Endogenous High-Dimensional Quantile Regression: A Control Function Approach

Kaixi Zhang, HKUST

University of Torino, Italy

June 27, 2025

1. Motivation
2. Estimator
3. Monte Carlo Simulation
4. Application
5. Conclusion

1 Motivation

2 Estimator

3 Monte Carlo Simulation

4 Application

5 Conclusion

A Running Example

Angrist and Krueger (1991): return to schooling

- three quarters of birth as instruments for education

A Running Example

Angrist and Krueger (1991): return to schooling

- three quarters of birth as instruments for education
- 180 instruments corresponding to the three quarter-of-birth dummies and their interactions with the 9 main effects for year-of-birth and 50 main effects for state-of-birth.(Hansen, Hausman and Newey, 2008)

A Running Example

Angrist and Krueger (1991): return to schooling

- three quarters of birth as instruments for education
- 180 instruments corresponding to the three quarter-of-birth dummies and their interactions with the 9 main effects for year-of-birth and 50 main effects for state-of-birth.(Hansen, Hausman and Newey, 2008)
- 1530 potential instruments with 530 exogenous controls (Belloni, Chernozhukov and Hansen, 2012)

Motivation

- Endogeneity problem
- High dimensional sparse regression models (HDSMs)
- Quantile regression (QR)

Review - Solving endogeneity in linear model

- **Instrumental Variable (IV) Regression**

$$Y_i = \alpha D_i + \epsilon_i$$

where $E(D_i \epsilon_i) \neq 0$ and $E(Z_i \epsilon_i) = 0$

$$E(Z_i(Y_i - \alpha D_i)) = 0 \quad \Rightarrow \quad \alpha = E(Z_i D_i)^{-1} E(Z_i Y_i)$$

Review - Solving endogeneity in linear model

- **Instrumental Variable (IV) Regression**

$$Y_i = \alpha D_i + \epsilon_i$$

where $E(D_i\epsilon_i) \neq 0$ and $E(Z_i\epsilon_i) = 0$

$$E(Z_i(Y_i - \alpha D_i)) = 0 \quad \Rightarrow \quad \alpha = E(Z_i D_i)^{-1} E(Z_i Y_i)$$

- **Control Function (CF) approach**

$$D_i = \pi Z_i + \theta X_i + \eta_i$$
$$Y_i = \alpha D_i + \gamma \hat{\eta}_i + \beta X_i + \epsilon_i$$

Review - Solving endogeneity in quantile model

- **Instrumental Variable Quantile Regression** (IVQR, Chernozhukov and Hansen ,2005, 2006)
  - Model
    $$Q_{Y|X}(\tau) = D'\alpha(\tau) + X'\beta(\tau)$$
  - Under suitable assumptions, the identification is given by
    $$P(Y \leq Q_{Y|X}(\tau)|X, Z) = \tau$$
    which implies unconditional moment conditions
    $$E[(\tau - 1\{Y < D'\alpha + X'\beta\})\varphi(X, Z)] = 0$$

Estimation

Proposed estimator

- **Endogenous Quantile Regression based on Control Function approach** (CFQR)
  - Model

$$Q_{D|X,Z}(v|x,z) = Z'\pi + X'\theta + Q_V(v)$$
$$Q_{Y|X,D,V}(u|x,d,v) = \alpha(u)D + \gamma(u)V + X'\beta(u)$$

ℓ_1-penalized quantile regression (Belloni and Chernozhukov, 2011)

Proposed estimator

- **Endogenous Quantile Regression based on Control Function approach** (CFQR)
  - Model

$$Q_{D|X,Z}(v|x,z) = Z'\pi + X'\theta + Q_V(v)$$
$$Q_{Y|X,D,V}(u|x,d,v) = \alpha(u)D + \gamma(u)V + X'\beta(u)$$

  ℓ_1-penalized quantile regression (Belloni and Chernozhukov, 2011)

  - Computationally simpler than IVQR

	Time (Second)
Our estimator	5.1
IVQR	414

Table 1: Comparison: $\tau = 0.5$ and $n = 100$

1 Motivation

2 Estimator

3 Monte Carlo Simulation

4 Application

5 Conclusion

Model

Triangular simultaneous equation model (Imbens and Newey, 2009)

$$Y = g(X, \epsilon) \quad (1)$$
$$D = h(Z, \eta) \quad (2)$$

where $X = (D, Z_1')'$ is a vector of observed variables, D is a endogenous variable and $Z = (Z_1', Z_2')'$ is a vector of exogenous variables. ϵ is a vector of disturbances and η is a scalar reduced-form error.

Theorem

Suppose (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 X and ϵ are independent conditional on $V = F_{D|Z}(D, Z) = F_{\eta}(\eta)$.

Estimation

- **Step 1**: We select the control variables to predict endogenous variable D, then obtain the residual $\hat{\eta}$ and $\hat{V}$.

$$\hat{\kappa} = \arg\min(D - \widetilde{Z}'\kappa)^2 + \lambda_1||\kappa||_1$$

where $\kappa = (\pi', \theta')'$ and $\widetilde{Z} = (Z', X')'$.

Estimation

- **Step 1**: We select the control variables to predict endogenous variable D, then obtain the residual $\hat{\eta}$ and $\hat{V}$.

$$\hat{\kappa} = \arg\min (D - \widetilde{Z}'\kappa)^2 + \lambda_1 ||\kappa||_1$$

where $\kappa = (\pi', \theta')'$ and $\widetilde{Z} = (Z', X')'$.

