Debiased/Double Machine Learning for Instrumental Variable Quantile Regressions

Jau-er Chen\textsuperscript{a,b} and Jia-Jyun Tien\textsuperscript{c}

\textsuperscript{a}Institute for International Strategy, Tokyo International University.
\textsuperscript{b}Center for Research in Econometric Theory and Applications, National Taiwan University.
\textsuperscript{c}Department of Economics, National Taiwan University.

Abstract

The aim of this paper is to investigate estimation and inference on a low-dimensional causal parameter in the presence of high-dimensional controls in an instrumental variable quantile regression. The estimation and inference are based on the Neyman-type orthogonal moment conditions, that are relatively insensitive to the estimation of the nuisance parameters. The Monte Carlo experiments show that the econometric procedure performs well. We also apply the procedure to reinvestigate two empirical studies: the quantile treatment effect of 401(k) participation on accumulated wealth, and the distributional effect of job-training program participation on trainee earnings.

Keywords: instrumental variable, quantile regression, treatment effect, LASSO, double machine learning.

\textit{JEL Classification:} C21; C26.

Correspondence: Jau-er Chen. E-mail: jechen@tiu.ac.jp Address: 1-13-1 Matobakita Kawagoe, Saitama 350-1197, Japan.

This version: September 2019. We are grateful to Masayuki Hirukawa, Tsung-Chih Lai, and Hsin-Yi Lin for discussions and comments. This paper has benefited from presentations at the 2nd International Conference on Econometrics and Statistics (EcoSta 2018), and the Ryukoku University. The authors declare no conflict of interest. The usual disclaimer applies.

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.
1 Introduction

Model selection and variable selection are widely discussed in the area of prediction. Much less attention, however, has been paid to the modification of prediction methods under the context of causal machine learning in economics, cf. Athey (2017) and Athey (2018). As one of the pioneering papers, within the linear framework of instrumental variable estimation, Belloni et al. (2014) proposed a double-selection procedure to correct for an omitted variable bias in a high-dimensional framework. Constructing a general framework encompassing results from the aforementioned Belloni’s paper, Chernozhukov et al. (2015) and Chernozhukov et al. (2018a) proposed a unified procedure, double/debiased machine learning (DML), which remains valid for nonlinear or semi-nonparametric models. The aim of this paper is to investigate estimation and inference on a low-dimensional causal parameter in the presence of high-dimensional controls in an instrumental variable quantile regression. In particular, our procedure follows the idea outlined by Chernozhukov et al (2018b). To the best of our knowledge, the present study is the first to investigate Monte Carlo performance and empirical studies of the double machine learning procedure within the framework of instrumental variable quantile regressions. The Monte Carlo experiments show that our econometric procedure performs well.

Causal machine learning has been actively studied in economics in recent years, which are based on two approaches: the double machine learning, cf. Chernozhukov et al. (2018), and the generalized random forests, cf. Athey, Tibshirani and Wager (2019). Chen and Hsiang (2019) investigate the generalized random forests model using instrumental variable quantile regression. In contrast to the DML for instrumental variable quantile regressions, their econometric procedure yields a measure of variable importance in terms of heterogeneity among control variables. Although related to our paper, Chen and Hsiang (2019) do not consider the setting of high-dimensional controls.

We apply the proposed procedure to empirically investigate causal quantile effects of the 401(k) participation on net financial assets. Our empirical results signify that the 401(k) participants with low savings propensity are more associated with the nonlinear income effect, which complements the findings concluded in Chernozhukov et al. (2018a) and Chiou et al. (2018). Another empirical example of the job training program participation is investigated as well.

The rest of the paper is organized as follows. The model specification and estimation procedure are introduced in Section 2. Section 3 presents Monte Carlo experiments. Section 4 presents two empirical applications. Section 5 concludes the paper.


2 The Model

We briefly review the conventional instrumental variable quantile regression (IVQR), and then the IVQR within the framework of high-dimensional controls. Our DML procedure for the IVQR is introduced in this section, which is constructed based on a tentative procedure suggested by Chernozhukov et al. (2018b).

