Endogenous High-Dimensional Quantile Regression: A Control Function Approach*

Kaixi Zhang†

Department of Economics, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR, China

This version: August 8, 2025

Abstract

This paper presents a double selection estimator based on the Control Function approach for estimating a high-dimensional Quantile Regression model with endogeneity (CFQR). We have several advantages over the Instrumental Quantile Treatment Effect model (henceforth IVQR), which was proposed by Chernozhukov and Hansen (2005). First, our model is computationally simpler than high-dimensional IVQR under general conditions. IVQR relies on a hypothesized value for coefficients of interest in the first step of estimation. The iterative nature of this process is costly, especially when dealing with big data. In contrast, our methods use the function of residuals obtained from the previous step as an additional exogenous control, which is more practical for application. Furthermore, we provide inference methods for the coefficients of interest in endogenous high-dimensional quantile regression models based on orthogonal score functions, which mitigate the effects of moderate selection errors. Monte Carlo simulations also show that our estimator performs well under high-dimensional controls. In addition, we could accommodate more than one endogenous treatment depending on different economic settings. Finally, we employ our model to investigate the impact of compulsory schooling on earnings using 1530 instruments for education based on Angrist and Krueger (1991)'s research. We find that high-dimensional quantile estimates are smaller than 2SLS and OLS estimates. Schooling returns appear to be lower than in previous studies.

Keywords: endogeneity, quantile regression, control function approach, high-dimensional model, returns to schooling

JEL Codes: C21, C36, J31

*We thank

†PhD student currently in 4th year. Email: kzhangbh@connect.ust.hk (K. Zhang).

1 Introduction

Endogeneity is a crucial issue that should be taken into account in empirical studies. A typical example is the study of schooling's impact on wages. Angrist and Krueger (1991) explain the endogeneity of schooling in the wage equation and provide three quarter-of-birth dummy variables as instruments in light of American compulsory schooling laws and the shape of the life cycle earnings profile. Hansen et al. (2008) based on these three quarter-of-birth dummies, added their interactions with year-of-birth and state-of-birth, totaling 180 instruments to further explore this problem. With many instrumental variables, however, conventional two-stage least squares asymptotic approximations can be poor at inference procedures. While Hansen et al. (2008) adjust the asymptotic theory accordingly, Belloni et al. (2011) consider the high-dimensional sparse regression models with 1530 instruments for the first-stage estimation as an alternative to any of the aforementioned methods. High-dimensional sparse regression models (HDSMs) arise from the wide availability of data sets that contain many potential control variables and the regression function is well approximated by an unknown set of controls. The number of controls p is typically very large, perhaps greater than n, the sample size, but only at most s controls contribute to the explanation of the response variable, where s grows slowly than n. Most literature primarily focuses on penalized mean regression to estimate HDSMs with l_1-norm serving as a penalty function.

This paper considers endogenous quantile regression in high-dimensional sparse models instead. For example, one might be more interested in the effect of schooling on the lower or higher tail of the wage distribution conditional on individual characteristics than in its impact on the mean. Quantile regression (QR) models are a valuable empirical tool in economic analysis for capturing the heterogeneous impact of controls on the conditional distribution of a response variable. As the asymptotic theory of ordinary quantile regression may not provide an accurate approximation of the behavior of the estimators among high-dimensional controls, Belloni and Chernozhukov (2011) present l_1-penalized quantile regression estimators, which penalize the l_1-norm of regression coefficients, as well as the post-penalized QR estimator, which applies ordinary quantile regression to the model selected by l_1-penalized QR. On the other hand, Chernozhukov and Hansen (2005) propose instrumental variable quantile regression (IVQR) to build a quantile model with endogeneity. To estimate the coefficients of a linear IVQR model, the most direct way is to use the unconditional moment conditions implied by their assumptions and theorems. Chernozhukov and Hansen (2006) also take a different approach called inverse quantile regression (IQR) for estimation rather than the conventional GMM framework. In a setting with high-dimensional controls, Neyman orthogonality conditions can be used to avoid the potentially poor performance of the estimators based directly on the methods discussed above (Chernozhukov et al. (2017)). Obviously, researchers can construct a quantile model with endogeneity by many other methods (Amemiya (1982), Powell (1983), Abadie et al. (2002), Blundell and Powell (2007)). Lee (2007) adjusts for endogeneity by adopting a control function approach developed in Imbens and Newey (2009) and presents a simple two-step estimator that exploits the partially linear structure of the quantile regression model. Therefore, this paper was motivated to fill the gap in the literature by developing an endogenous quantile

regression model using the control function (CF) approach among high-dimensional controls.

