SSIVQREG: Quantile selection models with endogenous regressors in Stata

Paul Bingley, **Christophe Kolodziejczyk**, Nicolai Kristensen

VIVE, The Danish Center for Social Science Research

August 29, 2025

Contact: `ckol@vive.dk`

THE DANISH CENTER FOR SOCIAL SCIENCE RESEARCH

Outline

- Introduction
- Model
- Estimation
- SSIVQREG
- Applications
- Conclusion

Motivation

- Old question in economics: estimate the returns to education
- Several issues
  - Sample selection: decision to work is a choice ; earnings/wages are only observed for participants.
  - Endogeneity: education is a choice correlated with unobserved factors driving the individual wages
  - Analysis of the wage distribution

Background

- Conditional mean: Heckman model and its variations
- Distribution: Arellano and Bonhomme (2017) (AB) propose estimation methods for quantile selection models.
- AB consider the case where covariates are exogenous and suggest one method for the endogenous case.
- AB's estimator for the case with **exogenous** covariates has been implemented in Stata with two commands
  - **arhomme** (Biewen and Erhardt, 2021)
  - **qregsel** (Muñoz and Siravegna, 2021)
- No commands for the endogeneity case

SSIVQREG

We introduce a new Stata command: **SSIVQREG**.

- Implement estimators for the case of endogenous covariates.
- New estimator based on the smoothing of the estimating equations of the model. Reduce computation time
- This estimator can be applied to the case with exogenous covariates.
- Computes analytical standard errors derived in AB's article.
- and more

THE DANISH CENTER FOR SOCIAL SCIENCE RESEARCH

Roadmap

- Explain the quantile selection model
- Present SSIVQREG
- Show a Monte Carlo simulation exercise
- Show an application with real data.

Quantile Selection Model with endogenous regressors

Consider the following model.

$$\begin{aligned} Y^* &= q(U, E, X), && (1) \\ D &= \mathbf{1}\{V \leqslant p(Z)\}, && (2) \\ Y &= Y^* \text{ if } D = 1, && (3) \end{aligned}$$

Y^* has a linear quantile form for a given rank τ:

$$q(\tau, E, X) = E\alpha_\tau + X\beta_\tau$$

Quantile Selection Model with endogenous regressors*

- Y is the outcome (wage)
- Y^* is the latent outcome and is only observed if $D = 1$
- E (education) is a potentially endogenous variable
- X is a set of exogenous variables and U the unobserved ability.
- V is the unobserved resistance to participate
- $p(Z)$ is the propensity to participate given observed Z

Selection

- Individuals participate if their propensity (given Z) exceeds V.
- U and V are potentially correlated. Those with a high ability are for example more likely to have a lower resistance to participate
- Modeled with a bivariate copula $C(U, V; \rho)$. ρ is the copula dependency parameter
- Parametric copula: Frank or Gaussian

Moment condition: rotated quantile

AB (2017) have the following identification result

$$
\begin{aligned}
P\left[Y^* \leqslant q(\tau, E, X) \mid D=1, Z\right] &= \frac{C_x(\tau, p(z))}{p(z)} \\
&= G_x(\tau, p(z))
\end{aligned}
$$

Note: if U and V are independent, then $G_x(\tau, p(z)) = \tau$ and we have the conditional moment for quantile regression. Appendix

IDENTIFICATION

In order to identify the model we need exclusion restrictions.

- (At least) one instrument for E
- One instrument for the participation decision D

Estimation I: AB's estimator/profiled GMM

AB's estimator consists of three steps

1. Estimate the propensity score
2. Choose the value of ρ which minimizes the objective function.
   - 2.1 For a fixed value of the dependency parameter estimate IV quantile regressions for a predefined grid of probabilities (e.g. 0.1, 0.2, ..., 0.9).
   - 2.2 and compute the moment condition for the dependency parameter.
   - 2.3 Probabilities are corrected for sample selection.
   - 2.4 The IVQR (Chernozhukov and Hansen, 2008) involves a grid search for the parameter of the endogenous variable.
3. (Optional) Estimate more quantiles with the estimated value of the dependency parameter.

Estimation II: Smoothing

- Applied by Kaplan and Sun (2017) as an alternative to the IVQR.
- Original problem is nonconvex
  - Use a smoothed version of the moment conditions instead of the original moments.
  - Use the GMM to estimate the parameters of this model
- Requires specifying the smoothing parameter or bandwidth.
- Optimize the objective function with the Gauss-Newton algorithm (with optimize or moptimize)

SSIVQREG's features

- Estimates quantile models with or without **endogenous** regressors.
- Two main estimators: profiled or smoothed GMM
- Computes analytical asymptotic standard errors and allows bootstrapping.
- Additional features: preprocessing (Pereda-Fernández, 2025), one-step estimator, simulated annealing, AMCMC (Baker, 2014).

