

Article

Causal Random Forests Model Using Instrumental Variable Quantile Regression

Jau-er Chen ^{1,2,*} and Chen-Wei Hsiang ³

¹ Institute for International Strategy, Tokyo International University, 1-13-1 Matobakita Kawagoe, Saitama 350-1197, Japan

² Center for Research in Econometric Theory and Applications, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 10617, Taiwan

³ Behavioral and Data Science Research Center, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 10617, Taiwan; r06323026@ntu.edu.tw

* Correspondence: jechen@tiu.ac.jp; Tel.: +81-49-232-1111

Received: 24 September 2019; Accepted: 11 December 2019; Published: 16 December 2019



Abstract: We propose an econometric procedure based mainly on the generalized random forests method. Not only does this process estimate the quantile treatment effect nonparametrically, but our procedure yields a measure of variable importance in terms of heterogeneity among control variables. We also apply the proposed procedure to reinvestigate the distributional effect of 401(k) participation on net financial assets, and the quantile earnings effect of participating in a job training program.

Keywords: quantile treatment effect; instrumental variable; quantile regression; causal machine learning; random forests

1. Introduction

Causal machine learning, which is based on two approaches: the double machine learning (DML), cf. [Chernozhukov et al. \(2018\)](#), and the generalized random forests method (GRF), cf. [Athey et al. \(2019\)](#), has been actively studied in economics in recent years. With the identification strategy of selection on observables, empirical applications have been investigated by using the aforementioned two approaches, including the works by [Gilchrist and Sands \(2016\)](#) and [Davis and Heller \(2017\)](#). When it comes to the identification strategy of selection on unobservables, few empirical papers using causal machine learning can be found in the existing literature. Those empirical applications very often lack important observed control variables or involve reverse causality, and thus researchers resort to the instrumental variable approach. Additionally, it remains unclear how the quantile treatment effect is to be estimated under the DML and GRF methods. In this paper, with the use of instrumental variables, we propose an econometric procedure for estimating quantile treatment effects based primarily on the generalized random forests of [Athey et al. \(2019\)](#).

[Chernozhukov and Hansen \(2005\)](#) propose an estimator that addresses endogeneity in quantile regressions via rank similarity, a crucial feature absent in the prior approaches. Using rank similarity, this estimator studies the heterogeneous quantile effects of an endogenous variable over the entire population (rather than for the compliers). Rank similarity thus identifies population-based quantile treatment effects, cf. [Frandsen and Lefgren \(2018\)](#). This approach does not require the monotonicity assumption used in [Abadie et al. \(2002\)](#) and allows for binary or continuous endogenous and instrumental variables. [Chernozhukov and Hansen \(2008\)](#) create a bridge between two-stage least squares (2SLS) estimator and

their 2005 estimator, and propose an estimator robust to weak instruments. However, it is noteworthy that these estimator are unable to estimate unconditional quantiles, which are, as discussed in [Guilhem et al. \(2019\)](#), quantities that should be of utmost interest to empirical researchers. In this paper, we use the instrumental variable quantile regression of [Chernozhukov and Hansen \(2008\)](#) as a vehicle for identifying the quantile treatment effect.

[Athey and Imbens \(2016\)](#) is the first paper that develops the regression tree model to estimate heterogeneous treatment effects using the honest splitting algorithm. [Wager and Athey \(2018\)](#) extend the regression tree model to causal forests. Recently, [Athey et al. \(2019\)](#) have developed the generalized random forests model, which is a unified framework in the sense that it is built on local moment conditions capable of encompassing many models. Therefore, we bring the first order condition of the instrumental variable quantile regression into the local moment conditions and then modify the GRF algorithm. Accordingly, the quantile treatment effect can be estimated under the framework of causal random forests. Thus, our proposed estimator and the generalized random forests model both share the advantage of estimating the conditional quantile treatment effect nonparametrically.

[Chen and Tien \(2019\)](#) investigate the instrumental variable quantile regression in the context of double machine learning. Although related to their paper, our procedure is not considering the same high-dimensional setting. Further, in contrast to the DML for instrumental variable quantile regressions, the proposed econometric procedure yields a measure of variable importance in terms of heterogeneity among control variables. The pattern of variable importance across quantiles can be revealed as well. We highlight the usage of exploring variable importance by reinvestigating two empirical studies - the distributional effect of 401(k) participation on net financial assets, and the quantile effect of participating in a job training program on earnings.

The rest of the paper is organized as follows. The model specification and practical algorithm are introduced in Section 2. Section 3 presents the measure of variable importance. Section 4 presents two empirical applications. Section 5 concludes the paper. The Appendix A discusses the usage of a doubly robust method along with the causal random forests structure for achieving more efficient estimation. The Appendix A also discusses the identifying restrictions and regularity conditions for the instrumental variable quantile regression and the generalized random forests, and further verifies conditions for establishing consistency and asymptotic normality of the proposed estimator.

2. The Model and Algorithm

We propose the causal random forests with the instrumental variable quantile regression (GRF-IVQR, hereafter). Estimation procedure of the GRF-IVQR is constructed as below, essentially based on the method developed in [Athey et al. \(2019\)](#).

2.1. Generalized Random Forests

The classification tree and regression tree (CART) and its extension, random forests [Breiman \(2001\)](#), are effective methods for flexibly estimating regression functions in terms of out-of-sample predictive power. Random forests have become particularly popular methods. A key attraction is that they require relatively little tuning and have superior performance to more complex methods such as deep learning neural networks, cf. Section 3.2 of [Athey and Imbens \(2019\)](#). Recently, random forests have garnered interest and have been extended to causal effects; that is, the generalized random forests estimator.

In what follows, we describe how we incorporate the instrumental variable quantile regression into the framework of GRF and modify the resulting estimator accordingly.

Given data $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$, we estimate the parameter of interest $\theta(x)$ via the following moment conditions

$$\mathbb{E}[\psi_{\theta(x), \nu(x)}(O_i) \mid X_i = x] = 0 \text{ for all } x \in \mathcal{X},$$

where $\psi(\cdot)$ stands for the score function and $\nu(x)$ are optional nuisance parameters. The above moment conditions, similar to the generalized method of moments (GMM), can be used to identify many objects of interest from an economic perspective. We seek forest-based estimates, $\hat{\theta}(x)$, which are the conditional quantile treatment effects, in the context of instrumental variable quantile regressions.

Chernozhukov and Hansen (2005) laid the theoretical foundations for the instrumental variable quantile regression (IVQR). With outcome Y_i , endogenous treatment variable D_i , instrumental variable Z_i , and control variables X_i , the IVQR can be represented as the following moment conditions

$$\begin{aligned} \mathbb{E}[\psi_{\theta(\tau), \nu(\tau)}(Y_i) \mid \{D_i, X_i, Z_i\}] \\ = \mathbb{E}[\{\tau - 1(Y_i \leq D_i\theta(\tau) + X_i\nu(\tau))\}(Z_i, X_i)' \mid \{D_i, X_i, Z_i\}], \end{aligned}$$

where $\theta(\tau)$ is the conditional quantile treatment effect, $\nu(\tau)$ are the nuisance parameters, $1(\cdot)$ is the indicator function, and τ is a quantile index.

The sample counterpart of the local moment conditions and the estimator of θ are introduced by Athey et al. (2019) and defined as below.

$$(\hat{\theta}(\tau, x), \hat{\nu}(\tau, x)) \in \underset{\theta(\tau), \nu(\tau)}{\operatorname{argmin}} \left\{ \left\| \sum_{i=1}^n \alpha_i(x) \psi_{\theta(\tau), \nu(\tau)}(Y_i) \right\|_2 \right\},$$

where $\alpha_i(x)$ are tree-based weights averaged by the forest, which measure how often each training example falls in the same leaf as x . In other words, these weights represent the relevance of each sample when we estimate θ . Specifically, the weights are obtained by a forest-based algorithm. For the point of interest x , let $L_b(x)$ represent the set of samples which fall in the same terminal leaf and contain x in b th tree, where $b \in \{1, 2, \dots, B\}$. That is to say, the weight $\alpha_i(x)$ of each sample for the point of interest x will be the frequency with which the i th sample is in the same terminal leaf among all trees $\{1, 2, \dots, B\}$. That is,

$$\begin{aligned} \alpha_{bi}(x) &= \frac{1(X_i \in L_b(x))}{|L_b(x)|}, \\ \alpha_i(x) &= \frac{1}{B} \sum_{b=1}^B \alpha_{bi}(x). \end{aligned}$$

With such forest-based weights and a pre-specified quantile index τ , we minimize the criterion function constructed using sample moment conditions, and then an estimate of the conditional quantile treatment effect $\hat{\theta}(\tau)$ is obtained. In the subsequent section, we discuss how to grow the trees and the forests with the instrumental variable quantile regression.

2.2. Tree Splitting Rules

Growing a tree is a recursive binary splitting process. The spirit of the tree-based algorithm is to split the data in the parent node P in half by maximizing the heterogeneity of the associated two children nodes $\{C_1, C_2\}$.

