# Tree-based method towards the Estimation of the Conditional Average Treatment Effect of Education on Earning

Huaning Liu [1] Noah Simpson [1] Xue Wang [1] Wenqian Zhao [1]

[1]University of California San Diego

## Causal Analysis of Education on Income

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 career prospects and high-status income levels[3]. Therefore, We intend 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[1]. The authors created and benchmarked an unbiased estimator of CATE across subsets of the population with different treatments, proposing an "honest" approach for estimation.

## Setup

### Problem Settings

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

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

$W_i \in \{0, 1\}$ is defined to be the indicator for whether unit $i$ received the treatment or not. Since for every unit $i$ we only observe one of the values of $W$, $Y_i^{obs}$ is as follows:

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

$X_i$ is defined to be a vector of $K$ features/covariates. Now, for every unit $i$, we have observations for $(Y_i^{obs}, W_i, X_i)$, and the conditional average treatment effect (CATE) is defined by:

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

### Dataset

We queried the IPUMS database to attain U.S. census microdata from around 1980 to 2020. To start with our investigation, we processed the data for them to fit in our model:

- Treatment Variable($W$): EDUC, education level transformed to binary scale, indicating whether the individual finished college or not.
- Outcome Variable($Y$): INCWAGE, natural log of the yearly wage of an individual.
- Covariates($X$): other features extracted such as AGE, SEX, RACE, etc.

### Research Problems

To further explore the interaction between the CATE and some of the covariates, we sub-sampled and created three sub-datasets for the purpose of answering the three research problems:

1. The CATE of college education on yearly income for different years, being 2010, 2000, and 1990.
2. The CATE of college education on yearly income for males and females in 2010.
3. The CATE of college education on yearly income for different age groups in 2010, being people in their 30s, 40s, and 50s.

## Explanatory Data Analysis

We plotted the distribution of the variables relevant to our research question to get a more general sense. From the EDAs, we can observe a clear difference in wages across decades and education levels, as well as a difference in education level across ages, which motivate our incentive to investigate the heterogenous CATE of education.

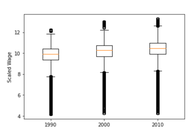


Figure 1. Wages by decades

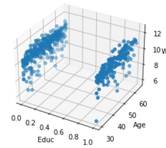


Figure 2. Educ-Age-Wage

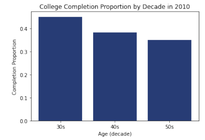


Figure 3. Complete rate

## Causal Tree & CATE Estimation

The "honest" approach used for building and validating the Causal Tree is an extension and divergence from the classification and regression trees(CART) algorithm. One of the most important concerns is to find criteria to evaluate and compare estimators for treatment effects.

Due to the fact that we cannot observe both $Y_i(1)$ and $Y_i(0)$ for an individual, the true treatment effect $\tau$ 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 $\tau_i$. Therefore, the paper proposed an estimator for $EMSE_\tau(\Pi)$ by modifying the $MSE_\mu$ 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 $-\widehat{EMSE}_\tau(S^{tr,cv}, N^{est}, \Pi)$.

## Results

### CATE across Decades

We observed that later decades would produce a deeper tree and a larger CATE of education on income. The corresponding average CATE estimation across leaves for the 1990s, 2000s, and 2010s are 0.7723, 0.547, and 0.481.

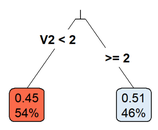


Figure 4. Tree by 1990

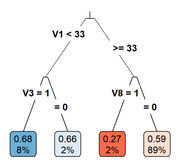


Figure 5. Tree by 2000

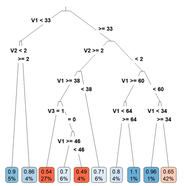

Figure 6. Tree by 2010

## Results(cont.)

### CATE across Ages

Similar to the former problem, the tree produced is deeper and the CATE is larger for the younger population, which is in our expectation. However, the difference in the tree depth and CATE is not as large as that over decades. The corresponding average CATE estimation across leaves for people in their 30s, 40s, and 50s are 0.74, 0.679, and 0.645.

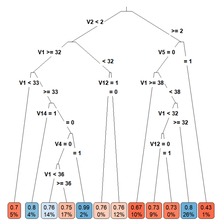


Figure 7. Tree from age 30s

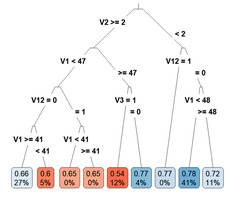


Figure 8. Tree from age 40s

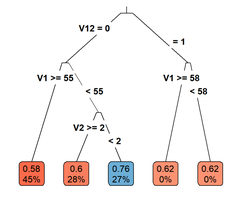


Figure 9. Tree from age 50s

### CATE across Genders

Different from the two above questions, there is no obvious difference in the tree depth and CATE across genders. The corresponding average CATE estimation across leaves for males and females in 2010 is 0.556 and 0.546.

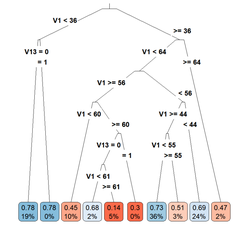


Figure 10. Tree by male's data at 2010

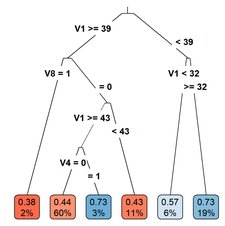


Figure 11. Tree by female's data at 2010

## Conclusions & Discussions

We have found positive CATE of education on earnings among all of the three research questions. The CATE is larger in recent decades and for younger people. Unexpectedly, there was no significant difference found between the CATE of males and females.

## References

[1] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, jul 2016.

[2] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

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