Syntax

Exogenous case:

ssivqreg *depvar* [*indepvars*] [*if*] [*in*] [*weight*] , select(*depvar* [=] [*indepvars*]) [quantile(#) nrho(#) copula(*string*) gmm rescale]

Endogenous case:

ssivqreg *depvar* (*varname*= *varlist*) [*indepvars*] [*if*] [*in*] [*weight*] , select(*depvar* [=] [*indepvars*]) [quantile(#) nrho(#) nalpha(#) copula(*string*) gmm rescale amcmc]

Monte Carlo simulation exercise

- We investigate the bias and the computation time of our estimators
- focus on the dependency parameter and a heterogeneous treatment effect.
- We run simulations for different values of the dependency parameter.
- 10,000 observations. $\sim$ 50 % of selected observations
- 500 replications (100 for the profiled GMM in the endogenous case)

THE DANISH CENTER FOR SOCIAL SCIENCE RESEARCH

Data generation

- Two Data generating processes (DGP): Exogenous/endogenous treatment.
- Heterogenous treatment effect E: Uniformly distributed between 0 and 1, Median effect is 0.5.
- Just-identified case: one instrument for the participation equation and one for the treatment effect.
- Sample selection: Frank and Gaussian copulas.

Results

- Estimators consistent for the dependency parameter Dependency
- Treatment effect:
  - exogenous cases: consistent TE exogenous
  - endogenous case: higher bias for high dependency and low/higher quantiles TE endogenous
- Computation time is considerably reduced when smoothing CPU time

Married women labor supply (Mroz, 1987)

- Data on married women labor supply in the US (Mroz, 1987)
- Small dataset used in Wooldrige's textbook (2010) to illustrate Heckman's model (753 obs.)
- Data on wage, education, husband's income, non-labor income and number of children
- Instruments:
  - Wage equation: Use parental education and the husband's education to instrument education
  - Participation equation: Non-labor income, parental education and the husband's education

Mroz data - Estimates returns to education

- Estimates returns to education from these data.
- Compare the different models available (QR, IVQR, SSQR, SSIVQR)
- Although the point-estimates of the sample selection models tend to be lower, we cannot reject the absence of sample selection.
- Point-estimates to returns to education tend to be lower when instrumenting education, but confidence intervals are larger
- Bear in mind that the sample is quite small.

Mroz table

Summary

1. Stata command SSIVQREG
   - Allows to estimate quantile selection models with or without endogenous covariates.
   - Three estimation methods
   - Analytical standard-errors and bootstrap
2. Monte Carlo study
   - Estimators seem to perform well except in the endogenous case when the correlation between unobserved variables is high, the bias for the treatment effect is rather high.
   - Smoothing is much faster.
3. I have illustrated the use of SSIVQREG with an application with real data.

Thank you!

References I

Arellano, M., and S. Bonhomme (2017): "Quantile Selection Models With an Application to Understanding Changes in Wage Inequality," *Econometrica*, 85(1), 1–28.

Baker, M. J. (2014): "Adaptive Markov Chain Monte Carlo Sampling and Estimation in Mata," *The Stata Journal*, 14(3), 623–661.

Biewen, M., and P. Erhardt (2021): "arhomme: An implementation of the Arellano and Bonhomme (2017) estimator for quantile regression with selection correction," *The Stata Journal*, 21(3), 602–625.

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

Kaplan, D. M., and Y. Sun (2017): "SMOOTHED ESTIMATING EQUATIONS FOR INSTRUMENTAL VARIABLES QUANTILE REGRESSION," *Econometric Theory*, 33(1), 105–157.

References II

Mroz, T. A. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55(4), 765–799.

Muñoz, E., and M. Siravegna (2021): "Implementing quantile selection models in Stata," *The Stata Journal*, 21(4), 952–971.

Pereda-Fernández, S. (2025): "Fast Algorithms for Quantile Regression with Selection," *Journal of Econometric Methods*, 14(1), 35–47.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, vol. 1 of *MIT Press Books*. The MIT Press.