This paper contributes to the literature in the following aspects: First, our approaches are computationally simpler compared to IVQR. Since the objective function of IVQR is both nonsmooth and noncovex in general, directly solving the optimization problem poses a substantial computational challenge. Inverse quantile regression (IQR, Chernozhukov and Hansen (2008)) is a feasible method for solving this problem, but the regression coefficients are obtained by grid search. A scalar endogenous variable is computationally inexpensive and can be formulated as a single parametric linear programming exercise, but they quickly become intractable for higher dimensional endogenous variables. In contrast, we propose an estimator for estimating a high-dimensional quantile regression model with endogeneity inspired by the control function approach. Iteration is avoided through the use of the function of residuals obtained from the previous step as an additional exogenous control. Moreover, as a natural extension of control function approach, we could use more than one endogenous treatment depending on the economic settings. Second, we extend the idea of double selection introduced by Belloni et al. (2014) to quantile analysis. Based on robust post-selection procedures, we construct estimates and confidence ranges for the endogenous coefficient of interest, filling a gap in the literature on high-dimensional sparse regression models. Third, we revisit the impact of compulsory schooling on earnings using 1530 instruments for education. High-dimensional quantile estimates of schooling return are smaller than 2SLS and OLS estimates, suggesting that schooling returns appear to be lower than in previous studies.

We organize the paper as follows. In the following section, in order to clarify our motivation, we briefly review the estimation procedures of two widely used IV quantile models for fixed p: the instrumental variable quantile regression (IVQR) model (Chernozhukov and Hansen (2005)) and the triangular models (Imbens and Newey (2009)). In section 3, we propose our double selection estimators, estimation procedures and asymptotic properties. Monte Carlo simulation of our estimator is presented in section 4. Section 5 applies our methods to estimate the impact of schooling on wages. The last part contains concluding comments. The theorem proofs are in the Appendix.

2 Review

2.1 IVQR model

Chernozhukov and Hansen (2005) focus on linear-in-parameter structural quantile regressions

$$Q_{Y|X}(\tau) = D'\alpha(\tau) + X'\beta(\tau)$$

where $\alpha(\tau)$ captures the causal effect of the endogenous variables D on the τ-th quantile of the conditional distribution of potential outcomes Y given exogenous covariates X. Under suitable assumptions, the identification is given by

$$P(Y \leq Q_{Y|X}(\tau)|X, Z) = \tau$$

which implies unconditional moment conditions

$$E[(\tau - 1\{Y < D'\alpha + X'\beta\})\varphi(X, Z)] = 0 \tag{2.1}$$

where Z is instruments. $\varphi(X, Z) = (\Phi(X, Z), X')'$ and $\Phi(X, Z)$ is a (known) transformation of X and Z. In practice, they take $\Phi(X, Z)$ to be the linear projection of (Z, X) onto D.

2.1.1 The method of inverse quantile regression (IQR)

IQR is based on the unconditional moment conditions 2.1, coupled with the linear quantile model, implies that the τ-th quantile of $Y - D'\alpha_0$ conditional on covariates X and instruments Z is equal to $X'\beta_0(\tau)$:

$$Q_{Y-D'\alpha}(\tau|X, Z) = X'\beta_0 + Z'\gamma_0 \quad \text{with } \gamma_0 \equiv 0$$

This implies that, at true value of α_0, the ordinary linear τ-quantile regression of $Y - D'\alpha_0$ on X and Z would yield coefficients on the instruments of exactly 0 in the population. We can then define the IQR estimator of α_0 as

$$\hat{\alpha} = \arg\min_{a \in A} W_N(a)$$

where A is the parameter space for α and

$$W_N(a) = N\hat{\gamma}(a)'\hat{\Omega}_N(a)^{-1}\hat{\gamma}(a)$$

where $\hat{\Omega}_N(a)$ denotes the estimated covariance matrix of $\sqrt{N}(\hat{\gamma}(a) - \gamma(a))$, and note that this covariance matrix is available in any common implementation of the ordinary quantile regression.

