Implementing Quantile Selection Models in Stata

Mariel Siravegna
Georgetown University

Ercio Munoz
The Graduate Center, CUNY

June 7, 2021

Non-random sample selection is a major issue in empirical work

- A simple sample selection model can be written as the latent model

$$Y^* = X'\beta + \mu$$

but Y^* is only observed if S=1

$$S = \mathbf{1}(Z'\gamma + \nu \geq 0)$$

- Since the seminal work of Heckman (1979), much progress has been made in methods that extend the original model or relax some of its assumptions

- Recently Arellano and Bonhomme (2017) proposed a copula-based method to correct for **sample selection in quantile regression**

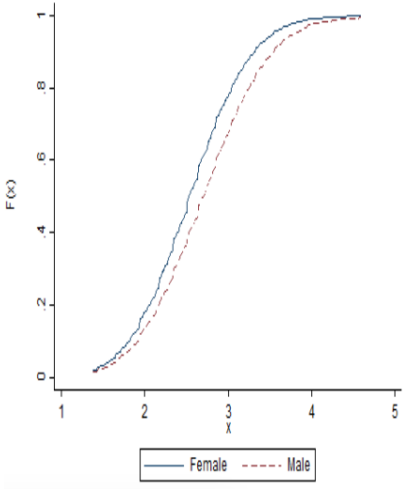
Two Recent Applications

Maasoumi and Wang (JPE 2019)

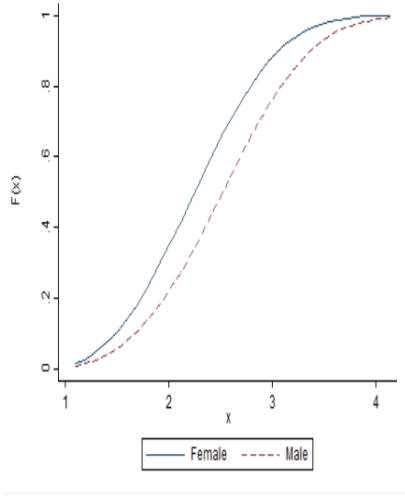
- In this paper the authors use the CPS between 1976-2013 to see how the gender wage gap vary across the wage distribution

- They assess how selective participation of individuals in the labor market affects the gender gap

Comparison of Female and Male Wage CDF



(Without correction)



(Correcting for Selection)

Bollinger et al. (JPE 2019)

- Survey earnings nonresponse is not random
- In this paper the authors match the survey earnings responses to administrative records to see how response vary across the earnings distribution
- They find that non-response rate follows an U shape across earnings and this produces an underestimation of inequality, which can be corrected using this copula-based approach with survey method as excluded instrument

Bollinger et al. (JPE 2019)

TABLE 7
PERFORMANCE OF SELECTION CORRECTION METHODS FOR NONRESPONSE IN THE ASEC ON INEQUALITY

	Inequality Measures			
Sample	Gini	90-10	90-50	50-10
ASEC	.461	10.099	2.607	3.870
ASEC, only respondents with IPW	.464	10.227	2.641	3.869
ASEC, only respondents with copula	.482	10.521	2.676	3.929
ASEC for respondents, DER for nonrespondents (benchmark)	.477	11.038	2.683	4.112

SOURCES.—US Census Bureau Current Population Survey, 2006–11 Annual Social and Economic Supplement; Social Security Administration Detailed Earnings Record, 2005–10.