2.1 The Inverse Quantile Regression as a GMM Estimator

The following conditional moment restriction yields an IVQR estimator.

\[ \mathbb{P}[Y \leq q(\tau, D, X)|X, Z] = \tau, \quad (1) \]

where \( q(\cdot) \) is the structural quantile function, \( \tau \) stands for the quantile index, \( D, X \) and \( Z \) are, respectively, the target variable, control variables and instruments. Condition (1) and linear structural quantile specification leads to the following unconditional moment restriction

\[ \mathbb{E}[(\tau - 1(Y - D'\alpha - X'\beta) \leq 0)\Psi] = 0 \quad (2) \]

where

\[ \Psi := \Psi(X, Z) \]

is a vector of a function of instruments and control variables. The parameters depend on the quantile of interest, but we suppress the \( \tau \) associated with \( \alpha \) and \( \beta \) for simplicity of presentation. Equation (2) leads to a particular moment condition for doing partialling out:

\[ g_\tau(V, \alpha; \beta, \delta) = (\tau - 1(Y - D'\alpha - X'\beta))\Psi(\alpha, \delta(\alpha))) \quad (3) \]

with “instrument”

\[ \Psi(\alpha, \delta(\alpha)) := (Z - \delta(\alpha)X) \]

\[ \delta(\alpha) = M(\alpha)J^{-1}(\alpha), \]

where \( \delta \) is a matrix parameter,

\[ M(\alpha) = \mathbb{E}[ZX'f_\epsilon(0|X, Z)], \quad J(\alpha) = \mathbb{E}[XX'f_\epsilon(0|X, Z)] \]

and \( f_\epsilon(0|X, Z) \) is the conditional density of \( \epsilon = Y - D'\alpha - X'/\beta(\alpha) \) with \( \beta(\alpha) \) defined by

\[ \mathbb{E}[(\tau - 1(Y - D'\alpha + X'/\beta(\alpha))X) = 0. \quad (5) \]
We construct the grid search interval for $\alpha$ first and profile out the coefficient for each $\alpha$ in the interval on the exogenous variable by equation (5). That is, 

$$
\hat{\beta}(a) = \arg \min_{b \in B} \frac{1}{N} \sum_{i=1}^{N} \rho_\tau(Y_i - D_i' a - X_i' b).
$$

We build sample counterpart of the population moment condition based on equations (2)–(5). That is, 

$$
\hat{g}_N(a) = \frac{1}{N} \sum_{i=1}^{N} g(V_i, a, \hat{\beta}(a), \hat{\delta}(a)), \quad (6)
$$

where 

$$
\hat{\delta}(a) = \hat{M}(a) \hat{J}^{-1}(a)
$$

for 

$$
\hat{M}(a) = \frac{1}{Nh_N} \sum_{i=1}^{N} Z_i X_i' K_{h_N}(Y_i - D_i' a - X_i' \hat{\beta}(a))
$$

$$
\hat{J}(a) = \frac{1}{Nh_N} \sum_{i=1}^{N} X_i X_i' K_{h_N}(Y_i - D_i' a - X_i' \hat{\beta}(a))
$$

where $K_{h_N}$ is a kernel function with bandwidth $h_N$. We thus can solve for the parameters through optimizing the GMM criterion function. Specifically, 

$$
\hat{\alpha}(\tau) = \arg \min_{a \in A} N \hat{g}_N(a)' \hat{\Sigma}(a, a)^{-1} \hat{g}_N(a) \quad (7)
$$

$$
\hat{\Sigma}(a_1, a_2) = \frac{1}{N} \sum_{i=1}^{N} g(V_i, a_1, \hat{\beta}(a_1)) g(V_i, a_2, \hat{\beta}(a_2))'
$$

where $\hat{\Sigma}(a_1, a_2)$ is a weighting matrix used in the GMM estimation. Notice that the estimator $\hat{\alpha}$ based on the inverse quantile regression (i.e. IVQR) is first-order equivalent to the estimator defined by the GMM.

### 2.2 Estimation with High-dimensional Controls

We modify the procedure introduced in Subsection 2.1 in order to deal with a dataset of high-dimensional control variables. We construct the grid search interval for $\alpha$ and profile out the coefficients on exogenous variable using the $L_1$-norm penalized quantile regression estimator: 

$$
\hat{\beta}(a) = \arg \min_{b \in B} \frac{1}{n} \sum_{i=1}^{N} \rho_\tau(Y_i - D_i' a - X_i' b) + \lambda \sum_{j=1}^{\text{dim}(b)} |b_j|. \quad (8)
$$
In addition, we estimate

\[ \hat{M}(a) = \frac{1}{Nh_N} \sum_{i=1}^{N} Z_i X_i' K_{h_N} (Y_i - D_i'a - X_i' \hat{\beta}(a)) \]

\[ \hat{J}(a) = \frac{1}{Nh_N} \sum_{i=1}^{N} X_i X_i' K_{h_N} (Y_i - D_i'a - X_i' \hat{\beta}(a)). \]

We also do dimension reduction on \( J \) because of the large dimension of \( X \). In particular, we implement the following regularization.

\[ \hat{\delta}_j(a) = \arg \min_{\delta} \frac{1}{2} \delta' \hat{J}(a) \delta - \hat{M}_j(a) \delta + \vartheta ||\delta||_1. \]