2.1.2 Estimation of IQR

The IQR method can be implemented as follows:

Step 1: For a given quantile τ of interest, define a grid of values $\alpha_j, j = 1, ..., J$ and run the ordinary τ-quantile regression of $Y - D'\alpha_j$ on X and $\varphi(X, Z)$ to obtain coefficients $\hat{\beta}(\alpha_j, \tau)$ and $\hat{\gamma}(\alpha_j, \tau)$, where $\hat{\gamma}(\alpha_j, \tau)$ is the coefficient on $\varphi(X, Z)$. The following implementation uses the linear projection of D on (X, Z) as an instrument.

Step 2: Choose $\hat{\alpha}(\tau)$ as the value among the grid $\alpha_j, j = 1, ..., J$ that makes $W_N(a)$ smallest. The estimate $\hat{\beta}(\tau)$ is then given by $\hat{\beta}(\hat{\alpha}(\tau), \tau)$.

2.2 Control function approach in quantile analysis

For a better understanding of the control function approach, we first consider the following triangular simultaneous equation model

$$Y = g(D, \epsilon)$$
$$D = h(Z, \eta)$$

where Y is the outcome variable, D is a continuous scalar endogenous variable, Z is a vector of continuous instruments, ϵ and η are both scalar reduced-form errors, and we ignore covariates X for simplicity. Quantile effects can be identified in this triangular system if the following conditions are met (Imbens and Newey (2009)): (i) $(\epsilon, \eta) \perp Z$ (ii) η is a continuously distributed scalar with CDF that is strictly increasing on the support of η and $h(Z, t)$ is strictly monotonic in t with probability 1. Then D and ϵ are independent conditional on $V = F_{D|Z}(D, Z) = F_\eta(\eta)$. Therefore, V can be used as a control variable with the conditional quantile restriction. The second stage estimates a quantile regression model for the response variable of interest, including the estimated control variable to deal with endogeneity.

2.2.1 Estimation of the control variable V

Chernozhukov et al. (2015) summarize several ways to estimate control variable V in the first step. Since

$$\hat{V} = \hat{F_{D|Z}}(d|z) = \hat{Q^{-1}_{D|Z}}(d|z)$$

in the classical additive location model

$$Q_{D|Z}(v|Z) = \pi_0 Z + Q_V(v) \tag{2.2}$$

we have $V = Q_V^{-1}(D - \pi_0 Z)$, which can be estimated by the empirical CDF of the OLS residuals. For example, if $D|Z \sim N(\pi_0 Z, \sigma^2)$, the control variable has the parametric form $V = \Phi^{-1}((D - \pi_0 Z)/\sigma)$. A non-additive quantile regression model has

$$Q_D(v|Z) = Z'\pi_0(v) \tag{2.3}$$

and

$$V = Q_D^{-1}(v|Z) = \int_{(0,1)} 1\{Z'\pi_0(v) \leq D\} dv$$

To avoid potential non-invertibility of $Q^{-1}_{D|Z}$ caused by nonmonotonicity of $v \to Z'\hat{\pi}(v)$, the estimator takes the form

$$\hat{V} = \tau + \int_{(\tau, 1-\tau)} 1\{Z'\hat{\pi}(v) \leq D\} dv$$

where $\hat{\pi}(v)$ is an ordinary quantile regression estimator, τ is a small positive trimming cut-off that prevents estimation of tail quantiles (Koenker (2005)), and the integral can be approximated numerically using a finite grid of quantiles.

The third method using distribution regression (Chernozhukov et al. (2013))

$$V = F_{D|Z}(d|z) = \Lambda(Z'\pi_0(d))$$

where Λ is known probit or logit link functions. The estimator is

$$\hat{V} = \Lambda(Z'\hat{\pi}(d))$$

where $\hat{\pi}(d)$ is the maximum likelihood estimator of $\pi_0(d)$ at each $d \in support(D)$.

3 Model

we consider the following triangular system of quantile model

$$Y = Q_Y(U|D, Z_1, V) \tag{3.1}$$
$$D = Q_D(\eta|Z) \tag{3.2}$$

where Y is a continuous dependent variable, D is a continuous endogenous variable of interest. $Z = (Z_1', Z_2')'$. Z is a $(d_{Z_1} \times 1)$ vector of exogenous covariates and Z is a $(d_{Z_2} \times 1)$ vector of instrument variables excluded from 3.1. We allow both d_{Z_1} and d_{Z_2} are comparable or larger than the sample size n. η is a unobserved regressor that accounts for endogeneity of D. The function $u \mapsto Q_Y(u|D, Z_1, V)$ is the conditional quantile function of Y given (D, Z_1, V) and $\eta \mapsto Q_D(\eta|Z)$ is the conditional quantile function of D given Z. U and η satisfies the independence assumptions