Estimation

Three-step Algorithm of Arellano and Bohnomme (2017)

Given an i.i.d sample (Y_i, Z_i, S_i), $i = 1, ..., N$ where $Z_i = (X_i, W_i)$ and assuming that quantile functions are linear:

$$q(\tau, x) = x'\beta_\tau, \quad \textit{for all } \tau \in (0, 1) \textit{ and } x \in X \tag{3}$$

the algorithm is as follows:

1. Estimation of the propensity score $p(z)$

2. Estimation of the dependence parameter or degree of selection ρ using this moment restriction:

$$\mathbb{E}[I(Y \leq X'\hat{\beta}_\tau) - G(\tau, p(z); \rho)|S = 1, Z = z] = 0$$

Second Step

Taken to the sample by choosing a ρ that minimizes the following objective function:

$$\hat{\rho} = argmin_{\rho} \| \sum_{i=1}^{N} \sum_{l=1}^{L} S_i \varphi_{\tau_l}(z_i) [\boldsymbol{I}\{Y_i \leq X_i' \tilde{\beta}_{\tau_l}(\rho)\} - G(\tau_l, p(z_i'); \rho)] \|$$

where $\|.\|$ is the Euclidean norm, $\tau_1 < \tau_2 < \cdots < \tau_L$ is a finite grid on $(0, 1)$, and the instrument functions are defined as $\varphi_{\tau_l}(z_i)$, $G(\tau_l, p(z_i'); \rho)$ is the conditional copula indexed by a parameter ρ, and:

$$\tilde{\beta}_{\tau}(\rho) = argmin_{\beta} \sum_{i=1}^{N} S_i [G_{\tau_i}(Y_i - X_i'\beta)^+ + (1 - G_{\tau,i}(Y_i - X_i'\beta)^-]$$

where $a^+ = max\{a, 0\}$, $a^- = max\{-a, 0\}$, and $G_{\tau,i} = G(\tau, p(z); \rho)$.

Third Step

3. Given the estimated $\hat{\rho}$, $\hat{\beta}_\tau$ can be estimated by minimizing a rotated check function of the form:

$$\hat{\beta_\tau} = argmin_\beta \sum_{i=1}^{N} S_i[\hat{G_{\tau,i}}(Y_i - X_i'\beta)^+ + (1 - \hat{G_{\tau,i}})(Y_i - X_i'\beta)^-]$$

where $\hat{\beta}_\tau$ will be a consistent estimator of the τ-th quantile regression coefficient.

Note that this step is unnecessary if the researcher is interested on the quantiles included in the finite grid of step 2.

Implementing the method in Stata

Syntax

`qregsel` *depvar* [*indepvars*] [*if*] [*in*] , *select*([$depvar_S =$] $varlist_S$)
`quantile(#)` [`copula(`*copula*`)` `noconstant` `finergrid` `coarsergrid` `rescale` `nodots`]

Postestimation

`predict newvarlist` [*if*] [*in*]

1. A counterfactual outcome variable

2. Binary indicator of selection

Empirical Example

Wages of women used in Heckman command

```
. global wage_eqn wage educ age
. global seleqn married children educ age
. qregsel $wage_eqn, select($seleqn) quantile(.1 .5 .9)
Grid for the copula parameter (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................
..................................................

Quantile selection model                        Number of obs      =      2000
                                                Selected           =      1343
                                                Nonselected        =       657
Copula parameter (gaussian):    -0.65
```

wage	Coef.
q10	
education	1.112866
age	.204362
_cons	-8.498507
q50	
education	1.017025
age	.2028979
_cons	.5828089
q90	
education	.8888879
age	.2272004
_cons	8.914994

