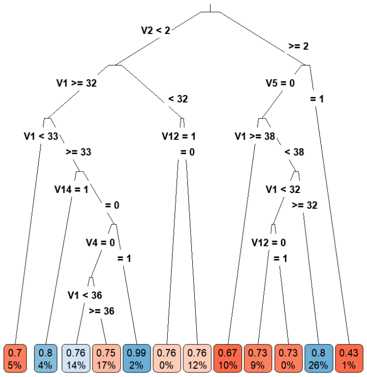

Figure 8: Tree from age 30s

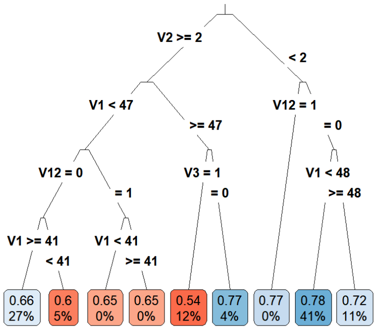

Figure 9: Tree from age 40s

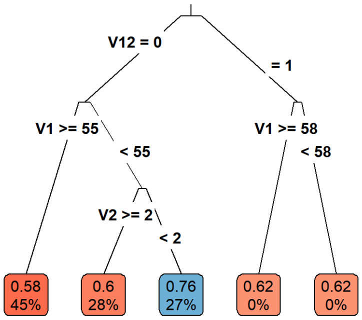

Figure 10: Tree from age 50s

Leafs	Age 30		Age 40		Age 50	
	CATE Est.	Std. Error	CATE Est.	Std. Error	CATE Est.	Std. Error
Leaf1	0.431	0.494	0.544	0.087	0.576	0.043
Leaf2	0.669	0.097	0.601	0.134	0.602	0.057
Leaf3	0.701	0.098	0.661	0.059	0.757	0.051
Leaf4	0.732	0.093	0.718	0.081	-	-
Leaf5	0.754	0.057	0.768	0.161	-	-
Leaf6	0.756	0.076	0.782	0.040	-	-
Leaf7	0.764	0.068	-	-	-	-
Leaf8	0.800	0.125	-	-	-	-
Leaf9	0.802	0.057	-	-	-	-
Leaf10	0.993	0.180	-	-	-	-

Table 3: CATE Estimations for difference age intervals in 2010

4.4 CATE by genders

Our last problem of interest aims to examine the conditional average treatment effect of college education on yearly wages for males and females. We similarly utilized data from 2010 for the latest possible data availability. Our analysis shows no significant variation in tree depth or CATE estimates by gender. The average CATE estimate for males and females in 2010 is 0.556 and 0.546, respectively. These results suggest that college education has a similar impact on income for both males and females.