The regularization above does a weighting LASSO for each instrument variable on control variables, and consequently the \( L_1 \) norm optimization obeys the Karush-Kuhn-Tucker condition

\[ ||\hat{\delta}_j(a)' \hat{J}(a) - \hat{M}_j(a)||_\infty \leq \vartheta, \quad \forall j. \]  

After implementing the double machine learning procedure outlined above for the IVQR, we now can solve for the low-dimensional causal parameter \( \alpha \) through optimizing the GMM defined as follows. The sample counterpart of the moment condition

\[ \hat{g}_N(a) = \frac{1}{N} \sum_{i=1}^{N} (\tau - 1(Y_i - D_i'a - X_i' \hat{\beta}(a) \leq 0)) \Psi(a, \hat{\delta}(a)). \]

(10)

Accordingly,

\[ \hat{\alpha} = \arg \min_{a \in A} N \hat{g}_N(a)' \hat{\Sigma}(a, a)^{-1} \hat{g}_N(a). \]

More importantly, the aforementioned double machine learning procedure (DML-IVQR hereafter) satisfies the Neyman orthogonality conditions, cf. Chernozhukov et al. (2018b).

2.3 Weak-Identification Robust Inference

Under the regularity conditions listed in Chernozhukov and Hansen (2008), the asymptotic normality of the GMM estimator with a nonsmooth objective function is guaranteed. We have

\[ \sqrt{n} \hat{g}_N(a) \xrightarrow{d} N(0, \Sigma(a, a)). \]

(11)

Consequently, it leads to

\[ N \hat{g}_N(a)' \hat{\Sigma}(a, a)^{-1} \hat{g}_N(a) \xrightarrow{d} \chi^2_{\dim(Z)}. \]
We define
\[ W_N \equiv N \hat{g}_N(a)' \hat{\Sigma}(a,a)^{-1} \hat{g}_N(a). \]
It then follows that a valid \((1-p)\) percent confidence region for the true parameter, \(\alpha_0\), may be constructed as the set
\[ CR := \{ \alpha \in \mathcal{A} : W_N(\alpha) \leq c_{1-p} \}, \]
where \(c_{1-p}\) is the critical point such that
\[ P[\chi_{\text{dim}(Z)}^2 > c_{1-p}] = p, \]
and \(\mathcal{A}\) can be numerically approximated by the grid \(\{\alpha_j, j = 1, \ldots, J\}\).

### 2.4 Algorithms for L1-norm Penalized Quantile Optimization

The suggested double machine learning algorithm involves solving L1-norm optimization which is a nontrivial task. Researchers often represent the L1-norm penalized quantile objective function as a linear programming problem. Specifically,
\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{N} \rho_\tau(Y_i - \theta_0 - W_i' \theta) + \lambda \| \theta \|_1 \\
\text{subject to} & \quad \theta_0 + x_i' \theta + \xi_i = y_i, \ i = 1, \ldots, n.
\end{align*}
\]

where \(\theta = [\alpha', \beta']'\) and \(W = [D', X']'\).

However, it turns out that the computation is challenging and time-consuming. For instance, it often meets the singular design within the high dimensional framework. As an alternative, we utilize the algorithm developed by Yi and Huang (2017) who use the Huber loss function to approximate the quantile loss function. In the equation (12), \(\rho_\tau\) is not differentiable, and
\[ \rho_\tau(t) = (1 - \tau)t_- + \tau t_+ = \frac{1}{2} |t| + (2\tau - 1)t. \]
Since $h_\tau(t) \to |t|$ as $\tau \to 0^+$, where $h_\tau(t)$ is the Huber loss function of $t$ defined in Yi and Huang (2017), we have $\rho_\tau(t) \approx \frac{1}{2}h_\tau(t) + (2\tau - 1)t$ for small $\tau$. Therefore the equation (12) can be approximated by

$$\min_{\theta_0 \in \mathbb{R}, \theta \in \mathbb{R}^p} \sum_{i=1}^{N} h_\tau(Y_i - \theta_0 - W_i'\theta) + (2\tau - 1)(Y_i - \theta_0 - W_i'\theta) + \lambda \|\theta\|_1. \quad (13)$$

The optimization above stands for the Huber approximation. This optimization problem is more computationally feasible for the sake of the differentiability of the loss function.