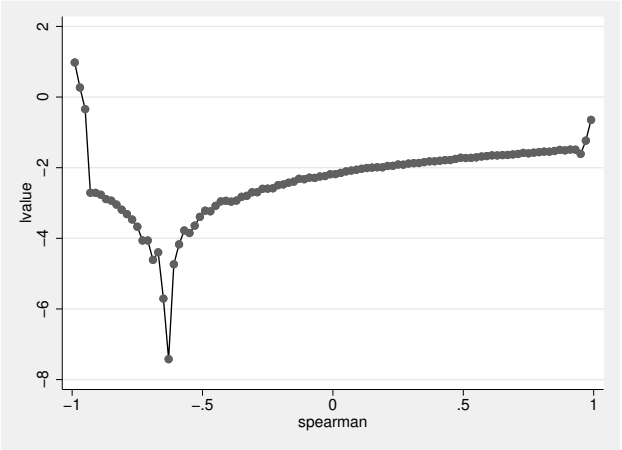
Inference

```
. bootstrap rho=e(rho) _b, reps(100) seed(2) notable: qregsel $wage_eqn, ///
>         select($seleqn) quantile(.1 .5 .9)
(running qregsel on estimation sample)

Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

Bootstrap results                               Number of obs     =      2,000
                                                Replications      =        100

      command:  qregsel wage educ age, select(married children educ age) quantile(.1 .5 .9)
    [_eq4]rho:  e(rho)

. estat bootstrap, percentile

Bootstrap results                               Number of obs     =      2,000
                                                Replications      =        100

      command:  qregsel wage educ age, select(married children educ age) quantile(.1 .5 .9)
    [_eq4]rho:  e(rho)
```

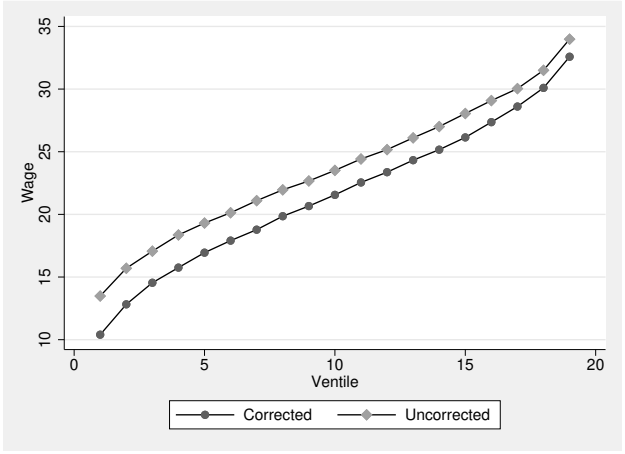
	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
q10						
education	1.1128663	-.0369692	.14707968	.7483546	1.322367	(P)
age	.20436202	-.0065281	.04903284	.0912168	.2998732	(P)
_cons	-8.4985072	.7444134	2.4852059	-11.27083	-2.926636	(P)
q50						
education	1.0170248	.009136	.07041415	.9073696	1.155043	(P)
age	.20289786	.0008091	.02794803	.1479627	.2588321	(P)
_cons	.58280893	-.1804622	1.3881311	-1.880296	2.965075	(P)
q90						
education	.88888792	.015074	.06247303	.7735702	1.034392	(P)
age	.22720039	-.0033785	.02609233	.1670902	.2715747	(P)
_cons	8.9149942	-.1022546	1.1223106	6.964433	10.89201	(P)
_eq4						
rho	-.64783484	-.0216367	.07354153	-.8230287	-.5277461	(P)

(P) percentile confidence interval

Grid for minimization



Counterfactual distribution: Corrected versus uncorrected quantiles



Conclusions

Conclusions

- We have introduced a new Stata command that implements a copula-based method to correct for sample selection in quantile regressions proposed in Arellano and Bonhomme (2017)

- This command may be useful for Stata users doing empirical work, as we have illustrated with the case of two recently published papers

- The code is available in our github or within Stata *(ssc install qregsel)*

- Questions, comments, and suggestions are welcome

References

- Arellano, M., and S. Bonhomme (2017), “Quantile Selection Models with an Application to Understanding Changes in Wage Inequality.” Econometrica 85(1)

- Bollinger, C., B. Hirsch, C. Hokayem, and J. Ziliak (2019), “Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch.” Journal of Political Economy 127(5).

- Maasoumi, E., and L. Wang (2019), “The Gender Gap between Earnings Distributions.” Journal of Political Economy 127(5).

- Munoz, E., and M. Siravegna (2021), “Implementing Quantile Selection Models in Stata.”