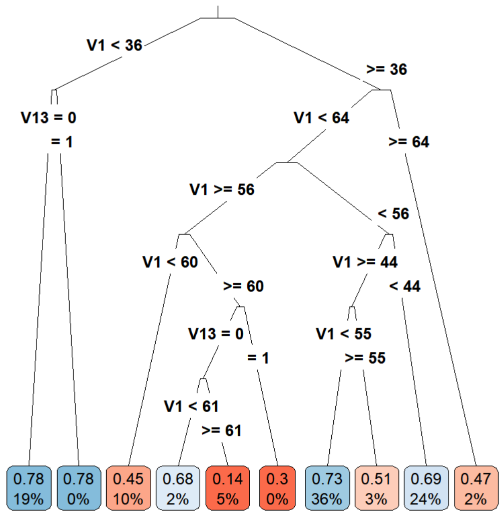

Figure 11: Tree by male's data at 2010

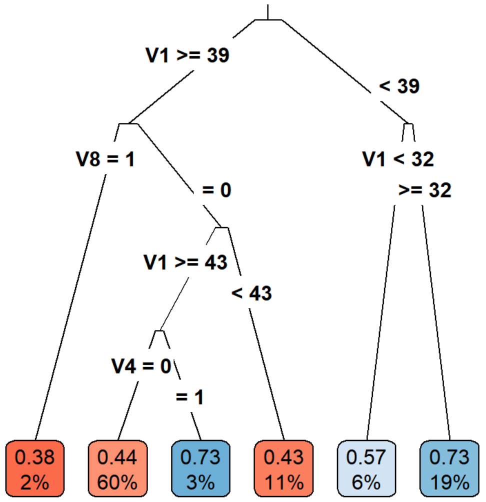

Figure 12: Tree by female's data at 2010

Leafs	Male		Female	
	CATE Est.	Std. Error	CATE Est.	Std. Error
Leaf1	0.14	0.135	0.379	0.191
Leaf2	0.45	0.087	0.430	0.090
Leaf3	0.468	0.258	0.439	0.040
Leaf4	0.511	0.177	0.567	0.116
Leaf5	0.681	0.163	0.729	0.066
Leaf6	0.694	0.052	0.730	0.201
Leaf7	0.730	0.044	-	-
Leaf8	0.777	0.053	-	-

Table 4: CATE Estimations for males and females cases in 2010

5 Forest method vs. Honest Tree

5.1 Training and Testing Details

Focusing on the three problems of interest above, we extend the practical application to forest-based method discussed in 3.2, and draw a comparison with its predicted treatment effects with the one from the tree. Given the number of trees as the only non-adaptive parameters in the random forest algorithm, we trained two models with 100 and 200 trees, and compare the mean and standard deviation of predicted treatment effects with the output from a single honest tree. Furthermore, we horizontally compare the forest-predicted CATE among the subgroups specified in the problems of interests. Empirically, for the inputted data, we sampled $50,000$ observations from the database as the training set *with no replacement*, and sampled another $5,000$ observations from the database as the test set (processed and encoded). Note that unlike in simulations or an experiment, so individuals do not undergo both treatments, and consequently we are unable to take the difference between the two treatments on the same individual i.e. the "true" individualized treatment effects. Therefore, the results are presented mostly for the purpose of CATE comparison based on our questions of interest.

5.2 CATE across decades

In Figure [13], we employed a fixed causal forest model with 100 trees to predict the conditional average treatment effects on a test set for each decade. Our analysis revealed that the predicted treatment effects did not exhibit substantial deviations across the decades. However, there were several outliers in the predicted values for the 2000s, which we attributed to the volatile global economic conditions during that period, including the emergence of global diseases and financial crises. Notably, our findings demonstrated an overall increasing trend in the treatment effects higher educations on earnings by decades, indicating an escalating demand for individuals with higher education in the job market, resulting in higher salaries for this group.

Turning to Table [5], we compared the distribution of predicted CATEs for three causal tree-based methods. We found no significant difference between the predicted CATE values of the forest model with 100 trees and the forest model with 200 trees. However, we observed a higher standard deviation in the predicted treatment effects from the honest tree method relative to the other two methods. This result suggests that ensembled models offer better stability compared to single-tree models. We emphasize that the choice of model and its hyperparameters can significantly affect the stability and performance of the model, highlighting the importance of careful selection and tuning of models for causal inference.