Specifically, for node j with data \mathcal{J} , we define the node parameters as follows.

$$(\hat{\theta}_j(\tau), \hat{\nu}_j(\tau)) (\mathcal{J}) \in \operatorname{argmin}_{\theta(\tau), \nu(\tau)} \left\{ \left\| \sum_{i \in \mathcal{J}: X_i \in j} \psi_{\theta(\tau), \nu(\tau)}(Y_i) \right\|_2 \right\},$$

where $j \in \{P, C\}$. In each node, we minimize the following criterion

$$\operatorname{err}(C_1, C_2) = \sum_{j=1,2} \mathbb{P}[X \in C_j \mid X \in P] \cdot \mathbb{E} \left[\left(\hat{\theta}_{C_j}(\tau) - \theta(\tau, X) \right)^2 \mid X \in C_j \right],$$

which is based on the GRF method. However, the minimization is infeasible due to the unknown value of $\theta(\tau, X)$. [Athey et al. \(2019\)](#) turn this minimization problem of $\operatorname{err}(C_1, C_2)$ into an accessible model-free maximization problem of

$$\Delta(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} (\hat{\theta}_{C_1}(\tau) - \hat{\theta}_{C_2}(\tau))^2,$$

where n_{C_1}, n_{C_2}, n_P are numbers of observations in children and parent nodes. Along the way of maximizing Δ , the $\theta_{C_j}(\tau)$ is estimated by the IVQR with respect to all possible splittings which correspond to the set $\{\text{all possible values for } X_i, \forall i\}$. Consequently, the estimation is computationally infeasible. To circumvent this difficulty, [Athey et al. \(2019\)](#) suggest a gradient tree algorithm which maximizes an approximate criterion $\tilde{\Delta}(C_1, C_2)$. In what follows, with two new ingredients A_p and ρ defined below, we construct $\tilde{\Delta}(C_1, C_2)$ step by step.

We first define A_p as the gradient of the expectation of the moment condition.

$$\begin{aligned} A_p &= \nabla \mathbb{E} \left[\psi_{\hat{\theta}_P(\tau), \hat{\nu}_P(\tau)}(Y_i) \mid \{D_i, X_i, Z_i\} \in P \right] \\ &= \nabla \mathbb{E} \left[(\tau - 1(Y_i \leq D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau))) (Z_i, X_i)' \mid \{D_i, X_i, Z_i\} \in P \right] \\ &= \nabla \left[(\tau - F(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau))) (Z_i, X_i)' \mid \{D_i, X_i, Z_i\} \in P \right] \\ &= \nabla \begin{bmatrix} \tau Z_i - F(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau)) Z_i \\ \tau X_{1i} - F(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau)) X_{1i} \\ \vdots \\ \tau X_{mi} - F(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau)) X_{mi} \end{bmatrix} \mid \{D_i, X_i, Z_i\} \in P \end{aligned},$$

where $F(\cdot)$ is a cumulative distribution function, and m is the dimension of X . For simplicity of derivation, we fix the following notations. $g_0 := \tau Z_i - F(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau)) Z_i$, $g_1 := \tau X_{1i} - F(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau)) X_{1i}$, \dots , $g_m := \tau X_{mi} - F(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau)) X_{mi}$, and we suppress $[\cdot \mid \{D_i, X_i, Z_i\} \in P]$ which means the estimation is conditional on the parent node. Accordingly, A_p can be written as the gradient of g_0, g_1, \dots, g_m with respect to the parent node parameters.

$$\begin{aligned}
A_p &= \nabla_{\hat{\theta}_p(\tau), \hat{\nu}_p(\tau)} \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_m \end{bmatrix}' = \begin{bmatrix} \frac{\partial g_0}{\partial \hat{\theta}_p(\tau)} & \frac{\partial g_1}{\partial \hat{\theta}_p(\tau)} & \frac{\partial g_2}{\partial \hat{\theta}_p(\tau)} & \cdots & \frac{\partial g_m}{\partial \hat{\theta}_p(\tau)} \\ \frac{\partial g_0}{\partial \hat{\nu}_{1,p}(\tau)} & \frac{\partial g_1}{\partial \hat{\nu}_{1,p}(\tau)} & \frac{\partial g_2}{\partial \hat{\nu}_{1,p}(\tau)} & \cdots & \frac{\partial g_m}{\partial \hat{\nu}_{1,p}(\tau)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_0}{\partial \hat{\nu}_{m,p}(\tau)} & \frac{\partial g_1}{\partial \hat{\nu}_{m,p}(\tau)} & \frac{\partial g_2}{\partial \hat{\nu}_{m,p}(\tau)} & \cdots & \frac{\partial g_m}{\partial \hat{\nu}_{m,p}(\tau)} \end{bmatrix} \\
&= \begin{bmatrix} -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) Z_i D_i & -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) X_{1i} D_i & \cdots & -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) X_{mi} D_i \\ -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) Z_i X_{1i} & -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) X_{1i} X_{1i} & \cdots & -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) X_{mi} X_{1i} \\ \vdots & \vdots & \ddots & \vdots \\ -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) Z_i X_{mi} & -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) X_{1i} X_{mi} & \cdots & -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) X_{mi} X_{mi} \end{bmatrix} \\
&= -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) \begin{bmatrix} Z_i D_i & X_{1i} D_i & \cdots & X_{mi} D_i \\ Z_i X_{1i} & X_{1i} X_{1i} & \cdots & X_{mi} X_{1i} \\ \vdots & \vdots & \ddots & \vdots \\ Z_i X_{mi} & X_{1i} X_{mi} & \cdots & X_{mi} X_{mi} \end{bmatrix},
\end{aligned}$$

where $f(\cdot)$ is the probability density function of $F(\cdot)$. Therefore, the inverse of A_p ,

$$\begin{aligned}
A_p^{-1} &= \left\{ -f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau)) \begin{bmatrix} Z_i D_i & X_{1i} D_i & \cdots & X_{mi} D_i \\ Z_i X_{1i} & X_{1i} X_{1i} & \cdots & X_{mi} X_{1i} \\ \vdots & \vdots & \ddots & \vdots \\ Z_i X_{mi} & X_{1i} X_{mi} & \cdots & X_{mi} X_{mi} \end{bmatrix} \right\}^{-1} \\
&= \frac{-1}{f(D_i \hat{\theta}_p(\tau) + X_i \hat{\nu}_p(\tau))} \begin{bmatrix} Z_i D_i & X_{1i} D_i & \cdots & X_{mi} D_i \\ Z_i X_{1i} & X_{1i} X_{1i} & \cdots & X_{mi} X_{1i} \\ \vdots & \vdots & \ddots & \vdots \\ Z_i X_{mi} & X_{1i} X_{mi} & \cdots & X_{mi} X_{mi} \end{bmatrix}^{-1}.
\end{aligned}$$

We then construct the pseudo-outcomes,

$$\rho_i = -\xi^\top A_p^{-1} \psi_{\hat{\theta}_p(\tau), \hat{\nu}_p(\tau)}(Y_i),$$

where ξ is a vector that picks out the $\theta(\tau)$ -coordinate from the $(\theta(\tau), \nu(\tau))$ vector. In the case with one treatment variable D , the ξ vector is $(1, 0, \dots, 0)$. Thus, the corresponding pseudo-outcomes are

$$\begin{aligned}
\rho_i &= - \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}' \left\{ \frac{-1}{f(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau))} \begin{bmatrix} \frac{1}{\#_C} \sum_{i \in C} (Z_i D_i) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} D_i) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} D_i) \\ \frac{1}{\#_C} \sum_{i \in C} (Z_i X_{1i}) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} X_{1i}) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} X_{1i}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\#_C} \sum_{i \in C} (Z_i X_{mi}) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} X_{mi}) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} X_{mi}) \end{bmatrix}^{-1} \right\} \\
&\quad \times \{ [\tau - 1(Y_i \leq D_i \theta(\tau) + X_i \nu(\tau))] (Z_i, X_i)' \} \\
&= \frac{1}{f(D_i \hat{\theta}_P(\tau) + X_i \hat{\nu}_P(\tau))} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}' \begin{bmatrix} \frac{1}{\#_C} \sum_{i \in C} (Z_i D_i) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} D_i) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} D_i) \\ \frac{1}{\#_C} \sum_{i \in C} (Z_i X_{1i}) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} X_{1i}) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} X_{1i}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\#_C} \sum_{i \in C} (Z_i X_{mi}) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} X_{mi}) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} X_{mi}) \end{bmatrix}^{-1} \\
&\quad \times \left\{ [\tau - 1(Y_i \leq D_i \theta(\tau) + X_i \nu(\tau))] \begin{bmatrix} Z_i \\ X_{1i} \\ \vdots \\ X_{mi} \end{bmatrix} \right\},
\end{aligned}$$

where the $\#_C$ denotes the number of observations in the children node C . The splitting rule is to maximize the following approximate criterion

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^2 \frac{1}{|\{i : X_i \in C_j\}|} \left(\sum_{\{i : X_i \in C_j\}} \rho_i \right)^2.$$

Notice that since some terms in ρ_i , such as $f(\cdot)$, do not affect the optimization of $\tilde{\Delta}(C_1, C_2)$, the ρ_i can be further simplified as follows.