$$U \sim U(0,1)|D, Z, V$$
$$\eta \sim U(0,1)|Z$$

The control variable V is the function of η. In particular, we impose a semiparametric restriction on the functional form of the conditional quantile function in 3.1. Assume that

$$y_i = d_i\theta(u) + g_u(z_{1i}, v_i) + \epsilon_i \tag{3.3}$$

we shall use p_1 controls $x_{1i} = \mathcal{X}(z_{1i})$ to achieve an accurate approximation to the function g_u, which takes the form

$$g_u(z_{1i}, v_i) = x_{1i}'\beta(u) + P_u(v_i) + r_{ui}$$

where $\mathcal{X}(z_{1i})$ and $P_u(v_i)$ are vector of transformations of the initial variables, such as power series or regression splines, and r_{ui} denotes an approximation error.

To simplify the analysis, we will start with a linear regression model for $Q_D(\eta|Z)$

$$d_i = x_i'\pi_0 + Q_\eta^{-1}(\eta) \tag{3.4}$$

where $x_i = \mathcal{X}(z_i)$ with p controls.

Example (Additive location model):

$$\begin{aligned} y &= \theta_{01}d + z_1'\gamma_{01} + (\theta_{02}d + z_1'\gamma_{02})\epsilon \\ d &= z'\pi_0 + v \end{aligned}$$

where (η, ϵ) are jointly standard bivariate normal with correlation $z_1'\rho$ conditional on $z = (z_1', z_2')'$, $(\theta_{02}d + z_1'\gamma_{02}) > 0$ a.s. By the properties of the normal distribution, $v = \Phi^{-1}(\eta)$ with $\eta \sim U(0,1)$ independent of z and $\epsilon = (z_1'\rho)\Phi^{-1}(\eta) + [1 - (z_1'\rho)^2]^{1/2}\Phi^{-1}(u)$ with $u \sim U(0,1)$ independent of (d, z, η). The corresponding conditional quantile functions are

$$\begin{aligned} Q_y(u|d, z_1, \eta) &= \theta(u)d + z_1'\gamma(u) + \gamma_{02}(z_1'\rho)\Phi^{-1}(\eta)d + \gamma_{02}(z_1'\rho)z_1'\Phi^{-1}(\eta) \\ Q_d(\eta|z) &= z'\pi_0 + \Phi^{-1}(\eta) \end{aligned}$$

In order to perform robust inference with respect to model selection mistakes, we construct a moment condition based on a score function that satisfies an additional orthogonality property that makes them immune to first-order changes in the value of the nuisance parameter. To construct the orthogonal moment condition, we use Belloni et al. (2019)'s method. Letting $f_i^{-1} \equiv f^{-1}_{\epsilon_i|d_i,z_i}(0)$ denotes the inverse function of the conditional density at 0 of the error term ϵ_i in outcome equation 3.3, we have

$$E[((1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)f_i^{-1}(d_i - x_i'\pi_0)] = 0 \tag{3.5}$$

which satisfies the following orthogonality condition with respect to first-order changes in the value of the nuisance parameters

$$\begin{aligned} &\partial_{P_u} E[(1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)f_i^{-1}(d_i - x_i'\pi_0)]|_{P_u = P_u(V)} \\ =&\partial_{P_u} E[E[1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)f_i^{-1}(d_i - x_i'\pi_0)|d, z]]|_{P_u = P_u(V)} \\ =&E[\partial_{P_u} E[(1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)|d, z]f_i^{-1}(d_i - x_i'\pi_0)]|_{P_u = P_u(V)} \\ =&E[f_i f_i^{-1}(d_i - x_i'\pi_0)] = E[d_i - x_i'\pi_0] = 0 \end{aligned}$$

$$\begin{aligned} &\partial_{\beta} E[(1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)f_i^{-1}(d_i - x_i'\pi_0)]|_{\beta = \beta(u)} \\ =&\partial_{\beta} E[E[(1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)f_i^{-1}(d_i - x_i'\pi_0)|d, z]]|_{\beta = \beta(u)} \\ =&E[\partial_{\beta} E[(1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)|d, z]f_i^{-1}(d_i - x_i'\pi_0)]|_{\beta = \beta(u)} \\ =&E[x_{1i}(d_i - x_i'\pi_0)] = 0 \end{aligned}$$

and

$$\begin{aligned}
&\partial_\pi E[(1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)f_i^{-1}(d_i - x_i'\pi_0)]|_{\pi=\pi_0} \\
=&\partial_\pi E[E[(1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u)f_i^{-1}(d_i - x_i'\pi_0)|d,z]]|_{\pi=\pi_0} \\
=&\partial_\pi E[E[1\{y_i \leq d_i\theta(u) + x_{1i}'\beta(u) + P_u(v_i)\} - u|d,z]f_i^{-1}(d_i - x_i'\pi_0)]|_{\pi=\pi_0} \\
=&\partial_\pi E[0 f_i^{-1}(d_i - x_i'\pi_0)]|_{\pi=\pi_0} = 0
\end{aligned}$$