3 Monte Carlo Experiments

We evaluate the finite-sample performance, in terms of RMSE and MAD, of the double machine learning for the IVQR. The following data generating process is modified from the one considered in Chen and Lee (2018).

$$\begin{bmatrix} u_i \\
\epsilon_i \end{bmatrix} \sim N \left( 0, \begin{bmatrix} 1 & 0.3 \\
0.3 & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} x_i \\
z_i \\
v_i \end{bmatrix} \sim N(0, I)$$

$$Z_i = z_i + v_i + x_i$$

$$D_i = \Phi(z_i + \epsilon_i)$$

$$X_i = \Phi(x_i)$$

$$Y_i = 1 + D_i + X_i^T 1 + D_i * u_i,$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Consequently,

$$\alpha(\tau) = 1 + F^{-1}_\epsilon(\tau),$$

where $F_\epsilon(\cdot)$ is the cumulative distribution function of $\epsilon$. 

7
3.1 Partialing out and nonPartialing out $Z$ on $X$

We focus on comparing MAD and RMSE resulting from different models under the exact specification (10 control variables). po-GMM stands for doing partialing out $Z$ on $X$. GMM stands for doing no partialing out $Z$ on $X$. Table 1 shows that doing partialing out $Z$ on $X$ leads to an efficiency gain across quantiles especially when sample size is moderate.

Table 1: Partialing out and nonPartialing out $Z$ on $X$

|                   | n = 500 |            | n = 1000 |            |
|-------------------|---------|------------|----------|------------|
|                   | RMSE    | MAD        | RMSE     | MAD        |
| $\alpha_{0.1}$ (po-GMM) | 0.1888  | 0.1510     | 0.1219   | 0.0950     |
| $\alpha_{0.1}$ (GMM)       | 0.4963  | 0.2559     | 0.1631   | 0.1138     |
| $\alpha_{0.25}$ (po-GMM)  | 0.1210  | 0.0966     | 0.0812   | 0.0654     |
| $\alpha_{0.25}$ (GMM)      | 0.1782  | 0.1179     | 0.0963   | 0.0754     |
| $\alpha_{0.5}$ (po-GMM)   | 0.0989  | 0.0716     | 0.0689   | 0.0436     |
| $\alpha_{0.5}$ (GMM)       | 0.1436  | 0.1016     | 0.0801   | 0.0542     |
| $\alpha_{0.75}$ (po-GMM)  | 0.1374  | 0.1066     | 0.0828   | 0.0676     |
| $\alpha_{0.75}$ (GMM)      | 0.2403  | 0.1710     | 0.1146   | 0.0848     |
| $\alpha_{0.90}$ (po-GMM)  | 0.2437  | 0.1839     | 0.1391   | 0.1067     |
| $\alpha_{0.90}$ (GMM)      | 0.8483  | 0.5340     | 0.3481   | 0.1967     |

The data generating process considers ten control variables. po-GMM stands for doing partialing out $Z$ on $X$. GMM stands for doing no partialing out $Z$ on $X$.

3.2 IVQR with High-dimensional Controls

We now evaluate the finite-sample performance of the IVQR with high-dimensional controls. The data generating process involves 100 control variables with an approximate sparsity structure. In particular, the exact model (true model) depends only on 10 relevant control variables out of the 100 controls. GMM uses 100 control variables without regularization. Table 2 shows that the RMSE and MAD stemmed from the DML-IVQR are close to those from the exact model. In addition, Figure 1 plots distributions of the IVQR estimator with/without double machine learning. The DML-IVQR stands for the double machine learning for the IVQR with high-dimensional controls. Histograms signify that the DML-IVQR estimator is more efficient and less biased than the IVQR using many control variables. Since a weak-identification robust inference procedure results naturally form the IVQR, cf. Chernozhukov and Hansen (2008), we construct the robust confidence regions for the GMM and the DML-IVQR estimators. Figure 2 signifies that,
### Table 2: IVQR with High-dimensional Controls

| α       | n = 500   | n = 1000  |
|---------|-----------|-----------|
|         | RMSE      | MAD       | RMSE      | MAD       |
| α₀.₁(GMM) | 0.7648    | 0.6645    | 0.3917    | 0.3442    |
| α₀.₁(exact-GMM) | 0.1888    | 0.1510    | 0.1219    | 0.0950    |
| α₀.₁(DML-IVQR) | 0.3112    | 0.2389    | 0.1376    | 0.1085    |
| α₀.₂₅(GMM)  | 0.2712    | 0.2212    | 0.1646    | 0.1361    |
| α₀.₂₅(exact-GMM) | 0.1210    | 0.0966    | 0.0812    | 0.0654    |
| α₀.₂₅(DML-IVQR) | 0.1562    | 0.1254    | 0.0991    | 0.0804    |
| α₀.₅(GMM)   | 0.1627    | 0.1234    | 0.1038    | 0.0754    |
| α₀.₅(exact-GMM) | 0.0989    | 0.0716    | 0.0689    | 0.0436    |
| α₀.₅(DML-IVQR) | 0.1168    | 0.0846    | 0.0775    | 0.0510    |
| α₀.₇₅(GMM)  | 0.3421    | 0.2806    | 0.1747    | 0.1452    |
| α₀.₇₅(exact-GMM) | 0.1374    | 0.1066    | 0.0828    | 0.0676    |
| α₀.₇₅(DML-IVQR) | 0.1495    | 0.1167    | 0.0930    | 0.0741    |
| α₀.₉(GMM)   | 0.9449    | 0.8032    | 0.4320    | 0.3681    |
| α₀.₉(exact-GMM) | 0.2437    | 0.1839    | 0.1391    | 0.1067    |
| α₀.₉(DML-IVQR) | 0.3567    | 0.2608    | 0.1649    | 0.1231    |