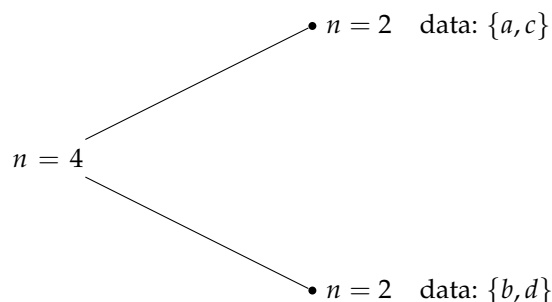
$$\begin{aligned}
\rho_i &= \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}' \begin{bmatrix} \frac{1}{\#_C} \sum_{i \in C} (Z_i D_i) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} D_i) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} D_i) \\ \frac{1}{\#_C} \sum_{i \in C} (Z_i X_{1i}) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} X_{1i}) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} X_{1i}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\#_C} \sum_{i \in C} (Z_i X_{mi}) & \frac{1}{\#_C} \sum_{i \in C} (X_{1i} X_{mi}) & \cdots & \frac{1}{\#_C} \sum_{i \in C} (X_{mi} X_{mi}) \end{bmatrix}^{-1} \\
&\quad \times \left\{ [1(Y_i > D_i \theta(\tau) + X_i \nu(\tau))] \begin{bmatrix} Z_i \\ X_{1i} \\ \vdots \\ X_{mi} \end{bmatrix} \right\}.
\end{aligned}$$

Using the modified ρ_i above, $\tilde{\Delta}(C_1, C_2)$ is our splitting rule for the instrumental variable quantile regression within the framework of generalized random forests. Based on the splitting rule, the tree is grown by recursively partitioning the data until a stopping criterion is met, cf. Section 2.4.

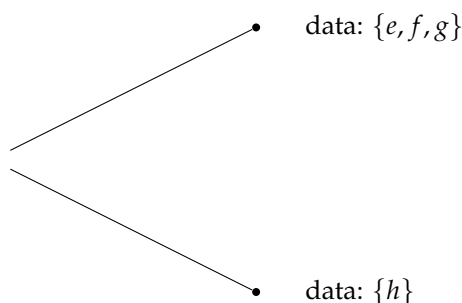
2.3. The Algorithm and an Example Illustrating Weights Calculation

With the splitting rule established, we can now grow the entire forests. In [Athey and Imbens \(2016\)](#) and [Wager and Athey \(2018\)](#), the concept of honest estimation is introduced, which is also included in the generalized random forests model. A model is honest if the information for the model construction and estimation is not the same. In the tree-forming case, the honesty here is consider as a sub-sample splitting between tree forming and weight calculation.

Here is an example of the implementation of honest estimation. Suppose we have eight samples in our data J , where $J = \{a, b, c, d, e, f, g, h\}$. We split the sample in half honestly, and we have two sub-samples $J_1 = \{a, b, c, d\}$ for tree forming and $J_2 = \{e, f, g, h\}$ weight calculation. By the splitting rule, we can construct the following tree with $J_1 = \{a, b, c, d\}$,



Next, we identify where the data of $J_2 = \{e, f, g, h\}$ is located in the tree.



Then we use this information to calculate the frequency and obtain the weights. Suppose we do not have any out of sample points of interest, we use each of the eight samples as point of interest, one at a time. If the point of interest is $\{a\}$, since a is in the same leaf with $\{e, f, g\}$, samples $\{e, f, g\}$ each gets $\frac{1}{3}$ of weight, $\{a, b, c, d, h\}$ get 0 of weight. If the point of interest is $\{b\}$, since b is in the same leaf with $\{h\}$, sample $\{h\}$ gets 1 of weight, $\{a, b, c, d, e, f, g\}$ get 0 of weight. By utilizing this method, we can get the weight for all data points. The following is the weight matrix for the above 1-tree model,

Point of interest	a	b	c	d	e	f	g	h
Weight for sample a	0	0	0	0	0	0	0	0
Weight for sample b	0	0	0	0	0	0	0	0
Weight for sample c	0	0	0	0	0	0	0	0
Weight for sample d	0	0	0	0	0	0	0	0
Weight for sample e	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
Weight for sample f	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
Weight for sample g	$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
Weight for sample h	0	1	0	1	0	0	0	1

Athey et al. (2019) prove that with proper honest sub-sampling rate and regularity conditions, the generalized random forests estimator $\hat{\theta}(x)$ is consistent and asymptotic normal to $\theta(x)$.

To build the random forests with honest tree, we first randomly select $\frac{1}{2}$ of sample for each tree. Then in each tree, we use $\frac{1}{2}$ subsampling rate for honest splitting. For the average quantile treatment effect, we adopt each point in the data as the point of interest using their own weights one by one and get the average of all results.

2.4. Practical Implementation

When implementing the generalized random forests algorithm, we first obtain baseline grids through the conventional IVQR estimator, and then utilize those grids to grow the tree. With the IVQR estimator $\hat{\gamma}_{pre}$ and its standard error $\hat{\sigma}_{pre}$, we construct the interval $[\hat{\gamma}_{pre} - 3\hat{\sigma}_{pre}, \hat{\gamma}_{pre} + 3\hat{\sigma}_{pre}]$. We divide this interval into 100 equal parts, and then obtain the baseline grid

$$\text{baseline grid} = [\hat{\gamma}_{pre} - 3\hat{\sigma}_{pre}, \dots, \hat{\gamma}_{pre} + 3\hat{\sigma}_{pre}].$$

For tree $b \in \{1, 2, \dots, B\}$ in the random forest estimation, half of the data is randomly selected. Consequently, we should reconstruct the grid for each tree. Similarly, we build the grid for the tree b

$$\text{grid for the tree } b = [\hat{\gamma}_{tree_b} - 3\hat{\sigma}_{tree_b}, \dots, \hat{\gamma}_{tree_b} + 3\hat{\sigma}_{tree_b}],$$

which is obtained via the randomly selected half of data in the tree b .

Following the concept of honest estimation, we further split the data into two parts denoted as data J_1 and J_2 . Data J_1 is used to grow the tree, and data J_2 is used to form the weight α_i . As to grow the tree with data J_1 , in what follows, we outline the splitting process in each node. We first estimate the parent node parameters $\hat{\theta}_P(\tau)$ and $\hat{\nu}_P(\tau)$ by optimizing

$$(\hat{\theta}_P(\tau, x), \hat{\nu}_P(\tau, x)) \in \underset{\theta_P(\tau), \nu_P(\tau)}{\operatorname{argmin}} \left\{ \left\| \sum_{\text{data in parent node}} \psi_{\theta(\tau)_P, \nu(\tau)_P}(Y_i) \right\|_2 \right\}$$

with the grid for the tree b . We then implement the splitting criterion

$$\max \tilde{\Delta}(C_1, C_2) = \sum_{j=1}^2 \frac{1}{|\{i : X_i \in C_j\}|} \left(\sum_{\{i : X_i \in C_j\}} \rho_i \right)^2$$

for every split.

The tree keeps splitting recursively until they reach the minimum-node-size constraint or a situation that the data in the parent node has little variation, therefore further splitting is infeasible. These two practical stopping criteria on splitting suffice for reasonable estimates.

Regarding estimation of the weight, we first identify where the observations in J_2 will be located in the tree constructed by the data J_1 . Using the algorithm discussed in the Section 2.3, we compute the weight for every data point. Accordingly, we have determined the estimation of growing a tree b .

By growing a total of B trees and averaging the weight in each tree, we obtain the weight of each observation. With the weight $\alpha_i(x)$, we estimate the conditional local quantile treatment effect

$$(\hat{\theta}(\tau, x), \hat{\nu}(\tau, x)) \in \underset{\theta(\tau), \nu(\tau)}{\operatorname{argmin}} \left\{ \left\| \sum_{i=1}^n \alpha_i(x) \psi_{\theta(\tau), \nu(\tau)}(Y_i) \right\|_2 \right\}.$$

To yield the local quantile treatment effect, we could average all x -pointwise conditional local quantile treatment effects. However, the averaging procedure can be further modified to get more efficient estimates, which is discussed in the Appendix A. Nevertheless, our empirical studies in Section 4 suggest that with a proper sampled data, the aforementioned practical procedure performs substantially well.

3. Variable Importance

Athey et al. (2019) and the associated grf R package develop a measurement for sorting variable importance which is a unique advantage of tree-based models. To explore the variable importance across quantiles, we adopt their measure of importance reproduced as follows.

$$\text{Importance}_i = \frac{\sum_{l=1}^{\text{max.depth}} \left(\frac{\sum_{b=1}^B \text{number of splitting in layer } l \text{ for } x_i \text{ in tree } b}{\sum_{b=1}^B \text{total number of splitting in layer } l \text{ in tree } b} \right) \cdot l^{-2}}{\sum_{l=1}^{\text{max.depth}} l^{-2}},$$

where the number of maximum depth is pre-specified by empirical researchers. Specifically, this measure of variable importance only considers the splitting frequency for variable X_i in trees $b = 1, \dots, B$.