3.1 Estimation

Several different estimators can be constructed using the orthogonal score function 3.5 that have the same first-order asymptotic properties but potentially different finite sample behaviors. We present the details of l_1-penalized least squares and l_1-penalized quantile regression as following

$$\hat{\pi} \in \arg\min_{\pi} \frac{1}{n}\sum_{i=1}^{n}(d_i - x_i'\pi)^2 + \lambda_1\|\pi\|_1$$

and

$$(\hat{\theta}_u, \hat{\beta}_u) \in \arg\min_{\theta,\beta} \frac{1}{n}\sum_{i=1}^{n}\rho_u(y_i - d_i\theta - x_{1i}'\beta - P_u(\hat{v}_i)) + \lambda_u\|\Gamma_u(\theta,\beta)\|_1$$

where ρ_u denotes the check function, Γ_u is a u-dependent diagonal matrix and $P_u(\hat{v}_i)$ is a transformation of v_i.

Following estimation procedure is based on the explicit construction of the orthogonal score function 3.5.

Algorithm:

- Step 0. Compute $\hat{v}_i$ from Lasso estimator of d_i on x_i.
- Step 1. Compute $(\hat{\theta}_u, \hat{\beta}_u)$ using l_1-penalized quantile regression.
- Step 2. Compute $(\tilde{\theta}_u, \tilde{\beta}_u)$ from quantile regression of y_i on d_i, $P_u(\hat{v}_i)$ and $\{x_{1j} : |\hat{\beta}_{uj}| \geq \lambda_u/(\frac{1}{n}\sum_{i=1}^{n} x_{1ij}^2)^{1/2}\}$.
- Step 3. Estimate the conditional density $\hat{f}_i^{-1}$ through 3.6 and 3.7.
- Step 4. Compute $\tilde{\pi}$ from the post-Lasso estimator of d_i on x_i.
- Step 5. Construct the score function

$$\hat{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(u - 1\{y_i - d_i\theta - \hat{x}_{1i}'\tilde{\beta}_u - P_u(\hat{v}_i)\})\hat{f}^{-1}(d_i - x_i'\tilde{\pi})$$

- Step 6. Compute $\check{\theta}_u$ via following GMM estimator

$$\check{\theta}_u = \arg\min_{\theta} n\hat{g}_n(\theta)'\hat{\Sigma}(\theta)^{-1}\hat{g}_n(\theta)$$

where

$$\hat{\Sigma}(\theta) = \frac{1}{n}\sum_{i=1}^{n}[(u - 1\{y_i - d_i\theta - \hat{x}_{1i}'\tilde{\beta}_u - P_u(\hat{v}_i)\})\hat{f}^{-1}(d_i - x_i'\tilde{\pi})]^2.$$

Remark: similar to Belloni et al. (2019), the estimation of conditional density function is based on the observation

$$f_i = \frac{1}{\partial Q(\tau|d_i, z_i)/\partial\tau}$$

then

$$f_i^{-1} = \frac{\hat{Q}(\tau + h|d_i, z_i) - \hat{Q}(\tau - h|d_i, z_i)}{2h} \tag{3.6}$$

when the conditional quantile function is three times continuously differentiable, this estimator is based on the first-order partial difference of the estimated conditional quantile function, and so it has the bias of order h^2. Under additional smoothness assumptions, an estimator that has a bias of order h^4 is given by

$$f_i^{-1} = \frac{\frac{3}{4}\{\hat{Q}(\tau + h|d_i, z_i) - \hat{Q}(\tau - h|d_i, z_i)\} - \frac{1}{12}\{\hat{Q}(\tau + 2h|d_i, z_i) - \hat{Q}(\tau - 2h|d_i, z_i)\}}{h} \tag{3.7}$$