Appendix: moment conditions

IV quantile regression and sample selection (L quantiles):

$$
\begin{aligned}
E\left[DZ'\left(1\{y \leq X\beta_{\tau_l}\} - G\left(\tau_l, ps; \rho\right)\right)\right] &= 0 \\
\sum_{l=1}^{L} E\left[D\varphi\left(\tau_l, Z\right)\left(1\{y \leq X\beta_{\tau_l}\} - G\left(\tau_l, ps; \rho\right)\right)\right] &= 0
\end{aligned}
$$

Back

Appendix: smoothed moments

$$
\begin{aligned}
E\left[DZ'\left(I\left(-\frac{y-\beta_{\tau_l}}{h_{\tau_l}}\right)-G\left(\tau_l, ps;\rho\right)\right)\right] &= 0 \\
\sum_{l=1}^{L} E\left[D\varphi\left(\tau_l, Z\right)\left(I\left(-\frac{y-\beta_{\tau_l}}{h_{\tau_l}}\right)-G\left(\tau_l, ps;\rho\right)\right)\right] &= 0
\end{aligned}
$$

$I\left(\cdot\right)$ is a smoothing function and h_{τ_l} is a bandwidth.

THE DANISH CENTER FOR SOCIAL SCIENCE RESEARCH

DGP I: No endogenous covariates

- d is a binary exogenous treatment
- u and v are correlated. Gaussian or Frank copula
- z_p is the instrument for the participation decision

$$\begin{aligned} y &= \alpha d + x_1 + x_2 + u \\ p &= 1\left(0.5x_1 + 0.5x_2 - 0.5z + 0.5z_p + v > 0\right) \end{aligned}$$

DGP II: 1 endogenous covariates

- d is endogenous since it is correlated with u
- z is the instrument for d

$$
\begin{aligned}
y &= \alpha d + x_1 + x_2 + u \\
p &= 1\left(0.5x_1 + 0.5x_2 - 0.5z + 0.5z_p + v > 0\right) \\
d &= 1(0.5x_2 + 0.5z + \epsilon > 0) \\
\epsilon &= 0.5u + 0.25w
\end{aligned}
$$

Bias dependency parameter

ρ	Exogenous Smoothed	Exogenous Profiled	Endogenous Smoothed	Endogenous Profiled
Gaussian				
-0.8	0.004	0.000	0.007	0.002
-0.5	0.001	-0.001	0.007	0.004
0.0	0.001	-0.001	0.005	-0.001
0.5	-0.004	-0.001	-0.005	-0.008
0.8	-0.005	-0.003	-0.007	-0.002
Replications	500	500	500	100

Back to results

Bias Treatment effect - Exogenous

	Smoothed			Profiled		
ρ	0.1	.5	0.9	0.1	.5	0.9
Gaussian						
-0.8	0.003	-0.002	-0.010	0.000	-0.001	-0.005
-0.5	0.002	-0.002	-0.007	-0.001	-0.002	-0.003
0.0	0.003	-0.001	-0.002	0.002	-0.000	-0.002
0.5	0.004	0.001	-0.002	0.001	0.000	0.001
0.8	0.003	0.002	-0.001	-0.001	0.001	-0.001
R	500	500	500	500	500	500

Back to results

Bias Treatment effect - Endogenous

	Smoothed			Profiled		
ρ	0.1	.5	0.9	0.1	.5	0.9
Gaussian						
-0.8	0.004	0.002	-0.153	0.008	0.004	-0.173
-0.5	0.002	0.001	-0.009	0.002	0.002	-0.004
0.0	0.006	0.002	-0.002	0.003	0.000	0.011
0.5	0.014	0.000	-0.006	0.027	0.005	-0.001
0.8	0.081	-0.001	0.001	0.054	0.011	0.005
R	500	500	500	100	100	100

Back to results

Computation time (in seconds)

	Exogenous			Endogenous		
	Smoothed	Profiled		Smoothed	Profiled	
ρ	mean	mean	ratio	mean	mean	ratio
-0.8	2	93	44	5	5592	1040
-0.5	2	66	36	2	3521	2058
0.0	2	98	59	2	5027	3284
0.5	2	65	35	2	3408	2099
0.8	2	95	45	3	3497	1398
Replications	500	500		500	100	

Back to results

Mroz data - Results

	QR	SSIVQREG exogenous	GMM	IVQR	SSIVQREG endogenous	GMM
τ_{20}	0.103***	0.0996***	0.0997***	0.0856*	0.0914*	0.0783*
	(4.32)	(3.69)	(3.62)	(2.47)	(2.01)	(2.33)
τ_{50}	0.116***	0.114***	0.109***	0.115***	0.111***	0.104***
	(6.78)	(7.28)	(7.01)	(4.75)	(4.74)	(4.91)
τ_{80}	0.118***	0.115***	0.116***	0.120***	0.127***	0.112***
	(8.83)	(7.90)	(8.63)	(5.64)	(6.60)	(5.27)
ρ		0.0695	0.118		-0.0695	0.148
		(0.34)	(0.63)		(-0.33)	(0.76)
N	428	753	753	428	753	753

t statistics in parentheses, # points $\alpha = 200$, # points $\rho = 100$
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Back to results

THE DANISH CENTER FOR SOCIAL SCIENCE RESEARCH