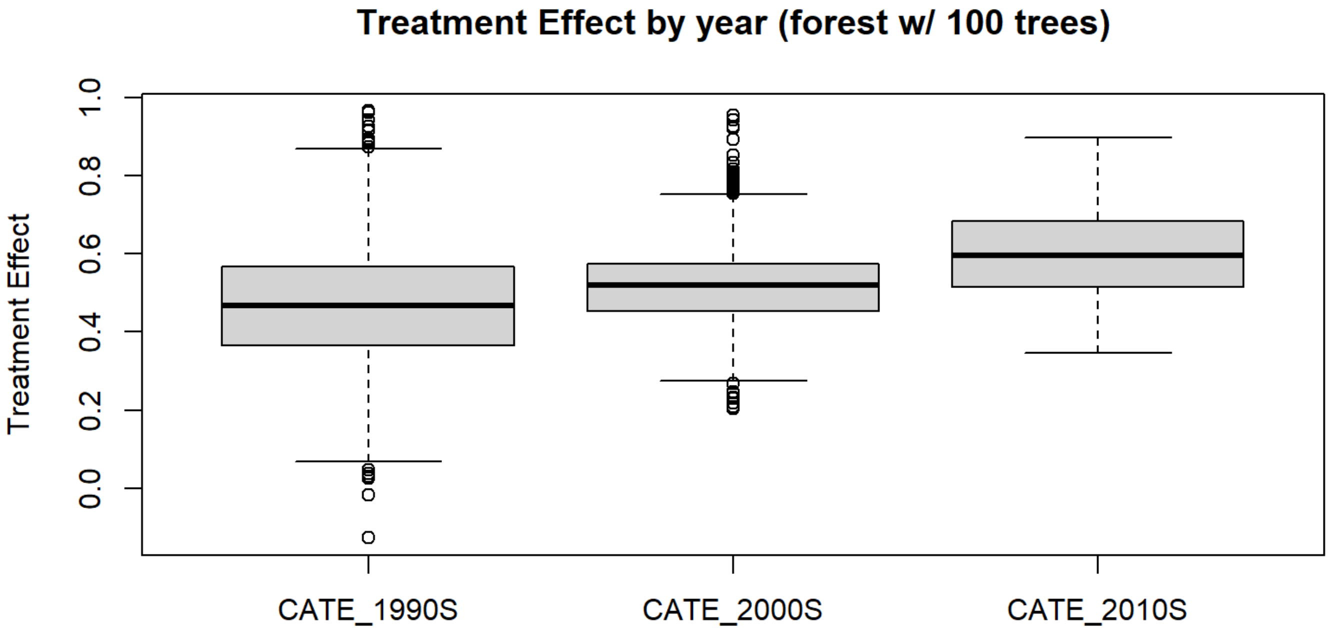

Figure 13: Treatment Effect comparison w.r.t decades

year \method	Forest (200 trees)		Forest (100 trees)		Honest Tree	
	std. pred	CATE Est.	std. pred	CATE Est.	std. pred	CATE Est.
1990s	0.162	0.457	0.166	0.457	0.078	0.567
2000s	0.089	0.519	0.092	0.522	0.097	0.579
2010s	0.115	0.584	0.119	0.602	0.156	0.625

Table 5: prediction CATE variability and ATE on test set

5.3 CATE across ages

In Figure [14], we observed a decreasing trend in the predicted treatment effects as the age of the population increased, as shown by the boxplots. This trend may be attributed to the decreasing demand for college graduates in the job market among the sub-population aged 50+ who have completed their advanced degrees. Our findings align with the analysis presented in section 5.2.

Examining Table [6], we observed a significantly larger standard deviation for the honest tree method compared to the other two methods. This result reinforces our previous finding that ensembled models, such as random forests, offer higher robustness with the bagging technique. Furthermore, we observed that the CATE estimated by the honest tree method tended to be higher than those estimated by the forest models, consistent with our previous findings. We note that the selection of model and hyperparameters can significantly impact the performance and stability of the model, underscoring the importance of careful consideration in the model selection process for causal inference.