across quantiles, the weak-identification (or weak-instrument) robust confidence region based on the DML-IVQR is relatively sharp. The Monte Carlo experiments show that the DML-IVQR procedure performs well.
Figure 1: Histograms of the IVQR Estimator with/without the DML

Notice: DML-IVQR results are plotted in green. Results from the GMM with many controls are in orange.
4 Empirical Applications

4.1 Quantile treatment effects of 401(k) participation on accumulated wealth

We reinvestigate impact of the 401(k) participation on accumulated wealth. Total wealth or net financial asset is the outcome variable $Y$. Treatment variable $D$ is a binary variable standing for participation in the 401(k) plan. Instrument $Z$ is an indicator for being eligible to enroll in the 401(k) plan. The vector of covariates $X$ consists of income, age, family size, married, an IRA individual retirement account, a defined benefit status indicator, a home ownership indicator and the different education-year indicator variables. The data consists of 9915 observations.

Following the regression specification in Chernozhukov and Hansen (2004), Table 3 presents quantile treatment effects obtained from different estimation procedures which have been defined in the previous section including IVQR, po-GMM and GMM. The corresponding results are similar. As to the high-dimensional analysis, we create 119 technical control variables including those constructed by the polynomial bases, interaction terms, and cubic splines (thresholds). To ensure each basis has equal length, we utilize the minimax normalization for all technical control variables. Consequently, we use the plug-in method to determine the value of penalty when doing the LASSO under the moment condition, and tune the penalty in the quantile L1-norm objective function based on the Huber approximation by 5-fold cross validation. The DML-IVQR also implements
Table 3: Estimations with Chernozhukov and Hansen (2004)’s Specification

| Quantiles | 0.1 | 0.15 | 0.25 | 0.5 | 0.75 | 0.85 | 0.9 |
|-----------|-----|------|------|-----|------|------|-----|
| TW(IVQR)  | 4400 | 5300 | 4900 | 6700 | 8000 | 8300 | 10800 |
| TW(po-GMM)| 4400 | 5100 | 4900 | 6300 | 7500 | 7500 | 9100 |
| TW(GMM)   | 4400 | 5200 | 4800 | 6300 | 8400 | 8000 | 8700 |
| NFTA(IVQR)| 3600 | 3600 | 3700 | 5700 | 13200 | 15800 | 17700 |
| NFTA(po-GMM)| 3500 | 3600 | 3700 | 5600 | 13900 | 15800 | 17700 |
| NFTA(GMM) | 3500 | 3600 | 3700 | 5700 | 13900 | 16100 | 18200 |

feature normalization of the outcome variable for the sake of computational efficiency. To make the estimated treatment effects across different estimation procedures roughly comparable, Table 4 shows the effect obtained through the DML-IVQR multiplied by the standard deviation of the outcome variable. Weak identification/instrument robust inference on quantile treatment effects are depicted in Figures 4 and 5. Yet, the robust confidence interval widens as the sample size becomes fewer at the upper quantiles; estimated quantile treatment effects are significantly different from zero. We could use the result from the DML-IVQR as a data-driven robustness check on those summarized in the Table 3.

Tables 5 and 6 present the selected important variables across different quantiles. The approximate sparsity is asymmetric across the conditional distribution in the sense that the number of selected variables decreases as the quantile index $\tau$ increases. However, it hinges on the relatively small number of observations at the upper quantiles as well. Our empirical results also signify that the 401(k) participants with low savings propensity are more associated with the nonlinear income effect than those with high savings propensity, which complements the results concluded in Chernozhukov et al. (2018a) and Chiou et al. (2018). In this particular example, $\tau$ captures the rank variable which governs the unobservable heterogeneity: savings propensity. Small values of $\tau$ represent participants with low savings propensity. The nonlinear income effects, across quantile ranging from $(0, 0.5]$, are picked up by the selected variables such as $\max(0, inc - 0.2)$, $\max(0, inc^2 - 0.2)$, $\max(0, inc^3 - 0.2)$ and etc. Technical variables in terms of age, education, family size, and income are more frequently selected. In addition, these four variables are also identified as important variables in the context of the generalized random forests, cf. Chen and Hsiang (2019).
Figure 3: Effects of Participation in the 401(K) on Total Wealth (TW) and Net Financial Assets (NFTA) respectively.
We create 119 technical control variables including those constructed by the polynomial bases, interaction terms, and cubic splines (thresholds). The DML-IVQR estimates the distributional effect which signifies an asymmetric pattern similar to the one identified in Chernozhukov and Hansen (2004).