This version of importance measurement shares similarity with the Gini importance widely used in random forests. Therefore, both algorithms prefer continuous variables since they have more potential splitting chances compared to binary variables. We thus shall be cautious when interpreting variable importance between a continuous variable and a categorical variable. Another important remark is that we should not conclude a particular covariate is unrelated to treatment effects simply because the tree did not split on it. There can be many different ways to pick out a subgroup of units with high or low treatment effects. Thus by comparing the average characteristics of units with high treatment effects to those with low treatment effects, researchers could obtain a fuller picture of the differences between these groups across all covariates.

Similar to the R-squared, variable importance signifies whether a variable yields enough explanatory power to the outcome variable in light of variation. Variable importance can also be used for model selection. In recent literature, e.g., O'Neill and Weeks (2018), researchers adopt variable importance measurement for policy making. Given hundreds of variables, the forest-based algorithm picks out important variables, which suffices for policy makers to identify their benchmark models.

4. Empirical Studies

In this section, we reinvestigate two empirical studies on quantile treatment effects: the effect of 401(k) participation on wealth, cf. Chernozhukov and Hansen (2004), and the effect of job training program participation on earnings, cf. Abadie et al. (2002). Not only does this conduct data-driven robustness checks on the econometric results, but the GRF-IVQR yields a measure of variable importance in terms of heterogeneity among control variables. This complements the existing empirical findings. In addition, we compare our empirical results with those from Chen and Tien (2019), the IVQR estimation based on the double machine learning approach, which is an alternative in causal machine learning literature.

As a critical note, we do not estimate and report the conditional quantile treatment effect (CQTE) in the applications. When the outcome level has an impact on the effect size and the conditional outcome variable are heterogeneous, then the CQTE could report spurious heterogeneity; see comprehensive summary of the problem in Strittmatter (2019). The same problem carries through to the importance measure. Therefore, the variable importance has to be interpreted with caution across different quantiles.

4.1. The 401(k) Retirement Savings Plan

Examining the effects of 401(k) plans on accumulated wealth is an issue of long-standing empirical interest. For example, based on the identification of selection on observables, [Chiou et al. \(2018\)](#) and [Chernozhukov and Hansen \(2013\)](#) suggest that the income nonlinear effect exists in the 401(k) study. Nonlinear effects from other control variables are identified as well. Few papers, however, investigate variable importance among control variables, cf. [Chen and Tien \(2019\)](#). In addition to estimating the quantile treatment effect of 401(k) participation, we fully explore variable importance across the conditional quantiles of accumulated wealth in light of the generalized random forests. The corresponding findings shed some light on the existing literature.

The data with 9915 observations are from the 1991 Survey of Income and Program Participation. The outcome variable is the net financial asset. The treatment variable is a binary variable standing for participation in the 401(k) plan. The instrument is an indicator for being eligible to enroll in the 401(k) plan. Control variables consist of age, income, family size, education, marital status, two-earner status, defined benefit pension status, individual retirement account (IRA) participation status, and homeownership status, which follow the model specification used in [Chernozhukov and Hansen \(2004\)](#).

Table 1 signifies that the quantile treatment effects estimated by the GRF-IVQR are similar to those calculated in [Chernozhukov and Hansen \(2004\)](#). The 401(k) participation has larger positive effects on net financial assets for people with higher savings propensity which corresponds to the upper conditional quantiles. The estimated treatment effects show a monotonically increasing pattern across the conditional distribution of net financial assets. Thus, the pattern identified by [Chernozhukov and Hansen \(2004\)](#) is assured through our data-driven robustness checks.

Table 1. Quantile treatment effects of the 401(k) participation on wealth.

	Quantile				
	0.10	0.25	0.50	0.75	0.90
CH	3209.209 (438.523)	3566.567 (525.499)	5523.524 (613.129)	9134.635 (1004.546)	14768.270 (2971.518)
GRF-IVQR	3117.674 (602.872)	3251.794 (653.277)	5547.822 (735.644)	10377.530 (892.624)	15410.360 (2078.207)

Note: GRF-IV: 11090.305 (1441.989). The GRF-IV stands for the 2SLS in the context of generalized random forests. CH and GRF-IVQR stand for, respectively, Chernozhukov and Hansen (2004) and our estimator. Numbers in parentheses are standard errors.

Based on the measure of variable importance introduced in Section 3, Table 2 and Figure 1 depict that income, age, education, and family size are the first four important variables in the analysis¹. On average, income and age are the most important variables accounting for heterogeneity, which lead to values of the variable importance 64.4% and 15.6%, respectively. We should interpret the variable importance measure with caution, because researchers could reduce the importance measure of one variable by adding a highly correlated additional variable to the model. Accordingly, in this case, the two highly correlated variables have to share the sample splits. However, even with the caution mentioned above, we now have an additional dimension, τ , which suffices to compare variable importance across quantiles. Particularly, the importance of age variable increases as the savings propensity (quantile index) goes up. The importance

¹ Following the default setting of the grf package, we set the max.depth equal to 4.

of income variable, however, decreases across conditional distribution of net financial assets. In addition, these four variables are also identified as important in the context of double machine learning, cf. [Chen and Tien \(2019\)](#).

Table 2. Variable importance.

	GRF-IV	GRF-IVQR at Specific Quantiles				
		0.10	0.25	0.50	0.75	0.90
Age	0.15607	0.17604	0.10666	0.19401	0.33202	0.48203
Income	0.64426	0.74348	0.83784	0.76596	0.62151	0.42814
Education	0.10005	0.03984	0.01790	0.01131	0.01310	0.04715
Family size	0.02908	0.02614	0.01638	0.01099	0.01244	0.02952
Married	0.00577	0.00288	0.00166	0.00317	0.00267	0.00348
Two-earner	0.01447	0.00349	0.00813	0.00619	0.00773	0.00352
Defined benefit pension	0.02110	0.00060	0.00035	0.00011	0.00042	0.00048
Participation in IRA	0.02032	0.00346	0.00655	0.00292	0.00278	0.00115
Home owner	0.00890	0.00408	0.00453	0.00535	0.00733	0.00453

Note: The GRF-IV stands for the 2SLS in the context of generalized random forests.

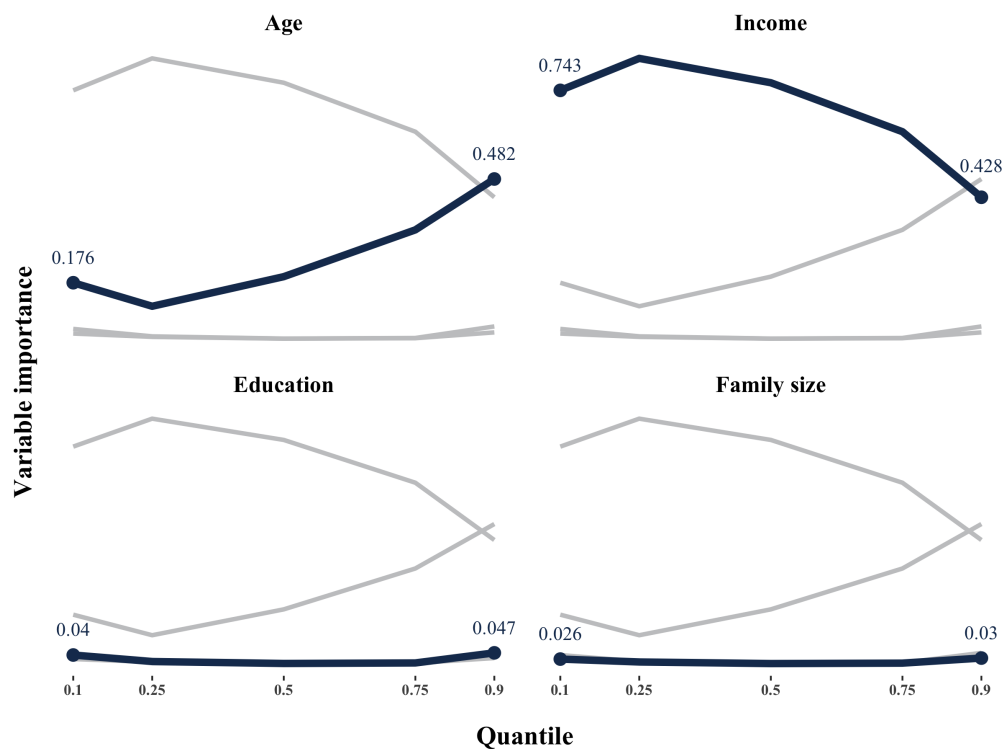


Figure 1. Variable importance across quantiles.

4.2. The Job Training Program

Abadie et al. (2002) use the Job Training Partnership Act (JTPA) data to estimate the quantile treatment effect of job training on the earning distribution². The data is from Title II of the JTPA in the early 1990's, which consists of 11,204 samples, 5102 of which are male, and 6102 of which are female. In the estimation, they take 30-month earnings as the outcome variable, enrollment for JTPA service as the treatment variable, and a randomized offer of JTPA enrollment as the instrumental variable. Control variables include education, race, marital status, previous year work status, job training service strategies, age, and whether earnings data is from the second follow-up survey. In the female group, an additional control, aid to families with dependent children (AFDC), is added. We follow the same model specifications when estimating the GRF-IVQR.

Tables 3 and 4 show that for females, job training program generates a significantly positive treatment effect on earnings at 0.5 and 0.75 quantiles. GRF-IVQR signifies similar results.