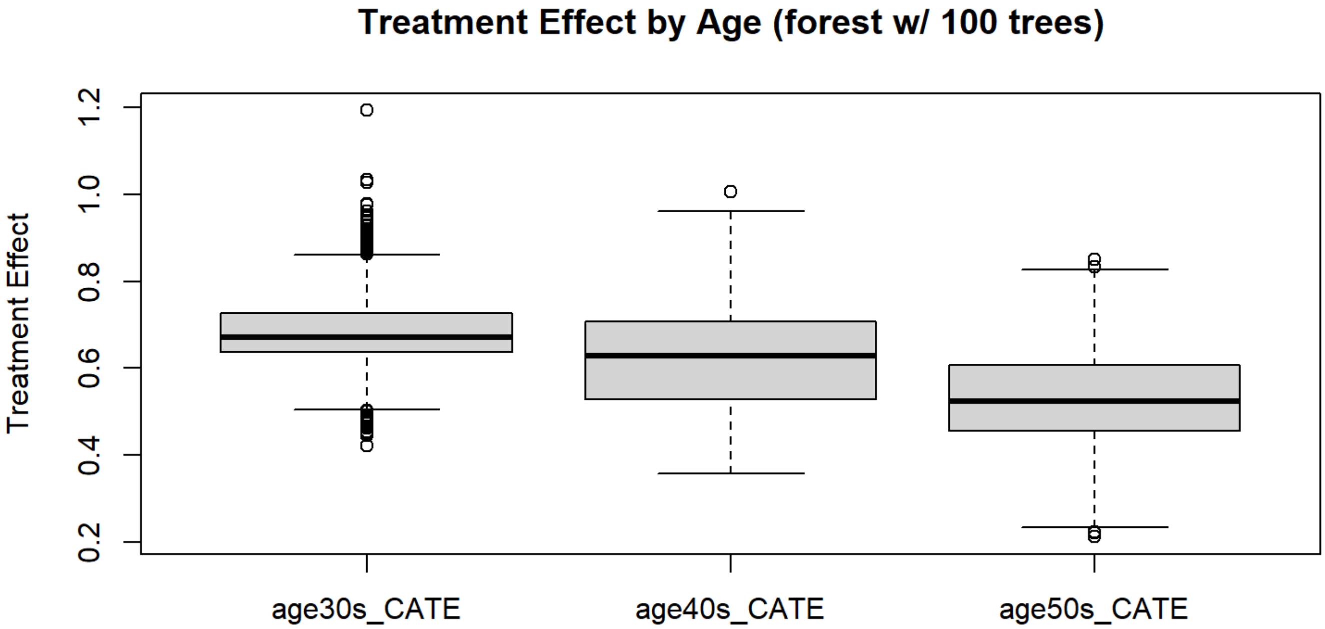

Figure 14: Treatment Effect comparison w.r.t ages

age \method	Forest (200 trees)		Forest (100 trees)		Honest Tree	
	std. pred	CATE Est.	std. pred	CATE Est.	std. pred	CATE Est.
30-40	0.079	0.68	0.074	0.706	0.105	0.715
40-50	0.085	0.61	0.099	0.635	0.14	0.65
50+	0.091	0.53	0.088	0.521	0.103	0.548

Table 6: prediction CATE variability and ATE on testset

5.4 CATE across genders

In Figure [15], we observed that the estimated treatment effects for males were higher than those for females, as evidenced by the boxplot. This finding suggests that obtaining a higher level of education leads to greater rewards for males than for females. This observation aligns with the well-documented career inequality due to gender in the United States, resulting in women receiving fewer benefits from education than men. Interestingly, we observed a greater number of outliers for females compared to males. This may be indicative of a large social class deviation among females, whereby girls from families with lower income or social status face even higher risks in investing in education. These findings underscore the need for policies and interventions that address the persistent gender and socio-economic disparities in education and the workforce.

In Table [7], we observed an unusually lower estimated average treatment effect for the honest tree method, which is in contrast to our findings in the previous sections. This result further underscores the fact that random forests tend to produce more robust estimates, while causal trees may yield CATE estimates with greater instability.

