Implementing Quantile Selection Models in Stata*

Ercio Muñoz† Mariel Siravegna‡

July 26, 2020

Abstract

This article describes **qregsel**, a Stata module to implement a copula-based sample selection correction for quantile regression recently proposed by Arellano and Bonhomme (2017, Econometrica 85(1): 1-28). We illustrate the use of **qregsel** with an empirical example using the data employed in the Stata base reference manual for the `heckman` command.

Keywords: *sample selection, quantile regression, copula method, python*

*We thank Jim Albrecht and Wim Vijverberg for useful comments and suggestions.
†CUNY Graduate Center; email: emunozsaavedra@gc.cuny.edu.
‡Georgetown University; email: mcs92@georgetown.edu.

1 Introduction

Non-random sample selection is a well known issue in empirical economics. Since the seminal work of Heckman (1979) addressing this problem, much progress has been made in methods that extend the original model or relax some of its assumptions. For example, Vella (1998) provides a survey of methods for estimating models with sample selection bias in this line.

Although most of the effort has been focused on models that estimate the conditional mean, the literature in econometrics has also tackled the problem of non-random sample selection in the context of quantile regression. For example, Arellano and Bonhomme (2017a) offer a survey of recently proposed methods with a focus on a copula-based sample selection model suggested in Arellano and Bonhomme (2017b).

As discussed in Arellano and Bonhomme (2017a), the flexible copula-based approach has an advantage over methodologies that are based on the control function approach. The latter impose conditions on the data that may not be compatible with quantile models if the model is non-additive with non-linear quantile curves on the selected sample (see Huber and Melly (2015)).

In this paper, we briefly discuss the copula-based approach proposed by Arellano and Bonhomme (2017b) and present a new Stata module called `qregsel` that implements it.[1] In addition, we illustrate the method with an empirical exercise in which we estimate a quantile regression model with sample selection using the Stata base reference manual example for the `heckman` command.

The paper is organized as follows. Section 2 describes the methodology. Section 3 describes the `qregsel` command and its syntax. In section 4 we illustrate the use of the command with the empirical example, and we conclude in Section 5.

2 Methodology

In this section we briefly review the quantile selection model of Arellano and Bonhomme (2017b). The goal is to obtain a consistent estimator when there is sample selection in a non-additive model, which precludes the use of the control function approach. The assumption of additivity does not hold in general, as argued by Huber and Melly (2015) in the context of testing.

2.1 The Model

Sample selection is modeled using a bivariate cumulative distribution function or copula of the percentile error in the latent outcome equation and the error in the sample selection equation. The copula parameters are estimated by minimizing a method-of-moments criterion that exploits variation in excluded regressors to achieve credible

1. A copula-based maximum-likelihood method for the conditional mean is already available in Stata (see Hasebe (2013)).

identification. Then the quantile regression parameters are obtained by minimizing a rotated check function, which preserves the linear programming structure of the standard linear quantile regression (see Koenker and Bassett (1978)).

Consider a general outcome equation specification where the quantile functions are linear:

$$Y^* = Q(U, X) = x'\beta(\tau) \tag{1}$$

where Y^* is the the latent outcome variable (e.g. wage offers), the function Q is the τ-th conditional quantile of Y^* given the covariates X (e.g. education, experience, etc.), and U is the error term of the outcome equation.

The participation equation is defined as:

$$D = I\{V \leq p(Z)\} \tag{2}$$

where D takes values equal to 1 when the latent variable is observable (e.g. employment) and 0 otherwise, Z contains X and at least one covariate B that do not appear in the outcome equation (e.g., a determinant of employment that does not affect wages directly), *p(Z)* is a propensity score, and V is an error term of the selection equation.

Define G_x as the conditional copula function, which measures the dependence between U and V as:

$$G_x(\tau, p) \equiv G(\tau, p; \rho) = \frac{C(\tau, p; \rho)}{p} \tag{3}$$

where the numerator is the unconditional copula of *(U,V)*, the denominator is the propensity score, and ρ is the copula parameter that governs the dependence between the error in the outcome equation and the error in the participation decision.[2] G_x maps rank τ in the distribution of latent outcomes (given X=x) to ranks $G_x(\tau, p(Z))$ in the distribution of observed outcomes conditional on participation (given Z=z). Namely, the conditional $G_x(\tau, p(z))$-quantile of observed outcome (that is when $D = 1$) coincides with the conditional τ-quantile of latent outcome, and this is true for each $\tau \in (0, 1)$. The key implication from this equation is that, if we are able to estimate the mapping $G_x(\tau, p)$ from latent to observed ranks, we are able to estimate the quantiles of observed outcomes corrected for selection.

