Implementing Quantile Selection Models in Stata

Mariel Siravegna
Georgetown University

Ercio Munoz
The Graduate Center, CUNY

July 30, 2020

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

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

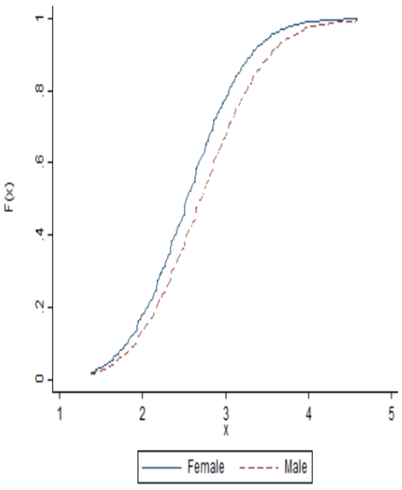
Two Recent Applications

Maasoumi and Wang (JPE 2019)

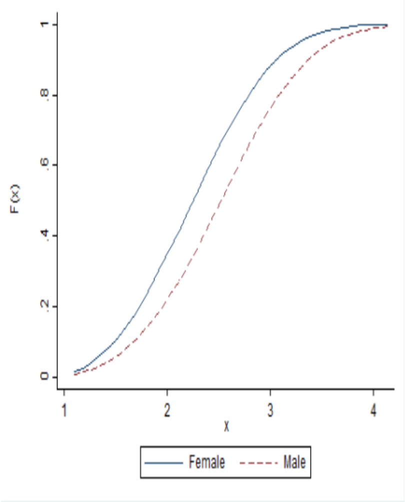
- In this paper the authors use the CPS between 1976-2013 to see how the gender wage gap vary across the wage distribution

- 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 response 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

Bollinger et al. (JPE 2019)

TABLE 7
Performance of Selection Correction Methods for Nonresponse in the ASEC on Inequality

Sample	Inequality Measures			
	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

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:

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

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(`*#*`) grid_min(grid_minvalue) grid_max(grid_maxvalue)`
`grid_length(grid_lengthvalue)` [`copula(`*copula*`) noconstant plot`]

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) copula(gaussian) ///
>  grid_min(-.9) grid_max(.9) grid_length(.05) plot

Quantile selection model                         Number of obs     =     1343

             |       q10        q50        q90
-------------+---------------------------------
   education |  1.083723   1.017025   .8888879
         age |   .204362   .2028979   .2272004
       _cons | -8.148793   .5828089   8.914994
```

Grid for minimization

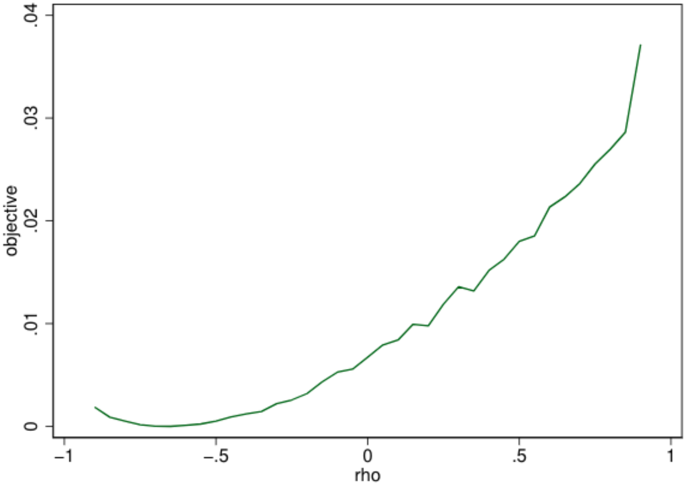


Figure 2: Grid for minimization

Counterfactual distribution: Corrected versus uncorrected quantiles

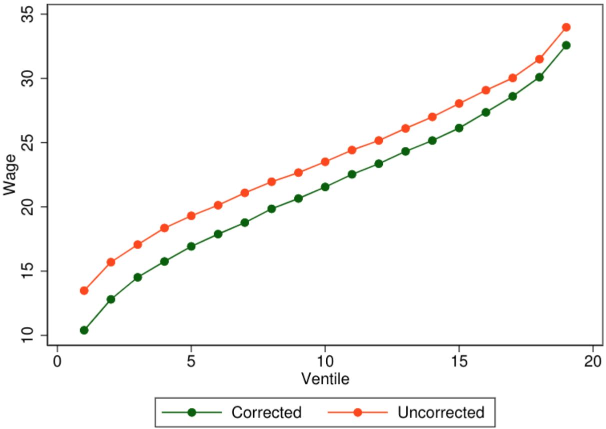


Figure 3: Corrected versus uncorrected quantiles

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 for now only available in our github repo
- 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 (2020), "Implementing Quantile Selection Models in Stata."