3.2 Asymptotic Results

Theorem 1: Under conditions (I) - (IV), we have

$$\sqrt{n}(\check{\theta}_\tau - \theta_\tau) \to^d N(0, E[d_i v_i]^{-1}\tau(1-\tau)E[\iota_{0i}^2]E[d_i v_i]^{-1})$$

where $v_i = d_i - x_i'\pi_{0\tau}$ and $\iota_i = f_i^{-1}v_i$.
(the conditions (I) - (IV) and the proof can be found in the appendix)

4 Monte Carlo Simulation

We adopt the following data generating processes with sample size of $n = c(100, 200, 400, 600, 1000)$. The dimension p of covariates X is 500, and the dimension s of the true model is 6. Each case replicated 100 times:

$$D = 1 + Z + (1/2)X_1 + (1/3)X_2 + (1/4)X_3 + (1/5)X_4 + \Phi^{-1}(V)$$
$$Y = 1 + D + X_1 + X_2 + X_3 + X_4 + \Phi^{-1}(U)$$

where the instrument Z and all controls $X = \{X_i\}_{i=1}^{498}$ are are independent standard normal $N(0, 1)$ and the error terms.

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim N(0, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix})$$

There are two methods for choosing penalty level λ. The first one proposed by Belloni and Chernozhukov (2011)

$$\lambda = 2\Lambda(1-\alpha|X)$$

where $\Lambda(1-\alpha|X) :=$ $(1-\alpha)$-quantile of Λ conditional on X. The random variable

$$\Lambda = n \sup_{u\in\mathcal{U}} \max_{1\leq j\leq dim(b)} |\frac{1}{n}\sum_{i=1}^{n}[\frac{x_{ij}(u-1\{u_i\leq u\})}{\hat{\sigma_j}\sqrt{u(1-u)}}]|$$

where $u_1, ..., u_n$ are $i.i.d.$ uniform $(0,1)$ random variables that are independently distributed from the controls $x_1, .., x_n$. The second method is cross validation.

Table 1 and table 2 show our simulation results compared to IVQR.

Table 1: Choice of λ, $\tau = 0.5$, $p = 500$ and $n < p$

		$\tau = 0.5$		
	Bias	RMSE	SD	Time(s)
$n < p$		$n = 100$		
Oracle-IQR	0.0187	0.1630	0.1647	1.00
CFQR-HD (λ = 2-fold Cross-Validation)	0.4878	0.4946	0.0831	5.10
CFQR-HD (λ = Belloni and Chernozhukov)	0.3414	0.3621	0.1227	0.52
IVQR-HD (λ = 2-fold Cross-Validation)	0.0359	0.2026	0.2029	414
IVQR-HD (λ = Belloni and Chernozhukov)	0.2733	0.5219	0.4522	30.26
DML-IVQR (λ = 2-fold Cross-Validation)	-0.1900	0.6902	0.6748	42.24
DML-IVQR (λ = Belloni and Chernozhukov)	0.1933	0.2864	0.2149	15.36
$n < p$		$n = 200$		
Oracle-IQR	-0.0307	0.1271	0.1255	1.08
CFQR-HD (λ = 2-fold Cross-Validation)	0.3758	0.3954	0.1249	7.62
CFQR-HD (λ = Belloni and Chernozhukov)	0.5141	0.5174	0.0593	0.72
IVQR-HD (λ = 2-fold Cross-Validation)	0.0897	0.1297	0.0954	514.80
IVQR-HD (λ = Belloni and Chernozhukov)	0.2827	0.3232	0.1593	43.70
DML-IVQR (λ = 2-fold Cross-Validation)	-0.0379	0.2512	0.2527	52.34
DML-IVQR (λ = Belloni and Chernozhukov)	0.0533	0.1155	0.1042	14.68
$n < p$		$n = 400$		
Oracle-IQR	-0.0160	0.0813	0.0811	1.33
CFQR-HD (λ = 2-fold Cross-Validation)	0.0182	0.0613	0.0595	10.28
CFQR-HD (λ = Belloni and Chernozhukov)	0.5297	0.5310	0.0382	1.08
IVQR-HD (λ = 2-fold Cross-Validation)	0.0455	0.0988	0.0893	836.40
IVQR-HD (λ = Belloni and Chernozhukov)	0.1027	0.1575	0.1215	64.70
DML-IVQR (λ = 2-fold Cross-Validation)	0.0100	0.0796	0.0803	80.26
DML-IVQR (λ = Belloni and Chernozhukov)	0.0300	0.0796	0.0750	16.32