---

**Table 4: DML-IVQR with High-dimensional Controls**

| Quantiles | 0.1  | 0.15 | 0.25 | 0.5  | 0.75 | 0.85 | 0.9  |
|-----------|------|------|------|------|------|------|------|
| NFTA(std-DML-IVQR ×63522) | 3176 | 3049 | 3303 | 5844 | 18802 | 26298 | 28076 |
| TW(std-DML-IVQR ×111529) | 2453 | 3011 | 3457 | 7695 | 15056 | 18736 | 16394 |
| NFTA(std-DML-IVQR) | 0.05 | 0.048 | 0.052 | 0.092 | 0.296 | 0.414 | 0.442 |
| TW(std-DML-IVQR) | 0.022 | 0.027 | 0.031 | 0.069 | 0.135 | 0.168 | 0.147 |

---

**Figure 4: Weak Instrument Robust Inference, P401(K) on TW with hqreg L1-norm**

---
Figure 5: Weak Instrument Robust Inference, P401(K) on NFTA with hqreg L1-norm
| Quantile | Selected Variables |
|----------|-------------------|
| 0.15     | \( \text{ira, educ, } \text{educ}^2, \text{age} \ast \text{ira, age} \ast \text{inc, fsize} \ast \text{educ, fsize} \ast \text{hmort} \) |
|          | \( \text{ira} \ast \text{educ, } \text{ira} \ast \text{inc, hval} \ast \text{inc, marr, male, i4, a3} \) |
|          | \( \text{twoearn, marr} \ast \text{fsize}, \text{pira} \ast \text{inc, max}(0, \text{age}^3 - 0.2) \) |
|          | \( \text{max}(0, \text{educ}^2 - 0.4), \text{max}(0, \text{educ} - 0.2), \text{max}(0, \text{age}^2 - 0.4) \) |
| 0.25     | \( \text{ira, age} \ast \text{fsize, age} \ast \text{ira, age} \ast \text{inc} \) |
|          | \( \text{fsize} \ast \text{educ, } \text{ira} \ast \text{educ, } \text{ira} \ast \text{inc} \) |
|          | \( \text{hval} \ast \text{inc, marr, male, i3, twoearn, marr} \ast \text{fsize} \) |
|          | \( \text{pira} \ast \text{inc, twoearn} \ast \text{fsize}, \text{max}(0, \text{inc} - 0.2) \) |
| 0.5      | \( \text{inc}^2, \text{age} \ast \text{fsize, age} \ast \text{ira, age} \ast \text{inc} \) |
|          | \( \text{fsize} \ast \text{educ, } \text{ira} \ast \text{educ, } \text{ira} \ast \text{inc} \) |
|          | \( \text{hval} \ast \text{inc, male, a1, a3 }, \text{pira} \ast \text{inc, twoearn} \ast \text{age}, \text{twoearn} \ast \text{fsize} \) |
|          | \( \text{twoearn} \ast \text{hmort, twoearn} \ast \text{educ, max}(0, \text{educ} - 0.6) \) |
| 0.75     | \( \text{inc, ira, age} \ast \text{ira, age} \ast \text{hval} \) |
|          | \( \text{age} \ast \text{inc, educ} \ast \text{inc, hval} \ast \text{inc, pira} \ast \text{inc, pira} \ast \text{age} \) |
| 0.85     | \( \text{inc, ira, age} \ast \text{hval, age} \ast \text{inc, ira} \ast \text{educ} \) |
|          | \( \text{educ} \ast \text{inc, hval} \ast \text{inc, pira} \ast \text{inc, pira} \ast \text{hval} \) |

Selected variables across \( \tau \), tuned via cross validation.