Table 3. Effects of JTPA enrollment on earning (male).

	Quantile				
	0.15	0.25	0.50	0.75	0.85
AAI	121.000 (475.000)	702.000 (670.000)	1544.000 (1073.000)	3131.000 (1376.000)	3378.000 (1811.000)
CH	−151.151 (535.146)	528.529 (627.293)	312.312 (957.707)	2697.698 (1547.084)	3190.190 (1536.335)
GRF-IVQR	−199.114 (540.548)	232.099 (651.584)	1068.086 (950.880)	2630.969 (1571.200)	2955.952 (1645.931)

Note: GRF-IV: 1814.755 (1022.473). The GRF-IV stands for the 2SLS in the context of generalized random forests. AAI, CH and GRF-IVQR stand for, respectively, Abadie, Angrist and Imbens (2002), Chernozhukov and Hansen (2005) and our estimator. Numbers in parentheses are standard errors.

Table 4. Effects of JTPA enrollment on earning (female).

	Quantile				
	0.15	0.25	0.50	0.75	0.85
AAI	324.000 (175.000)	680.000 (282.000)	1742.000 (645.000)	1984.000 (945.000)	1900.000 (997.000)
CH	35.536 (266.445)	398.398 (313.555)	1566.567 (626.065)	2493.493 (910.474)	1845.345 (1059.988)
GRF-IVQR	185.141 (270.490)	571.842 (336.180)	1892.934 (610.466)	2431.793 (894.658)	1716.304 (1119.506)

Note: GRF-IV: 2127.544 (607.943). The GRF-IV stands for the 2SLS in the context of generalized random forests. AAI, CH and GRF-IVQR stand for, respectively, Abadie, Angrist and Imbens (2002), Chernozhukov and Hansen (2005) and our estimator. Numbers in parentheses are standard errors.

For the male group, Table 5 and Figure 2 depicts that work less than 13 weeks (wlkess13) and on-the-job training and/or job search assistance (ojt_jsa) are the most important variables. However, there is no apparent pattern suggesting that variable importance differs across quantiles. The pattern of variable importance resulting from the GRF-IV and the GRF-IVQR are different as well.

² Since Abadie, Angrist and Imbens (2002) and Chernozhukov and Hansen (2005) impose different identification strategies, the corresponding estimated quantile treatment effect are, in general, for distinct sub-populations.

Table 5. Variable importance (male).

	GRF-IV	GRF-IVQR at Specific Quantiles				
		0.15	0.25	0.50	0.75	0.85
High school or GED	0.11710	0.10009	0.10155	0.08882	0.07741	0.08376
Black	0.04914	0.06883	0.04729	0.09594	0.09909	0.10482
Hispanic	0.05177	0.00000	0.00000	0.00000	0.00000	0.00000
Married	0.12656	0.11679	0.12841	0.09669	0.07070	0.08854
Work less than 13 week in past year	0.09076	0.19681	0.16594	0.19491	0.07749	0.08512
Classroom training	0.05013	0.02939	0.02939	0.03967	0.08669	0.04849
On-the-job training and/or job search assistance	0.08262	0.24453	0.36400	0.27075	0.41578	0.37083
Age from 22 to 25	0.03710	0.04702	0.03930	0.04970	0.03646	0.04310
Age from 26 to 29	0.06769	0.02204	0.02204	0.02498	0.02792	0.06612
Age from 30 to 35	0.04465	0.03233	0.03820	0.04114	0.03673	0.03967
Age from 36 to 44	0.07998	0.02792	0.01910	0.03527	0.03086	0.02351
Age from 45 to 54	0.09737	0.00000	0.00000	0.00000	0.00000	0.00000
Whether data are from second follow-up survey	0.10514	0.11425	0.04476	0.06214	0.04087	0.04604

Note: The GRF-IV stands for the 2SLS in the context of generalized random forests.

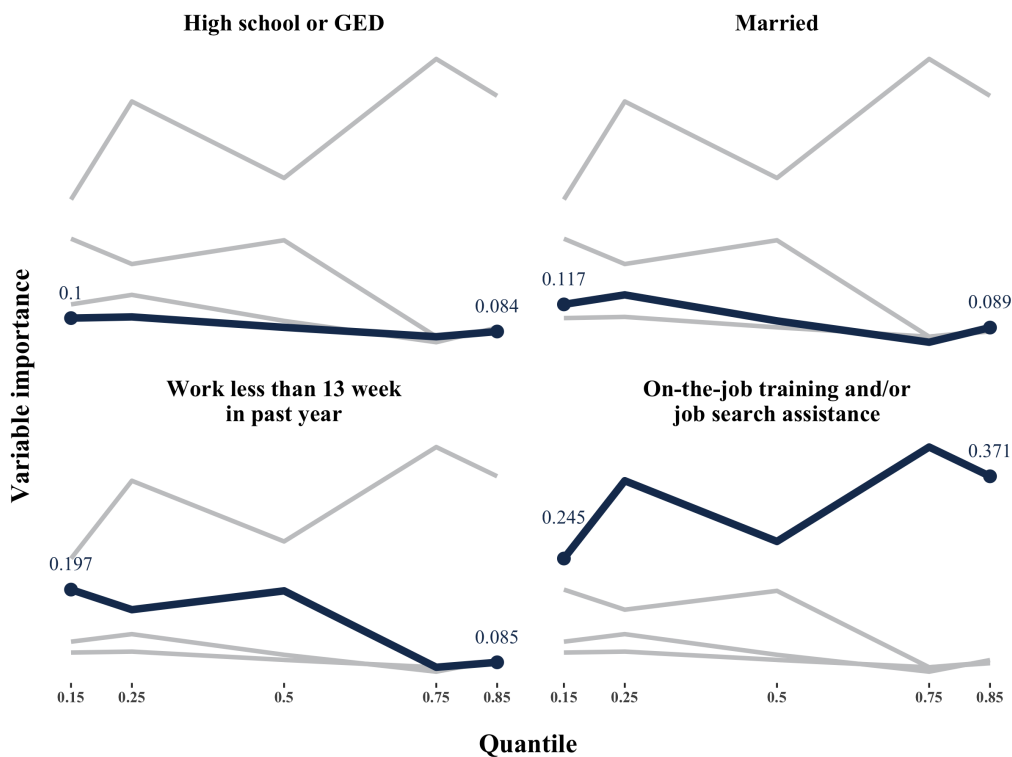


Figure 2. Top 4 variable importance (male).

As to the GRF-IVQR, in Table 5, the variance importance for Hispanics and the age group 45 to 54 are 0 across all quantiles, while the GRF-IV suggests these two variables are of some importance. Possible explanations are as follows. Compared to the GRF-IVQR's moment condition, the GRF-IV still performs well in nodes with a relatively small amount of data. Consequently, the GRF-IVQR is more restrictive for growing a shallower tree than the GRF-IV. Therefore, some variables used to make a split in deeper nodes will not be chosen by the GRF-IVQR algorithm. Besides, at deeper nodes, the data is very similar in each node. Specifically, this situation occurs frequently with a large number of binary variables, and thus leads

to no variation in a certain variable. Therefore, in practical estimation, the GRF-IVQR grows a relatively small tree.

For the female group, Table 6 and Figure 3 depicts that classroom training (*class_tr*) and on-the-job training and/or job search assistance (*ojt_jsa*) are the most important variables. The importance of on-the-job training and/or job search assistance decreases across quantiles, which is different from the pattern in the male group. The issue concerning no variation of a binary variable in deeper nodes becomes severe in the female group.

The variance importance for Hispanics and several age binary variables are 0 across all quantiles, which indicates that in the female group, the aforementioned characteristics variables are more homogeneous over the conditional distribution of earnings.

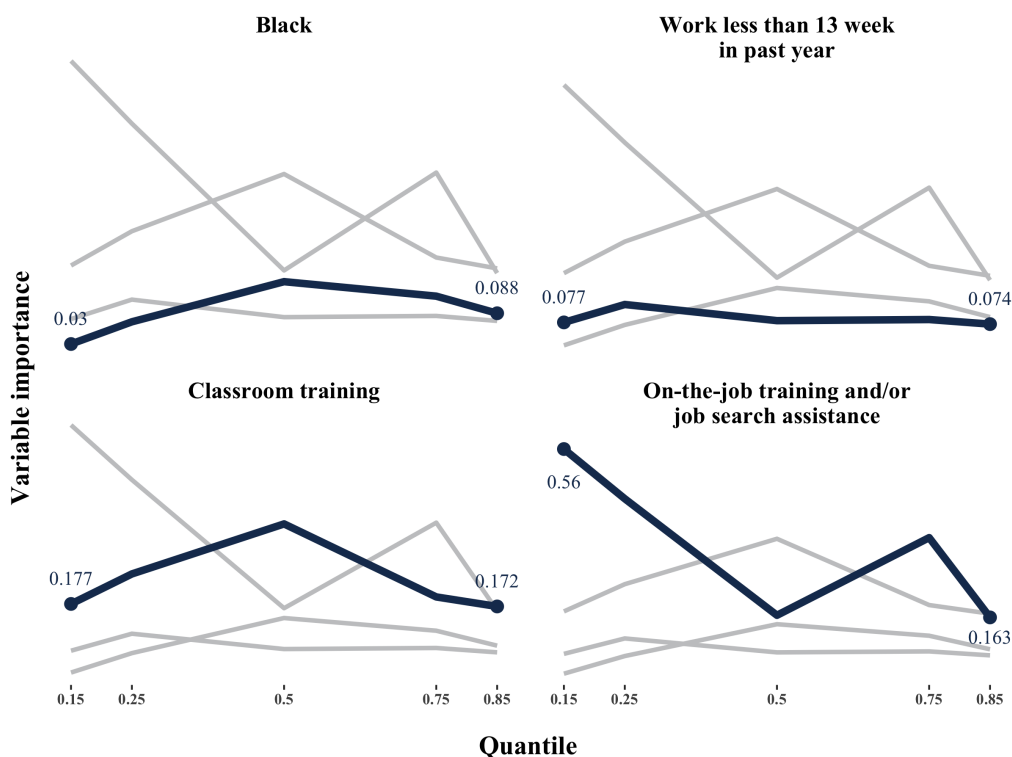


Figure 3. Top 4 variable importance (female).