Table 2: Choice of λ, $\tau = 0.5$, $p = 500$ and $n > p$

	$\tau = 0.5$			
	Bias	RMSE	SD	Time(s)
$n > p$		$n = 600$		
Oracle-IQR	0.0160	0.0620	0.0609	1.78
CFQR-HD (λ = 2-fold Cross-Validation)	-0.0056	0.0454	0.0458	13.70
CFQR-HD (λ = Belloni and Chernozhukov)	0.4866	0.4876	0.0327	2.48
IVQR-HD (λ = 2-fold Cross-Validation)	0.0338	0.0830	0.0773	975.6
IVQR-HD (λ = Belloni and Chernozhukov)	0.1160	0.1541	0.1031	86.80
DML-IVQR (λ = 2-fold Cross-Validation)	0.0300	0.0753	0.0702	101.88
DML-IVQR (λ = Belloni and Chernozhukov)	0.0367	0.0753	0.0669	19.58
$n > p$		$n = 1000$		
Oracle-IQR	-0.0027	0.0400	0.0406	3.12
CFQR-HD (λ = 2-fold Cross-Validation)	-0.0262	0.0379	0.0278	17.88
CFQR-HD (λ = Belloni and Chernozhukov)	0.4713	0.4722	0.0284	2.40
IVQR-HD (λ = 2-fold Cross-Validation)	0.0314	0.0693	0.0629	1458
IVQR-HD (λ = Belloni and Chernozhukov)	0.0893	0.1288	0.0944	132
DML-IVQR (λ = 2-fold Cross-Validation)	0.0148	0.0471	0.0456	147.60
DML-IVQR (λ = Belloni and Chernozhukov)	0.0100	0.0408	0.0403	27.32

5 Application

In this section, we reexamine an empirical study on quantile effects: The impact of compulsory schooling on earnings (Angrist and Krueger (1991)). Since primary school in America requires a student to have turned age six by January 1 of the year in which he or she enters school and compulsory schooling laws generally require students to attend school until they are sixteen or seventeen, children born in the first quarter of the year typically have a lower average level of education than children born later in the year. As a result, Angrist and Krueger (1991) use quarters of birth as instruments for education.

Their model is described as follows

$$y_i = \theta d_i + x_i'\gamma + \epsilon_i$$
$$d_i = z_i'\beta + x_i'\delta + \eta_i$$

where y_i is the log(wage) of individual i, d_i denotes education, x_i is a vector of control variables, and z_i is a vector of instrumental variables that affect education but do not directly affect the wage.

The data was collected from 1980 U.S. Censuses consist of $329,509$ men born between 1930 and 1939. They use three quarter-of-birth dummies interacted with nine year-of-birth dummies as instruments z_i in 2SLS estimation. Race dummies, a dummy for residence in an SMSA (center city),

a marital status dummy, and eight region-of-residence dummies are also included as exogenous controls x_i. Table 3 and Table 4 are summary statistics and regression results in Angrist and Krueger's research. In table 5, we present estimates of the same set of models using ordinary quantile regression.

In our high-dimensional setting, x_i is a set of 510 variables: A race dummy, 9 year-of-birth dummies, 50 state-of-birth dummies, and 450 state-of-birth $\times$ year-of-birth interactions. As instruments z_i, we consider three cases:

(1) Three quarter-of-birth dummies;

(2) Three quarter-of-birth dummies and their interactions with 9 main effects for year-of-birth and 50 main effects for state-of-birth, totaling 180 instruments;

(3) Three quarter-of-birth dummies and their interactions with the set of state-of-birth and year-of-birth controls, resulting in 1530 instruments.

Table 6 presents the results based on our method.

6 Conclusion

In this paper, we propose a double selection estimator to estimate a high-dimensional quantile regression model in the presence of endogeneity based on the control function (CF) approach. First, we extend the idea of double selection to quantile analysis. Second, we find that under general conditions, our model is computationally simpler than instrumental variable quantile regression (IVQR). Monte Carlo simulations also show that our estimator performs well under high-dimensional controls. Third, we employ our model to investigate the impact of compulsory schooling on earnings using 1530 instruments for education based on Angrist-Krueger data, and we find that high-dimensional quantile estimates are smaller than 2SLS and OLS estimates.

References

Abadie, A., J. Angrist, and G. Imbens (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica 70*(1), 91–117.