ira: individual retirement account (IRA), inc: income, fsize: family size, hequity: home equity, hval: home value, educ: education years, marr: married, smcol: college, db: defined benefit pension, hown: home owner, hmort: home mortgage, a1: less than 30 years old, a2: 30-35 years old, a3: 36-44 years old, a4: 45-54 years old, a5: 55 years old or older, i1: < $10K, i2: $10 – 20K, i3: $20 – 30K, i4: $30 – 40K, i5: $40 – 50K, i6: $50 – 75K, i7: $75K+. 
Table 6: Net Financial Assets

| Quantile | Selected Variables |
|----------|---------------------|
| 0.15     | $ira, educ^2, fsize^3, hval^4, educ^5, age * educ, age * hmort$ $age * inc, fsize * hmort, fsize * inc, ira * educ, ira * inc$ $hval * inc, marr, db, male, i2, i3$ $i4, i5, twoearn, marr * fsize$ $pira * inc, pira * educ, twoearn * inc, twoearn * ira$ $max(0, age^3 - 0.2), max(0, age^2 - 0.2), max(0, age - 0.6)$ $max(0, inc^3 - 0.2), max(0, inc^2 - 0.2), max(0, educ - 0.2)$ |
| 0.25     | $ira, hmort, age * hmort, age * inc, fsize * hmort, fsize * inc$ $ira * educ, ira * inc, hval * inc, db, smcol, male$ $i2, i3, i4, i5, a2, a3$ $twoearn, pira * inc, pira * age$ $pira * fsize, twoearn * inc, twoearn * ira$ $twoearn * hmort, max(0, age^2 - 0.2)$ $max(0, age - 0.6), max(0, inc^2 - 0.2), max(0, inc - 0.4)$ $max(0, inc - 0.2), max(0, educ - 0.2)$ |
| 0.5      | $age, ira, age * fsize, age * ira, age * inc$ $fsize * educ, fsize * hmort, ira * educ, ira * inc, hval * inc, hown$ $male, i3, i4, a1, a2, a4, pira * inc$ $pira * fsize, twoearn * inc, twoearn * fsize$ $twoearn * hmort, twoearn * educ, max(0, inc - 0.2)$ |
| 0.75     | $ira, age * inc, hval * inc, pira * inc, pira * age$ |
| 0.9      | $ira, age * inc, educ * inc, hval * inc, pira * inc$ |

Selected variables across $\tau$, tuned via cross validation.

$ira$: individual retirement account (IRA), $inc$: income, $fsize$: family size, $hequity$: home equity, $hval$: home value, $educ$: education years, $marr$: married, $smcol$: college, $db$: defined benefit pension, $hown$: home owner, $hmort$: home mortgage, $a1$: less than 30 years old, $a2$: 30-35 years old, $a3$: 36-44 years old, $a4$: 45-54 years old, $a5$: 55 years old or older, $i1$: $< \$10K$, $i2$: $\$10 - \$20K$, $i3$: $\$20 - \$30K$, $i4$: $\$30 - \$40K$, $i5$: $\$40 - \$50K$, $i6$: $\$50 - \$75K$, $i7$: $\$75K+$.  

17
4.2 Effects of subsidized training on male and female trainee earnings

Abadie, Angrist and Imbens (2002) use the Job Training Partnership Act (JTPA) data to estimate the quantile treatment effect of job training on the earning distribution. The data are from Title II of the JTPA in early 1990s, which consist of 11,204 samples, 5,102 of them are male, and 6,102 of them are female. In estimation, they take thirty-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. The control variables include the binary variables of black and Hispanic applicants, high-school graduates, married applicants, 5 age-group, AFDC receipt (for women), whether the applicant worked at least 12 weeks in the 12 months preceding random assignment, the dummies for the original recommended service strategy (classroom, OJT/JSA, other) and a dummy for whether earnings data are from the second follow-up survey.

Table 7 presents quantile treatment effects for male and female groups respectively obtained from several estimation procedures including IVQR, po-GMM, and GMM. As to the high-dimensional analysis, we create 85 technical control variables including those constructed by the polynomial bases, interaction terms, and cubic splines (thresholds). Table 8 shows the quantile treatment effect obtained through the DML-IVQR. Table 7 together with the existing findings in the literature suggest that for female only, job training program generates significantly positive treatment effect on earnings at 0.5 and 0.75 quantiles. The DML-IVQR signifies similar results, which can be confirmed by the identification-robust confidence intervals depicted in Figures 6 and 7. The selected variables are collected in the online appendix. Thus, the existing empirical conclusions in the literature is reassured by the IVQR using double machine learning procedure.