Table 6. Variable importance (female).

	GRF-IV	GRF-IVQR at Specific Quantiles				
		0.15	0.25	0.50	0.75	0.85
High school or GED	0.06385	0.02720	0.03360	0.05760	0.04320	0.11040
Black	0.05128	0.03040	0.07200	0.14720	0.12000	0.08800
Hispanic	0.07033	0.00000	0.00000	0.00000	0.00000	0.00000
Married	0.09909	0.03520	0.01120	0.08000	0.06880	0.18240
ADFC	0.14744	0.02658	0.02880	0.08070	0.07881	0.10153
Work less than 13 week in past year	0.05033	0.07717	0.11360	0.08070	0.08301	0.07367
Classroom training	0.15284	0.17735	0.24160	0.34876	0.19237	0.17197
On-the-job training and/or job search assistance	0.06905	0.56049	0.44320	0.16823	0.35140	0.16323
Age from 22 to 25	0.03948	0.00160	0.00480	0.00000	0.00160	0.00000
Age from 26 to 29	0.05355	0.00000	0.00000	0.00000	0.00000	0.00000
Age from 30 to 35	0.05257	0.03040	0.03520	0.01280	0.01600	0.04000
Age from 36 to 44	0.05325	0.00000	0.00000	0.00000	0.00000	0.00000
Age from 45 to 54	0.04129	0.00000	0.00000	0.00000	0.00000	0.00000
Whether data are from second follow-up survey	0.05564	0.03360	0.01600	0.02400	0.04480	0.06880

Note: The GRF-IV stands for the 2SLS in the context of generalized random forests.

5. Conclusions

Based on the generalized random forests of [Athey et al. \(2019\)](#), we propose an econometric procedure to estimate the quantile treatment effect. Not only does this method estimate the treatment effect nonparametrically, but our procedure yields a measure of variable importance, in terms of heterogeneity among control variables. We provide the practical algorithm and the associated R codes. We also apply the proposed procedure to reinvestigate the distributional effect of 401(k) participation on net financial assets, and the quantile effect of participating a job training program on earnings. Income, age, education, and family size are identified as the first-four important variables in the 401(k) analysis. In the job training program example, our procedure suggests that the previous year work status and the job training service strategies are important control variables.

Author Contributions: Both authors contributed equally to the paper.

Funding: This research was partly funded by the personal research fund from Tokyo International University, and financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 107L900203) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Acknowledgments: We are grateful to the two anonymous referees for their constructive comments that have greatly improved this paper. We thank Patrick DeJarnette, Masaru Inaba, Min-Jeng Lin and Shinya Tanaka for discussions and comments. This paper has benefited from presentation at the Aoyama Gakuin University, Kansai University, and 2019 Annual Conference of Taiwan Economic Association. The usual disclaimer applies.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DML	Double machine learning
GRF	Generalized random forests
IVQR	Instrumental variable quantile regression

Appendix A

Appendix A.1. Improving Efficiency by Doubly Robust Estimators

Chernozhukov et al. (2018) and Athey and Wager (2018) pioneered the use of the doubly robust estimator embedded in a framework of causal machine learning. The resulting estimator becomes more accurate and gains efficiency. In light of their idea, it might be beneficial to incorporate the doubly robust estimation in our methodology.

A doubly robust augmented inverse propensity weighted (AIPW) estimator was introduced by Robins et al. (1994). The AIPW estimator for average treatment effect is constructed by two components as follows.

$$\widehat{ATE}_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left\{ \left[\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \hat{e}(X_i)} \right] - \frac{(D_i - \hat{e}(X_i))}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \left[(1 - \hat{e}(X_i)) \hat{E}(Y_i | D_i = 1, X_i) + \hat{e}(X_i) \hat{E}(Y_i | D_i = 0, X_i) \right] \right\},$$

where $e(x) = P[D_i | X_i = x]$ being the propensity score. The first line in the equation represents the inverse probability weighted estimator, and the second line depicts a weighted regression. The AIPW estimator is doubly robust because the estimator will be consistent, provided that at least one of the two components is correctly specified.

Appendix A.2. The Doubly Robust Estimation for Causal Forests

Athey and Wager (2019) and their **grf** R package implement a variant of doubly robust AIPW estimators for causal forests. Specifically, for estimating average treatment effect, their doubly robust estimator is shown as follows.

$$\begin{aligned} \widehat{ATE} &= \hat{\gamma} = \frac{1}{n} \sum_{i=1}^n \hat{\Gamma}_i, \\ \hat{\Gamma}_i &= \hat{\gamma}^{(-i)}(X_i) \\ &\quad + \frac{D_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)(1 - \hat{e}^{(-i)}(X_i))} \left(Y_i - \hat{m}^{(-i)}(X_i) - (D_i - \hat{e}^{(-i)}(X_i)) \hat{\gamma}^{(-i)}(X_i) \right), \end{aligned}$$

where $\hat{\gamma}(X_i)$ is the conditional average treatment effect estimator based on causal forest, $\hat{\Gamma}_i$ is the conditional average treatment effect estimator adjusted by inverse probability weighting, $\hat{m}(x)$ and $\hat{e}(x)$ are the estimators of $E[Y|X = x]$ and $E[D|X = x]$ which are based on random forest with honest splitting, and the average treatment effect estimator $\hat{\gamma}$ is simply the sample average of those adjusted conditional average treatment effect estimates.

Glynn and Quinn (2009) provide some evidence that the doubly robust estimator performs better in terms of efficiency than inverse probability weighting estimators, matching estimators, and regression estimators. To explore how adapting the doubly robust method in the causal forest estimator affects the efficiency and accuracy, we follow their DGP designs and conduct Monte Carlo experiments with different degree of confoundedness. In the simulation, X_1, X_2 , and X_3 are covariates following $N(0, 1)$, D is the treatment variable, Y is the outcome variable, and ϵ is the disturbance which follows $N(0, 1)$. Two data generating processes are considered. Degree of confoundedness are modeled in three levels: low, moderate, and severe.

Table A1. Simulation setting.

	Outcome (control)	Outcome (treatment)
Simple DGP	$Y = X_2 + X_3 + \epsilon$	$Y = 5 + 3X_2 + X_3 + \epsilon$
Complicate DGP	$Y = X_2 + X_3 + \epsilon$	$Y = 5 + 3X_2 + X_3 + 2X_2^2 + 2X_3^2 + \epsilon$
Degree of confoundedness	True treatment assignment probabilities	
Low	$P(D = 1 X) = \Phi(0.1X_1 + 0.1X_2 + 0.05X_1X_2)$	
Moderate	$P(D = 1 X) = \Phi(X_1 + X_2 + 0.5X_1X_2)$	
Severe	$P(D = 1 X) = \Phi(1.5X_1 + 1.5X_2 + 0.75X_1X_2)$	

With three different sample sizes, 250, 500, and 1000, three degrees of confoundedness, and two DGP settings, the Monte Carlo results are tabulated in Table A2. The results confirm that the causal forest with doubly robust estimation indeed has efficiency gains over the conventional causal forest.

Table A2. Finite-sample performance: causal forests with doubly robust estimation.