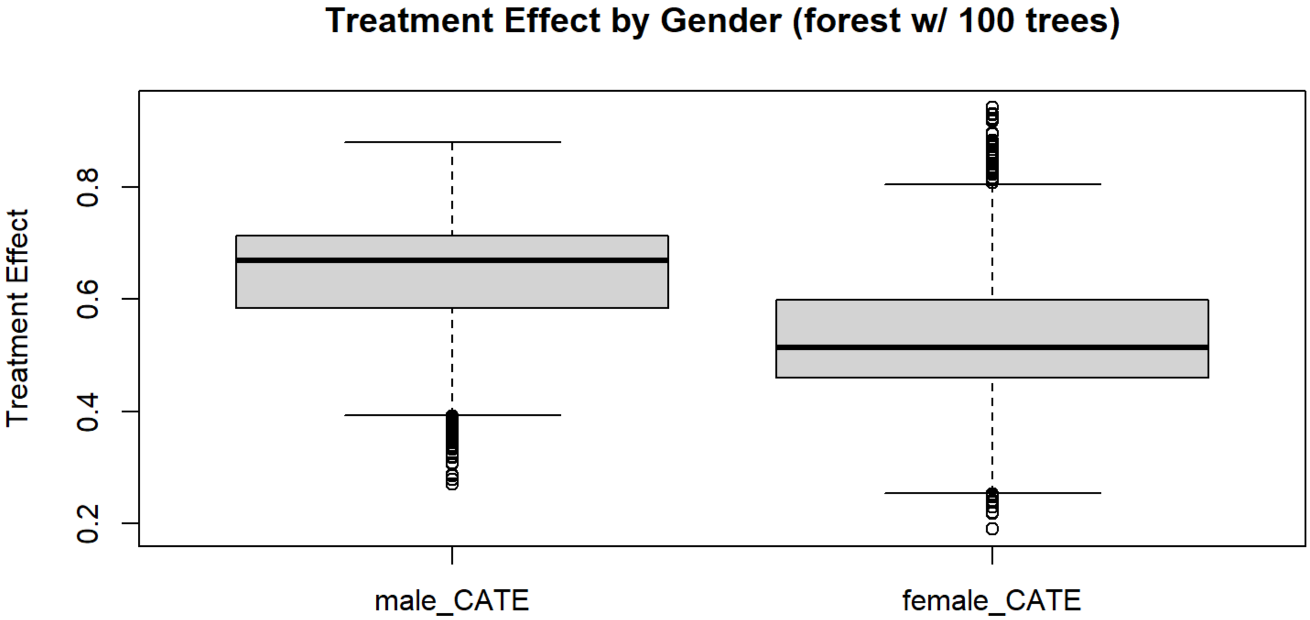

Figure 15: Treatment Effect comparison w.r.t genders

gender \method	Forest (200 trees)		Forest (100 trees)		Honest Tree	
	std. pred	CATE Est.	std. pred	CATE Est.	std. pred	CATE Est.
Male	0.094	0.638	0.097	0.64	0.139	0.629
Female	0.112	0.53	0.13	0.53	0.13	0.562

Table 7: prediction CATE variability and ATE on test set

6 Continued discussion & Summary

In sections 4 and 5, we addressed three important problems in the context of causal inference and summarized and visualized the estimated conditional average treatment effects within subpopulations using both the honest tree method and the causal forest algorithm. Our analysis revealed that the data became more complex in recent decades, as the tree was grown deeper with the inclusion of the 2010 data, compared to the 1990 data. This suggests that recent-year cohorts face more choices and challenges in the job market. Additionally, we found that trees for samples of individuals in their 30s were grown deeper, which is consistent with our previous conclusion that higher education is becoming increasingly important for job market success. Our analysis also showed that higher education had a greater impact on the earnings of young people compared to middle-aged individuals, as evidenced by the boxplots in section 5.

But there are also factors to be considered other than the finite covariates we involved. Another possible explanation for the higher treatment effects on earning in recent years will be the impact of economic and monetary changes on the data. Over the past several years, the global economy has experienced several recessions, and this has had a significant impact on many industries and individuals. For example, the value of the dollar has depreciated, which could have affected the data we collected. It's possible that this depreciation could have led to changes in purchasing power, which could have influenced the results of our analysis. Additionally, the recession could have led to changes in employment rates or income levels, which could have had an impact on the variables we studied.

In this research, we conducted a comparative analysis of various causal tree-based techniques and observed that the causal forest approach, as an ensemble method, was inclined towards producing more consistent predictions compared to those generated by an individual honest tree. Our findings highlight the criticality of selecting appropriate models and methods for performing causal inference analyses. In situations where the data are complex, and the number of confounding factors is substantial, ensemble methods such as the causal forest algorithm may yield more robust outcomes. Furthermore, it is essential to meticulously consider the hyperparameters of the chosen method to optimize its performance. However, it is pertinent to acknowledge the trade-off between time and performance when comparing ensemble methods and plain techniques. Our research involved an extensive training procedure during random forest analysis, and the training time for a single tree was already significant. Therefore, our group is excited about the algorithmic acceleration or more informative trees is the subsequent stage of this study in the future study of causal inference. While the former involves a more computer science oriented approach, the latter leans more heavily on the contributions of mathematicians and economists.