Table 7: Estimations with Abadie et. al. (2002)’s Specification

| Quantiles | 0.1 | 0.15 | 0.25 | 0.5 | 0.75 | 0.85 | 0.9 |
|-----------|-----|------|------|-----|------|------|-----|
| Male(IVQR) | 0   | -200 | 400  | 500 | 3300 | 3100 | 1700|
| Male(po-GMM) | 0   | -100 | 500  | 1900 | 5000 | 6800 | 7800|
| Male(GMM) | 0   | -100 | 500  | 1600 | 5100 | 5800 | 7200|
| Female(IVQR) | 0   | 0    | 400  | 1600 | 5100 | 5800 | 1400|
| Female(po-GMM) | 0   | 200  | 700  | 3300 | 5200 | 6500 | 6900|
| Female(GMM) | 100 | 200  | 700  | 3200 | 5200 | 6500 | 6900|

1Selected variables for the male group: [https://github.com/FieldTien/DML-QR/blob/master/Empirical_work/hqreg_data/selected_male.csv](https://github.com/FieldTien/DML-QR/blob/master/Empirical_work/hqreg_data/selected_male.csv) selected variables for the female group: [https://github.com/FieldTien/DML-QR/blob/master/Empirical_work/hqreg_data/selected_female.csv](https://github.com/FieldTien/DML-QR/blob/master/Empirical_work/hqreg_data/selected_female.csv)
Table 8: DML-IVQR with High-dimensional Controls

| Quantiles | 0.1 | 0.15 | 0.25 | 0.5 | 0.75 | 0.85 | 0.9 |
|-----------|-----|------|------|-----|------|------|-----|
| Male(std-DML-IVQR ×19400) | 0   | -97  | -97  | 0   | 2735 | 2735 | 488 |
| Female(std-DML-IVQR ×13400) | 0   | 67   | 335  | 1274 | 2213 | 1676 | 268 |
| Male(std-DML-IVQR) | 0   | -0.005 | -0.005 | 0   | 0.14 | 0.14 | 0.025 |
| Female(std-DML-IVQR) | 0   | 0.005 | 0.025 | 0.095 | 0.165 | 0.125 | 0.02 |

We create 85 technical control variables including those constructed from the polynomial bases, interaction terms, and cubic splines (thresholds).

Figure 6: Weak Instrument Robust Inference. The male group.
Figure 7: Weak Instrument Robust Inference. The female group.
5 Conclusion

The performance of a debiased/double machine learning algorithm within the framework of high-dimensional IVQR is investigated. The simulation results signify that the proposed procedure performs more efficiently than those based on the conventional estimator with many controls. Furthermore, we evaluate the corresponding weak-identification robust confidence interval of the low-dimensional causal parameter. Given a large number of technical controls, we reinvestigate quantile treatment effects of the 401(k) participation on accumulated wealth and then highlight the non-linear income effects driven by the the savings propensity.
References

Abadie A. Angrist J. and G. Imbens. 2002. “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” Econometrica, 70(1): 91–117.

Athey, S. 2017. “Beyond Prediction: Using Big Data for Policy Problems,” Science, 355: 483–485.

Athey, S. 2018. “The Impact of Machine Learning on Economics,” working paper, Stanford GSB.

Athey, S., Tibshirani, J., and S. Wager. 2019. “Generalized Random Forests,” The Annals of Statistics, 47(2): 1148–1178.

Belloni, A., Chernozhukov V., and C. Hansen. 2014. “High-Dimensional Methods and Inference on Structural and Treatment Effects,” Journal of Economic Perspectives, 28: 29–50.

Chen, J.-E. and C.-W. Hsiang. 2019. “Causal Random Forests Model using Instrumental Variable Quantile Regression,” working paper, Center for Research in Econometric Theory and Applications, National Taiwan University.

Chen, L.-Y., and S. Lee. 2018. “Exact Computation of GMM Estimators for Instrumental Variable Quantile Regression Models,” Journal of Applied Econometrics, forthcoming.

Chernozhukov, V., and C. Hansen. 2004. “The Impact of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis,” Review of Economics and Statistics, 86: 735–751.

Chernozhukov, V. and C. Hansen. 2008. “Instrumental Variable Quantile Regression: A Robust Inference Approach,” Journal of Econometrics, 142: 379–398.

Chernozhukov, V., Hansen C., and M. Spindler. 2015. “Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach,” Annual Review of Economics, 7: 649–688.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and J. Robins. 2018a. “Double/debiased Machine Learning for Treatment and Structural
Chernozhukov, V., Hansen, C., and K. Wüthrich. 2018b. “Instrumental Variable Quantile Regression,” *Handbook of Quantile Regression*.

Chiou, Y.-Y., Chen, M.-Y., and J.-E. Chen. 2018. “Nonparametric Regression with Multiple Thresholds: Estimation and Inference,” *Journal of Econometrics*, 206(2): 472–514.

Yi, C., and J. Huang. 2017. “Semismooth Newton Coordinate Descent Algorithm for Elastic-net Penalized Huber Loss Regression and Quantile Regression,” *Journal of Computational and Graphical Statistics*, 26(3): 547–557.