		Linear DGP		Nonlinear DGP	
		Causal forest	Causal forest with doubly robust	Causal forest	Causal forest with doubly robust
Sample size	Confoundedness degree	RMSE	RMSE	RMSE	RMSE
250	low	0.3730	0.1693	0.7542	0.3147
250	moderate	0.4200	0.2099	0.9295	0.3914
250	severe	0.4562	0.2218	1.0205	0.3997
500	low	0.3206	0.1081	0.6911	0.1855
500	moderate	0.3634	0.1417	0.8711	0.2320
500	severe	0.4107	0.1497	0.9529	0.2505
1000	low	0.2745	0.0717	0.6041	0.1124
1000	moderate	0.3244	0.1008	0.7755	0.1540
1000	severe	0.3742	0.1098	0.8919	0.1709

Appendix A.3. The Doubly Robust Estimation for Instrumental Causal Forests

With instrumental variables, [Athey and Wager \(2018\)](#) provide a doubly robust estimator for local average treatment effect; namely

$$\widehat{\text{LATE}} = \hat{\gamma} = \frac{1}{n} \sum_{i=1}^n \hat{\Gamma}_i,$$

$$\hat{\Gamma}_i = \hat{\gamma}^{(-i)}(X_i) + \frac{1}{\hat{\Delta}^{(-i)}(X_i)} \frac{Z_i - \hat{z}^{(-i)}(X_i)}{\hat{z}^{(-i)}(X_i)(1 - \hat{z}^{(-i)}(X_i))} \left(Y_i - \hat{m}^{(-i)}(X_i) - (D_i - \hat{e}^{(-i)}(X_i)) \hat{\gamma}^{(-i)}(X_i) \right),$$

where $\hat{\gamma}(X_i)$ is the conditional local average treatment effect estimator based on instrumental forest, $\hat{\Gamma}_i$ is the conditional local average treatment effect estimator adjusted by inverse probability weighting, $\hat{m}(x)$, $\hat{e}(x)$, $\hat{z}(x)$, and $\hat{\Delta}(x)$ are the estimators of $E[Y|X = x]$, $E[D|X = x]$, $E[Z|X = x]$, and $P(D|Z = 1, X = x) - P(D|Z = 0, X = x)$ which are based on random forest with honest splitting, and the local average treatment effect estimator $\hat{\gamma}$ is simply the sample average of those adjusted conditional local average treatment effect estimates.

Appendix A.4. An Unsolved Task: The Doubly-Robust GRF-IVQR

Researchers would like to incorporate the doubly robust estimation in the GRF-IVQR model, following similar ideas introduced above. However, it remains unclear how to do it. We leave this unsolved task as future work.

Appendix A.5. Identifying Restrictions and Regularity Conditions for the GRF-IVQR

Following Chernozhukov and Hansen (2008), we consider the instrumental variable quantile regression characterizing the structural relationship:

$$\begin{aligned} Y &= D'\theta(U) + X'\nu(U), \quad U|X, Z \sim \text{Uniform}(0, 1) \\ D &= \delta(X, Z, V) \quad \text{where } V \text{ is statistically dependent on } U \\ \tau &\mapsto D'\theta(\tau) + X'\nu(\tau) \quad \text{strictly increasing in } \tau \end{aligned}$$

where

- Y is the scalar outcome variable of interest.
- U is a scalar random variable (rank variable) that aggregates all of the unobserved factors affecting the structural outcome equation.
- D is a vector of endogenous variables determined by $\delta(X, Z, V)$.
- V is a vector of unobserved disturbances determining D and correlated with U .
- Z is a vector of instrumental variables.
- X is a vector of included control variables.

The one-dimensional rank variable and the rank similarity (rank preservation) condition imposed on the outcome equation play an important role in identifying the quantile treatment effect. To derive the standard error of the IVQR estimator, the following assumptions are needed as well.

Assumption CH1. Y_i, D_i, X_i, Z_i are iid defined on the probability space Ω, F, P and have compact support.

Assumption CH2. For the given τ , $(\theta(\tau), \nu(\tau))$ is in the interior of the parameter space.

Assumption CH3. Density $f_Y(Y|X, D, Z)$ is bounded by a constant \bar{f} a.s.

Assumption CH4. $\partial E[1(Y < D'\theta + X'\nu + Z\gamma)\Psi]/\partial(\nu', \gamma')$ at $(\nu, \gamma) = (\nu(\theta, \tau), \gamma(\theta, \tau))$ has full rank for each θ in Θ , for $\Psi = V_i(Z_i', X_i')'$.

Assumption CH5. $\partial E[1(Y < D'\theta + X'\nu)\Psi]/\partial(\theta', \nu')$ has full rank at $(\theta(\tau)', \nu(\tau)')$.

Assumption CH6. The function $(\theta, \nu) \mapsto E[\{\tau - 1(Y < D'\theta + X'\nu)\Psi\}]$ is one-to-one over parameter space.

Assumptions CH1–CH6 are compatible with those imposed in Athey et al. (2019); for example, both sets of assumptions do not apply to time-series data.

Assumption ATW1 (Lipschitz x -signal). For fixed values of (θ, ν) , we assume that $M_{\theta, \nu}(x) := \mathbb{E}[\psi_{\theta, \nu}(O)|X = x]$ is Lipschitz continuous in x .

Assumption ATW2 (Smooth identification). When x is fixed, we assume that the M -function is twice continuously differentiable in (θ, ν) with a uniformly bounded second derivative, and that $V(x) := V_{\theta(x), \nu(x)}(x)$ is invertible for all $x \in \mathcal{X}$, with $V_{\theta, \nu} := \frac{\partial}{\partial(\theta, \nu)} M_{\theta, \nu}(x)|_{\theta(x), \nu(x)}$.

Assumption ATW3 (Lipschitz (θ, ν) -variogram). The score functions $\psi_{\theta, \nu}(O_i)$ have a continuous covariance structure. Writing γ for the worst-case variogram and $\|\cdot\|_F$ for the Frobenius norm, then for some $L > 0$,

$$\gamma \left(\begin{pmatrix} \theta \\ \nu \end{pmatrix}, \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right) \leq L \left\| \begin{pmatrix} \theta \\ \nu \end{pmatrix} - \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right\|_2 \quad \text{for all } (\theta, \nu), (\theta', \nu')$$

$$\gamma \left(\begin{pmatrix} \theta \\ \nu \end{pmatrix}, \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right) := \sup_{x \in \mathcal{X}} \{ \| \text{Var}[\psi_{\theta, \nu}(O_i) - \psi_{\theta', \nu'}(O_i) | X_i = x] \|_F \}$$

Assumption ATW4 (Regularity of ψ). The ψ -functions can be written as $\psi_{\theta, \nu}(O) = \lambda(\theta, \nu; O_i) + \zeta_{\theta, \nu}(g(O_i))$, such that λ is Lipschitz-continuous in (θ, ν) , $g : O_i \rightarrow \mathbb{R}$ is a univariate summary of O_i , and $\zeta_{\theta, \nu} : \mathbb{R} \rightarrow \mathbb{R}$ is any family of monotone and bounded functions.

Assumption ATW5 (Existence of solutions). We assume that, for any weights α_i with $\sum \alpha_i = 1$, the estimating equation $(\hat{\theta}(x), \hat{\nu}(x)) \in \text{argmin}_{\theta, \nu} \left\{ \left\| \sum_{i=1}^n \alpha_i(x) \psi_{\theta, \nu}(O_i) \right\|_2 \right\}$ returns a minimizer $(\hat{\theta}, \hat{\nu})$ that at least approximately solves the estimating equation $\left\| \sum_{i=1}^n \alpha_i \psi_{\hat{\theta}, \hat{\nu}}(O_i) \right\|_2 \leq C \max \{\alpha_i\}$, for some constant $C \geq 0$.

Assumption ATW6 (Convexity). The score function $\psi_{\theta, \nu}(O_i)$ is a negative sub-gradient of a convex function, and the expected score $M_{\theta, \nu}(X_i)$ is the negative gradient of a strongly convex function.

Given Assumptions ATW1-ATW6, the Theorems 3 and 5 of [Athey et al. \(2019\)](#) guarantee that the GRF estimator achieves consistency and asymptotic normality. In what follows, we check each assumptions for the proposed GRF-IVQR estimator.

Observe that the score function of the IVQR

$$\psi_{\theta, \nu}(O_i) = [\tau - 1(Y_i \leq D_i \theta(\tau, x) + X_i \nu(\tau, x))] (Z_i, X_i)'$$

In [Chernozhukov and Hansen \(2008\)](#), the moment functions are conditional on $\{X_i, D_i, Z_i\}$. For simplicity, we write conditional functions as $[\cdot | X_i = x]$ when considering splitting in X_i within the framework of generalized random forests.

Checking Assumption ATW1.

$$\begin{aligned} & \mathbb{E} [\psi_{\theta(\tau, x), \nu(\tau, x)}(O_i) | X_i = x] \\ &= \mathbb{E} [\tau - 1(Y_i \leq D_i \theta(\tau, x) + X_i \nu(\tau, x))] (Z_i, X_i)' | X_i = x \quad \text{for all } x \in \mathcal{X}. \end{aligned}$$

Thus the expected score function

$$\begin{aligned} M_{\theta, \nu}(x) &= \mathbb{E} [\psi_{\theta, \nu}(O_i) | X_i = x] \\ &= \mathbb{E} [\tau - 1(Y_i \leq D_i \theta + X_i \nu)] (Z_i, X_i)' | X_i = x \\ &= [\tau - F(Y_i \leq D_i \theta + X_i \nu | X_i = x)] (Z_i, x)'. \end{aligned}$$

We want the conditional cumulative distribution function is Lipschitz continuous. Since every function with bounded first derivatives is Lipschitz, we need the conditional density is bounded.

Assumption CH3 states that the conditional density $f_Y(Y|X, D, Z)$ is bounded by a constant \bar{f} a.s.. In particular, $f_Y(Y|X, D, Z)$ is a density of a convolution of a continuous random variable and a discrete random variable, we also need the continuous variable not to be degenerate.