7 Appendix

Race Codes:

Code	Label
1	White
2	Black/African American
3	American Indian or Alaska Native
4	Chinese
5	Japanese
6	Other Asian or Pacific Islander
7	Other race, nec
8	Two major races
9	Three or more major races

State FIPS Codes:

Alabama	01	Missouri	29
Alaska	02	Nebraska	31
Arizona	04	Nevada	32
Arkansas	05	New Hampshire	33
California	06	New Jersey	34
Colorado	08	New Mexico	35
Connecticut	09	New York	36
Delaware	10	North Carolina	37
District of Columbia	11	North Dakota	38
Florida	12	Ohio	39
Georgia	13	Oklahoma	40
Hawaii	15	Oregon	41
Idaho	16	Pennsylvania	42
Illinois	17	Rhode Island	44
Indiana	18	South Carolina	45
Iowa	19	South Dakota	46
Kansas	20	Tennessee	47
Kentucky	21	Texas	48
Louisiana	22	Utah	49
Maine	23	Vermont	50
Maryland	24	Virginia	51
Massachusetts	25	Washington	53
Michigan	26	West Virginia	54
Minnesota	27	Wisconsin	55
Mississippi	28	Wyoming	56

References

[1] Joshua D. Angrist and Alan B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, Nov 1991.

[2] Omar Arias and Walter W McMahon. Dynamic rates of return to education in the u.s. *Economics of Education Review*, 20(2):121–138, April 2001.

[3] Orley Ashenfelter and David Zimmerman. Estimates of the returns to schooling from sibling data: Fathers, sons, and brothers. *The Review of Economics and Statistics*, 79(1):1–9, Feb 1997.

[4] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests, 2016.

[5] Albert E. Beaton. The influence of education and ability on salary and attitudes. In ed. F. Thomas Juster, editor, *Education, Income, and Human Behavior*, pages 365–396. NBER, 1975.

[6] Gary S. Becker. Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70(5), Oct 1962.

[7] Donna Bobbitt-Zeher. The gender income gap and the role of education. *Sociology of Education*, 80:1–22, January 2007.

[8] Raymond Boudon. Educational growth and economic equality. *Quality Quantity*, 9:1–10, March 1974.

[9] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[10] Edward F. Denison. Measuring the contribution of education. *The residual factor and economic growth*, 1964.

[11] Christian Dustmann and Uta Schönberg. The causal effect of education on earnings revisited. *Oxford Economic Papers*, 64(4):689–733, 2012.

[12] Bloom et al. *A Taxonomy of Educational Objectives: Handbook I The Cognitive Domain*, chapter 1. Longman, Green Co., New York, 1956.

[13] Yingjie Feng. Causal inference in possibly nonlinear factor models, 2020.

[14] Tim Genewein, Tom McGrath, Grégoire Déletang, Vladimir Mikulik, Miljan Martic, Shane Legg, and Pedro A. Ortega. Algorithms for causal reasoning in probability trees, 2020.

[15] Claudia Goldin and Lawrence F. Katz. Education and income in the early twentieth century: Evidence from the prairies. *Journal of Economic History*, 62(3):752–777, 2002.

[16] Max Goplerud, Kosuke Imai, and Nicole E. Pashley. Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis, 2022.

[17] William H. Greene. *Econometric Analysis*. Pearson Education, Inc., 2012.

[18] P. Richard Hahn, Jared S. Murray, and Carlos Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects, 2017.

[19] Colm Harmon and Ian Walker. Estimates of the economic return to schooling for the united kingdom. *American Economic Review*, 85(5):1278–1286, Dec 1995.

[20] Michael Hout. Social and economic returns to college education in the united states. *Annual Review of Sociology*, 38:379–400, Aug 2012.

[21] John B. Willett John H. Tyler, Richard J. Murnane. Estimating the labor market signaling value of the ged. *The Quarterly Journal of Economics*, 115(2):431–468, May 2000.

[22] Markus Jäntti and Ritva Reinikka-Soininen. Education and earnings in a transition economy: the case of vietnam. *World Development*, 24(7):1133–1146, 1996.

[23] Jiuyong Li, Saisai Ma, Thuc Le, Lin Liu, and Jixue Liu. Causal decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):257–271, feb 2017.

[24] Jacob A. Mincer. *Schooling, Experience, and Earnings*. NBER, 1974.

[25] Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially private estimation of heterogeneous causal effects, 2022.

[26] Ulrich Teichler. *Higher Education and the World of Work: Conceptual Frameworks, Comparative Perspectives, Empirical Findings*. Sense Publishers, 2009.

[27] Neelam Younas, Amjad Ali, Hafsa Hina, Muhammad Hamraz, Zardad Khan, and Saeed Aldahmani. Optimal causal decision trees ensemble for improved prediction and causal inference. *IEEE Access*, 10:13000–13011, 2022.