2.2 Estimation

Given all the above, Arellano and Bonhomme (2017a)'s estimation algorithm can be summarized in 3 steps: estimation of the propensity score, estimation of the degree of selection via the cumulative distribution function of the percentile error in the outcome equation and the error in the participation decision, and then, using the estimated parameter, the computation of quantile estimates through rotated quantile regression.

2. We focus only on the implementation of models with single-parameter copulas.

The first step consists of estimating the propensity score γ by a probit regression:

$$\hat{\gamma} = argmax_a \sum_{i=1}^{N} D_i ln\Phi(Z_i'a) + (1 - D_i) ln\Phi(-Z_i'a) \tag{4}$$

The second step is to estimate ρ by generalized method of moments, which allow us to obtain an observation-specific measure of dependence between the rank error in the equation of interest and the rank error in the selection equation. This step consists of working with a parametric copula and deriving moment restrictions on the copula parameter:

$$\hat{\rho} = argmin_c \| \sum_{i=1}^{N} \sum_{l=1}^{L} D_i \varphi(\tau_l, Z_i) [\mathbf{1}\{Y_i \leq X_i' \hat{\beta}_{\tau_l}(c)\} - G(\tau_l, p(Z_i'; \hat{\gamma}), c)] \| \tag{5}$$

where $\|.\|$ is the Euclidean norm, $\tau_1 < \tau_2 < \ldots < \tau_L$ is a finite grid on $(0, 1)$, and the instrument functions are defined as $\varphi(\tau, Z_i)$ where the dim $\varphi \leq$ dim ρ and:

$$\hat{\beta}_\tau(c) = argmin_{b(\tau)} \sum_{i=1}^{N} D_i [G(\tau, p(Z_i'\hat{\gamma}); c)(Y_i - X_i'b)^+ + \tag{6}$$

$$(1 - G(\tau, p(Z_i'\hat{\gamma}; c))(Y_i - X_i'b)^-] \tag{7}$$

where $a^+ = max\{a, 0\}$, $a^- = max\{-a, 0\}$, and the grid of τ values on the unit interval as well as the instrument function are chosen by the researcher.

Lastly, using $\hat{\gamma}$ and $\hat{\rho}$ obtained above, the third step consists in computing $\hat{G}_{\tau i} = G(\tau, p(Z_i'\hat{\gamma}); \hat{\rho})$ for all i to estimate $\beta(\tau)$ by minimizing a rotated check function of the form:

$$\hat{\beta}(\tau) = argmin_{b(\tau)} \sum_{i=1}^{N} D_i [\hat{G}_{\tau i}(Y_i - X_i'b(\tau))^+ + (1 - \hat{G}_{\tau i})(Y_i - X_i'b(\tau))^-] \tag{8}$$

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

2.3 Copulas

The Arellano and Bonhomme (2017a) analysis covers the case where the copula is left unrestricted but for the implementation they focus on the case of identification where the copula depends on a low-dimensional vector of parameters.

In our empirical implementation, we only consider the case of a reduced set of one-dimensional copulas. We include the Gaussian, Frank, Farlie-Gumbel-Morgenstern (FGM), and Ali-Mikhail-Haq (AMH). Table 1 provides their respective functional forms.

Table 1: Copula functions

Copula name	$C(U,V;\rho)$	Range of ρ
Gaussian	$\Phi_2\{\Phi^{-1}(U),\Phi^{-1}(V);\rho\}$	$-1 \leq \rho \leq 1$
Frank	$-\rho^{-1}log\{1+\frac{(e^{-\rho U}-1)(e^{-\rho V}-1)}{(e^{-\rho}-1)}\}$	$-\infty \leq \rho \leq \infty$
FGM	$UV\{1+\rho(1-U)(1-V)\}$	$-1 \leq \rho \leq 1$
AHM	$UV\{1-\rho(1-U)(1-V)\}^{-1}$	$-1 \leq \rho \leq 1$

2.4 Rotated quantile regression

As previously mentioned, the quantile estimates are obtained by minimizing a rotated check function (See equation 8). The minimization problem can be written as the following linear programming problem:[3]

$$Min_{\beta_\tau,u,v} \sum_{i=1}^{N} \hat{G}_{\tau i}u_i + (1-\hat{G}_{\tau i})v_i \tag{9}$$

such that:

$$\mathbf{y} - \mathbf{X}\beta_\tau = \mathbf{u} - \mathbf{v} \tag{10}$$

$$\mathbf{u} \geq \mathbf{0}_n \tag{11}$$

$$\mathbf{v} \geq \mathbf{0}_n \tag{12}$$

where $\mathbf{0}_n$ is a vector of 0s, $\mathbf{X}$ is the matrix of observations of the covariates, $\mathbf{y}$ is the vector of observations of the outcome, and $\mathbf{u}$ and $\mathbf{v}$ are added to the inequality constraint to transform it into an equality.