- **Step 2**: we regress Y on the endogenous variable D, $\hat{V}$ and the control variables X with ℓ_1-penalized quantile regression

$$(\hat{\alpha}, \hat{\gamma}, \hat{\beta}) = \arg\min \rho_\tau (Y - \alpha D - \gamma \hat{V} - X'\beta) + \lambda_2 ||\beta||_1$$

Comparison with IVQR

- Estimation of IVQR (IQR, Chernozhukov and Hansen, 2006)

IVQR

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

at true value of α_0, the ordinary linear QR of $Y - D'\alpha_0$ on X and Z would yield coefficients on the instruments of exactly zero.

Comparison with IVQR

- Estimation of IVQR (IQR, Chernozhukov and Hansen, 2006)
  IVQR

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

  at true value of α_0, the ordinary linear QR of $Y - D'\alpha_0$ on X and Z would yield coefficients on the instruments of exactly zero.
- The advantages of CFQR

Comparison with IVQR

- Estimation of IVQR (IQR, Chernozhukov and Hansen, 2006)
  IVQR

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

  at true value of α_0, the ordinary linear QR of $Y - D'\alpha_0$ on X and Z would yield coefficients on the instruments of exactly zero.

- The advantages of CFQR
  - Computationally simpler
    i.e. running 1000 times vs. running twice

Comparison with IVQR

- Estimation of IVQR (IQR, Chernozhukov and Hansen, 2006)
  IVQR

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

  at true value of α_0, the ordinary linear QR of $Y - D'\alpha_0$ on X and Z would yield coefficients on the instruments of exactly zero.

- The advantages of CFQR
  - Computationally simpler
    i.e. running 1000 times vs. running twice
  - Allow for more endogenous variables
    i.e. adding $D_1, D_2, \ldots$ to the second step

1 Motivation

2 Estimator

3 Monte Carlo Simulation

4 Application

5 Conclusion

Data generation process

We adopt the following data generating processes with dimension p of covariates X is 500 and sample size $n = c(100, 200, 400, 600, 1000)$. Each case replicated 30 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

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

Choice of penalty level λ

- Belloni and Chernozhukov (2011)

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

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

$$\Lambda = n \sup_{u \in \mathcal{U}} \max_{1 \le j \le p} |\frac{1}{n} \sum_{i=1}^{n} [\frac{x_{ij}(u - 1\{u_i \le u\})}{\hat{\sigma}_j \sqrt{u(1-u)}}]|$$

where $\hat{\sigma}_j{}^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2$, $u_1, ..., u_n$ are $i.i.d.$ uniform $(0,1)$ random variables that are independently distributed from the controls $x_1, .., x_n$.

- K-fold cross validation

Simulation Results

	$\tau = 0.5$			
	Bias	RMSE	SD	Time(s)
$n < p$		$n = 100$		
CFQR-HD (λ = 2-fold Cross-Validation)	0.4878	0.4946	0.0831	5.1
CFQR-HD (λ = Belloni and Chernozhukov)	0.3414	0.3621	0.1227	0.515
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
$n < p$		$n = 200$		
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	514.8
IVQR-HD (λ = Belloni and Chernozhukov)	0.2827	0.3232	0.1593	43.7
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
$n < p$		$n = 400$		
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	836.4
IVQR-HD (λ = Belloni and Chernozhukov)	0.1027	0.1575	0.1215	64.7
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: Comparison 1: Choice of λ, $\tau = 0.5$, $p = 500$ and $n < p$

Simulation Results

	$\tau = 0.5$			
	Bias	RMSE	SD	Time(s)
$n > p$		$n = 600$		
CFQR-HD (λ = 2-fold Cross-Validation)	-0.0056	0.0454	0.0458	13.7
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	86.8
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
$n > p$		$n = 1000$		
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	2.4
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	147.6
DML-IVQR (λ = Belloni and Chernozhukov)	0.0100	0.0408	0.0403	27.32

Table 3: Comparison 2: Choice of λ, $\tau = 0.5$, $p = 500$ and $n > p$

1 Motivation

2 Estimator

3 Monte Carlo Simulation

4 Application

5 Conclusion

Angrist and Krueger (1991) collected data from the 1980 U.S. Censuses consist of 329,509 men born between 1930 and 1939.

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: Summary statistics

OSL and 2SLS regression results

	Dependent variable:	
	LWKLYWGE	
	OLS	*2SLS*
	(1)	(2)
EDUC	0.063***	0.081***
	(0.0003)	(0.016)
RACE (1 = Black)	−0.257***	−0.230***
	(0.004)	(0.026)
MARRIED (1 = married)	0.248***	0.244***
	(0.003)	(0.005)
SMSA (1 = center city)	−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
Note:	*p<0.1; **p<0.05; ***p<0.01	

- The quantile regression model is described as follows

$$D_i = Z_i'\pi_0 + X_i'\theta_0$$
$$Q_{Y|X,D}(u|x,d) = \alpha_0(u)D_i + X_i'\beta_0(u)$$

- 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.

In high-dimensional model 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

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

Effects of return to schooling in the Angrist-Krueger data

	Dependent variable:			
	LWKLYWGE			
	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
Number of instruments	(1)	(2)	(3)	(4)
Ordinary QR	0.06155	0.05967	0.05955	0.06626
	(0.00032)	(0.00027)	(0.00028)	(0.00045)
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
Observations	329,509	329,509	329,509	329,509

1 Motivation

2 Estimator

3 Monte Carlo Simulation

4 Application

5 Conclusion

Summary

- We extend the idea of double selection to quantile analysis
- Our model is computationally simpler than instrumental variable quantile regression (IVQR)
- We employ our model to investigate the impact of compulsory schooling on earnings using 1530 instruments for education based on Angrist-Krueger data

Thanks!