Amemiya, T. (1982). Two stage least absolute deviations estimators. *Econometrica: Journal of the Econometric Society*, 689–711.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Belloni, A. and V. Chernozhukov (2011). $l1$-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*(1), 82–130.

Belloni, A., V. Chernozhukov, and C. Hansen (2011). Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Belloni, A., V. Chernozhukov, and K. Kato (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association 114*(526), 749–758.

Blundell, R. and J. L. Powell (2007). Censored regression quantiles with endogenous regressors. *Journal of Econometrics 141*(1), 65–83.

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics 186*(1), 201–221.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013). Inference on counterfactual distributions. *Econometrica 81*(6), 2205–2268.

Chernozhukov, V. and C. Hansen (2005). An IV model of quantile treatment effects. *Econometrica 73*(1), 245–261.

Chernozhukov, V. and C. Hansen (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics 132*(2), 491–525.

Chernozhukov, V. and C. Hansen (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics 142*(1), 379–398.

Chernozhukov, V., C. Hansen, and K. Wüthrich (2017). Instrumental variable quantile regression. *Handbook of quantile regression*, 119–143.

Hansen, C., J. Hausman, and W. Newey (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics 26*(4), 398–422.

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Koenker, R. (2005). *Quantile regression*, Volume 38. Cambridge university press.

Lee, S. (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics 141*(2), 1131–1158.

Powell, J. L. (1983). The asymptotic normality of two-stage least absolute deviations estimators. *Econometrica: Journal of the Econometric Society*, 1569–1575.

Table 3: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
LWKLYWGE	329,509	5.900	0.679	−2.342	10.532
AGE	329,509	44.645	2.940	40	50
EDUC	329,509	12.770	3.281	0	20
RACE	329,509	0.082	0.274	0	1
SMSA	329,509	0.186	0.389	0	1
MARRIED	329,509	0.863	0.344	0	1
QOB	329,509	2.506	1.112	1	4
POB	329,509	30.693	14.218	1	56
ENOCENT	329,509	0.201	0.401	0	1
ESOCENT	329,509	0.065	0.247	0	1
MIDATL	329,509	0.162	0.368	0	1
MT	329,509	0.049	0.217	0	1
NEWENG	329,509	0.056	0.230	0	1
WNOCENT	329,509	0.078	0.268	0	1
WSOCENT	329,509	0.097	0.296	0	1
SOATL	329,509	0.168	0.374	0	1

Table 4: OSL and 2SLS regression results

	Dependent variable:	
	LWKLYWGE	
	OLS	*2SLS*
	(1)	(2)
EDUC	0.063***	0.081***
	(0.0003)	(0.016)
RACE	−0.257***	−0.230***
	(0.004)	(0.026)
MARRIED	0.248***	0.244***
	(0.003)	(0.005)
SMSA	−0.176***	−0.158***
	(0.003)	(0.017)
9 Year-of-birth dummies	Yes	Yes
8 Region of residence dummies	Yes	Yes
Constant	4.986***	4.744***
	(0.007)	(0.229)
Observations	329,509	329,509
R^2	0.165	0.158
Adjusted R^2	0.165	0.158
Residual Std. Error (df = 329487)	0.620	0.623
F Statistic	3,101.110*** (df = 21; 329487)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 5: Quantile regression results

	Dependent variable:			
	LWKLYWGE			
	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
	(1)	(2)	(3)	(4)
EDUC	0.06155***	0.05967***	0.05955***	0.06626***
	(0.00032)	(0.00027)	(0.00028)	(0.00045)
RACE	−0.24153***	−0.21838***	−0.20779***	−0.20371***
	(0.00425)	(0.00339)	(0.00331)	(0.00482)
MARRIED	0.27655***	0.19754***	0.15765***	0.13113***
	(0.00380)	(0.00298)	(0.00283)	(0.00431)
SMSA	−0.18011***	−0.15779***	−0.14501***	−0.12736***
	(0.00294)	(0.00245)	(0.00249)	(0.00376)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region of residence dummies	Yes	Yes	Yes	Yes
Constant	6.60725***	4.84521***	4.39191***	5.20814***
	(1.27394)	(1.02427)	(1.12665)	(1.79119)
Observations	329,509	329,509	329,509	329,509
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Effects of return to schooling in the Angrist-Krueger data

High-dimensional CFQR regression results				
Number of instruments	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
3	0.003065	0.003432	0.005259	0.003880
180	0.003885	0.004605	0.005035	0.004670
1530	0.002100	0.003604	0.003256	0.001631