Checking Assumption ATW2.

$$\begin{aligned} M_{\theta, \nu}(x) &= \left[\tau - F(D_i\theta + X_i\nu | X_i = x) \right] (Z_i, x)'. \\ V_{\theta, \nu}(x) &= \frac{\partial}{\partial(\theta, \nu)} M_{\theta, \nu}(x) \Big|_{\theta(\tau, x), \nu(\tau, x)} \\ &= \frac{\partial}{\partial(\theta, \nu)} \left\{ \left[\tau - F(D_i\theta + X_i\nu | X_i = x) \right] (Z_i, x)' \right\} \\ &= \begin{bmatrix} -f(D_i\theta + X_i\nu | X_i = x) Z_i' D_i & -f(D_i\theta + X_i\nu | X_i = x) x' D_i \\ -f(D_i\theta + X_i\nu | X_i = x) Z_i' x & -f(D_i\theta + X_i\nu | X_i = x) x' x. \end{bmatrix} \end{aligned}$$

We want V is invertible and therefore $\begin{bmatrix} Z_i' D_i & x' D_i \\ Z_i' x & x' x \end{bmatrix}$ needs to be invertible. In addition, the conditional density $f(D_i\theta + X_i\nu | X_i = x)$ is required to have continuous uniformly bounded first derivative. If $f(D_i\theta + X_i\nu | X_i = x)$ is continuously differentiable, then its first derivative is uniformly bounded. Those conditions are implied by Assumptions CH4 and CH5. Thus A_p is invertible as well.

Checking Assumption ATW3.

$$\begin{aligned} &\gamma \left(\begin{pmatrix} \theta \\ \nu \end{pmatrix}, \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right) \\ &= \sup_{x \in \mathcal{X}} \left\{ \left\| \text{Var}[\psi_{\theta, \nu}(O_i) - \psi_{\theta', \nu'}(O_i) | X_i = x] \right\|_F \right\} \\ &= \sup_{x \in \mathcal{X}} \left\{ \left\| \text{Var} \left[[\tau - 1(Y_i \leq D_i\theta + X_i\nu)] (Z_i, X_i)' - [\tau - 1(Y_i \leq D_i\theta' + X_i\nu')] (Z_i, X_i)' \mid X_i = x \right] \right\|_F \right\} \\ &= \sup_{x \in \mathcal{X}} \left\{ \left\| (Z_i, x)' (Z_i, x) \text{Var} \left[-1(Y_i \leq D_i\theta + X_i\nu) + 1(Y_i \leq D_i\theta' + X_i\nu') \mid X_i = x \right] \right\|_F \right\} \\ &= \sup_{x \in \mathcal{X}} \left\{ \left\| (Z_i, x)' (Z_i, x) \left[F(D_i\theta + X_i\nu | X_i = x) - F(D_i\theta' + X_i\nu' | X_i = x) \right] \right. \right. \\ &\quad \left. \left. \left[1 - [F(D_i\theta + X_i\nu | X_i = x) - F(D_i\theta' + X_i\nu' | X_i = x)] \right] \right\|_F \right\}. \end{aligned}$$

Taylor expansion implies the following approximation of γ .

$$\gamma \left(\begin{pmatrix} \theta \\ \nu \end{pmatrix}, \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right) \approx \sup_{x \in \mathcal{X}} \left\{ \left\| (Z_i, x)' (Z_i, x) \left[f(y | X_i = x) (D_i(\theta - \theta') + X_i(\nu - \nu')) \right] \right\|_F \right\}.$$

Since the conditional probability density function is bounded, there exists a $L > 0$, such that

$$\gamma\left(\begin{pmatrix} \theta \\ \nu \end{pmatrix}, \begin{pmatrix} \theta' \\ \nu' \end{pmatrix}\right) \leq L \left\| \begin{pmatrix} \theta \\ \nu \end{pmatrix} - \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right\|_2.$$

Checking Assumption ATW4. The score function can be written as

$$\begin{aligned} \psi_{\theta,\nu}(O_i) &= [\tau - 1(Y_i \leq D_i\theta + X_i\nu)](Z_i, X_i)' \\ &= \lambda(\theta, \nu; O_i) + \zeta_{\theta,\nu}(g(O_i)), \end{aligned}$$

where

$$\begin{aligned} g(O_i) &= Y_i, \text{ and} \\ \zeta_{\theta,\nu}(g(O_i)) &= \begin{pmatrix} [\tau - 1(Y_i \leq D_i\theta + X_i\nu)]Z_i \\ [\tau - 1(Y_i \leq D_i\theta + X_i\nu)]X_i \end{pmatrix}. \end{aligned}$$

Checking Assumption ATW5. Since Assumption ATW5 is used to ensure the existence of solutions, it is required.

Checking Assumption ATW6. With a V-shaped check function of the instrumental variable quantile regression, the corresponding score function $\psi_{\theta,\nu}(O_i)$ is a negative subgradient of a convex function, and the expected score function $M_{\theta,\nu}(x)$ is a negative gradient of a strongly convex function. Therefore, Assumption ATW6 holds.

Corollary. (Consistency and asymptotic normality of the GRF-IVQR estimator) *Given Assumptions ATW1-6, Assumptions CH1-6, and Theorems 3 and 5 of Athey et al. (2019), the GRF-IVQR is consistent and asymptotically normal:*

$$\frac{\hat{\theta}_n(x) - \theta(x)}{\sigma_n(x)} \rightarrow N(0, 1).$$

The variance estimator

$$\hat{\sigma}_n^2 = \xi^\top \hat{V}_n(x)^{-1} \hat{H}_n(x) (\hat{V}_n(x)^{-1})^\top \xi,$$

where $\hat{V}_n(x)$ and $\hat{H}_n(x)$ are consistent estimators for the $V_{\theta,\nu}(x)$ and $H_n(x) = \text{Var} \sum_{i=1}^n \alpha_i \psi_{\theta,\nu}(O_i)$ respectively.

References

- Abadie, Alberto, Joshua Angrist, and Guido Imbens. 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70: 91–117. [\[CrossRef\]](#)
- Athey, Susan, and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113: 7353–60. [\[CrossRef\]](#) [\[PubMed\]](#)
- Athey, Susan, and Guido Imbens. 2019. Machine learning method that economists should know about. *Annual Review of Economics* 11: 685–725. [\[CrossRef\]](#)
- Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. Generalized random forests. *The Annals of Statistics* 47: 1148–78. [\[CrossRef\]](#)
- Athey, Susan, and Stefan Wager. 2018. Efficient policy learning. *arXiv*, arXiv:1702.02896v4.

- Athey, Susan, and Stefan Wager. 2019. Estimating treatment effects with causal forests: An application. *arXiv*, arXiv:1902.07409.
- Breiman, Leo. 2001. Random forests. *Machine Learning* 45: 5–32. [CrossRef]
- Chen, Jau-er, and Jia-Jyun Tien. 2019. *Debiased/Double Machine Learning for Instrumental Variable Quantile Regressions*. Working Paper. Taipei, Taiwan: Center for Research in Econometric Theory and Applications, National Taiwan University.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21: C1–C68. [CrossRef]
- Chernozhukov, Victor, and Christian Hansen. 2004. The effects of 401(k) participation on the wealth distribution: An Instrumental quantile regression analysis. *Review of Economics and Statistics* 86: 735–51. [CrossRef]
- Chernozhukov, Victor, and Christian Hansen. 2005. An IV model of quantile treatment effects. *Econometrica* 73: 245–61. [CrossRef]
- Chernozhukov, Victor, and Christian Hansen. 2008. Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics* 142: 379–98. [CrossRef]
- Chernozhukov, Victor, and Christian Hansen. 2013. NBER 2013 Summer Institute: Econometric Methods for High-Dimensional Data. Available online: http://www.nber.org/econometrics_minicourse_2013/ (accessed on 15 July 2013).
- Chiou, Yan-Yu, Mei-Yuan Chen, and Jau-er Chen. 2018. Nonparametric regression with multiple thresholds: Estimation and inference. *Journal of Econometrics* 206: 472–514. [CrossRef]
- Davis, Jonathan M. V., and Sara B. Heller. 2017. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review* 107: 546–50. [CrossRef]
- Frandsen, Brigham R., and Lars J. Lefgren. 2018. Testing rank similarity. *Review of Economics and Statistics* 100: 86–91. [CrossRef]
- Gilchrist, Duncan Sheppard, and Emily Glassberg Sands. 2016. Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy* 124: 1339–82. [CrossRef]
- Glynn, Adam N., and Kevin M. Quinn. 2009. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* 18: 36–56. [CrossRef]
- Guilhem, Bascle, Louis Mulotte, and Jau-er Chen. 2019. Addressing Strategy endogeneity and performance heterogeneity: Evidence from firm multinationality. *Academy of Management Proceedings* 2019: 12733.
- O'Neill, Eoghan, and Melvyn Weeks. 2018. Causal tree estimation of heterogeneous household response to time-of-use electricity pricing schemes. *arXiv*, arXiv:1810.09179.
- Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89: 846–66. [CrossRef]
- Strittmatter, Anthony. 2019. Heterogeneous earnings effects of the job corps by gender: A translated quantile approach. *Labour Economics* 61: 101760. [CrossRef]
- Wager, Stephan, and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113: 1228–42. [CrossRef]