This linear programming problem could be solved using the LinearProgram() class in Stata or alternatively using the Stata integration with Python. However, we implement an interior point algorithm developed by Portnoy and Koenker (1997) by translating the Matlab code used by Arellano and Bonhomme (2017a) to Mata language.[4]

3 The qregsel command

In this section we describe the `qregsel` command to implement a copula-based sample selection correction in quantile regression.

3. This closely follows the quantile regression example for linear programming available in the Mata reference manual (see example 3 for LinearProgram() in StataCorp (2019a)).
4. The Matlab's routine was written by Daniel Morillo and Roger Koenker in Ox, translated to Matlab by Paul Eilers, and slightly modified by Roger Koenker. It can be found in the supplemental material of Arellano and Bonhomme (2017b), and in Roger Koenker's website.

3.1 Syntax

The syntax of the `qregsel` command is:

`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`]

3.2 Options

`select(`[$depvar_S$ =] *varlist*$_S$`)` specifies the selection equation. If $depvar_S$ is specified, it should be coded as 0 and 1, with 0 indicating an outcome not observed for an observation and 1 indicating an outcome observed for an observation. *select()* is required.

`quantile(#)` estimate # quantiles. *quantile()* is required.

`grid_min(grid_minvalue)` specifies the minimum value to be considered in the grid search. *grid_min()* is required.

`grid_max(grid_maxvalue)` specifies the maximum value to be considered in the grid search. *grid_max()* is required.

`grid_length(grid_length)` specifies the length of the (evenly spaced) grid to be considered in the grid search. *grid_length()* is required.

`copula(`*copula*`)` specifies a copula function governing the dependence between the errors in the outcome equation and selection equation. *copula* may be one of the following: *gaussian*, *frank*, *fgm*, and *amh*. The default is *copula(gaussian)*.

`noconstant` suppresses the constant term in the outcome equation.

`plot` generates a graph of the value of the objective function being minimized over the values of rho (the parameter of the copula) considered in the grid search.

3.3 Returned values

`qregsel` saves the following in e():

Scalars

`e(N)`	Number of observations
`e(rank)`	Number of parameters
`e(df_r)`	Degrees of freedom
`e(rho)`	Copula parameter
`e(it)`	Number of iterations

Macros

`e(copula)`	Specified copula
`e(depvar)`	Dependent variable
`e(indepvars)`	Independent variables
`e(cmdline)`	Command line
`e(outcome_eq)`	Outcome equation
`e(select_eq)`	Selection equation
`e(predict)`	Predict command name
`e(cmd)`	Command name
`e(title)`	Quantile selection model

Matrices

`e(coefs)`	Coefficient matrix
`e(grid)`	Values of the objective function minimized over the grid

Functions

`e(sample)`	Marks estimation sample

3.4 Prediction

After the execution of `qregsel`, the `predict` command is available to compute a counterfactual of the outcome variable corrected for sample selection. Here is its syntax:

`predict newvar` [*if*] [*in*]

The counterfactual outcomes are constructed by randomly generating an integer q between 1 and 99 for each individual in the full sample, and then using the quantile coefficients associated with each draw of q to produce a prediction of the qth quantile of the outcome distribution. This approach follows the conditional quantile decomposition method of Machado and Mata (2005) and has been recently applied for example in Bollinger et al. (2019).

3.5 Inference

Confidence intervals for any of the parameters can be estimated using methods such as the conventional nonparametric bootstrap, or alternatively using subsampling (Politis et al. (1999)) as done in Arellano and Bonhomme (2017b) due to the computational advantage when using large sample sizes.

In our empirical application we illustrate how to use bootstrap to create a confidence interval for the copula parameter.

3.6 Dependency of qregsel

`qregsel` depends on the Mata function mm_cond(), which is part of the moremata package (Jann 2005). If not already installed, you can install it by typing ssc install moremata.

4 Empirical Example

We use the fictional data set used in the documentation of the Heckman selection model in the Stata base reference manual (see StataCorp (2019b)) to study wages of women. As in the example, we assume that the hourly wage is a function of education and age, whereas the likelihood of working (and hence the wage being observed) is a function of marital status, the number of children at home, and (implicitly) the wage (via the inclusion of age and education). We do not take the logarithm of wage as it is usually done, however the variable in the fictional data set has already a bell-shaped histogram (see Figure 1). In addition, we follow the example in the Stata 16 base reference manual by not including squared age as it is standard in this type of regression.

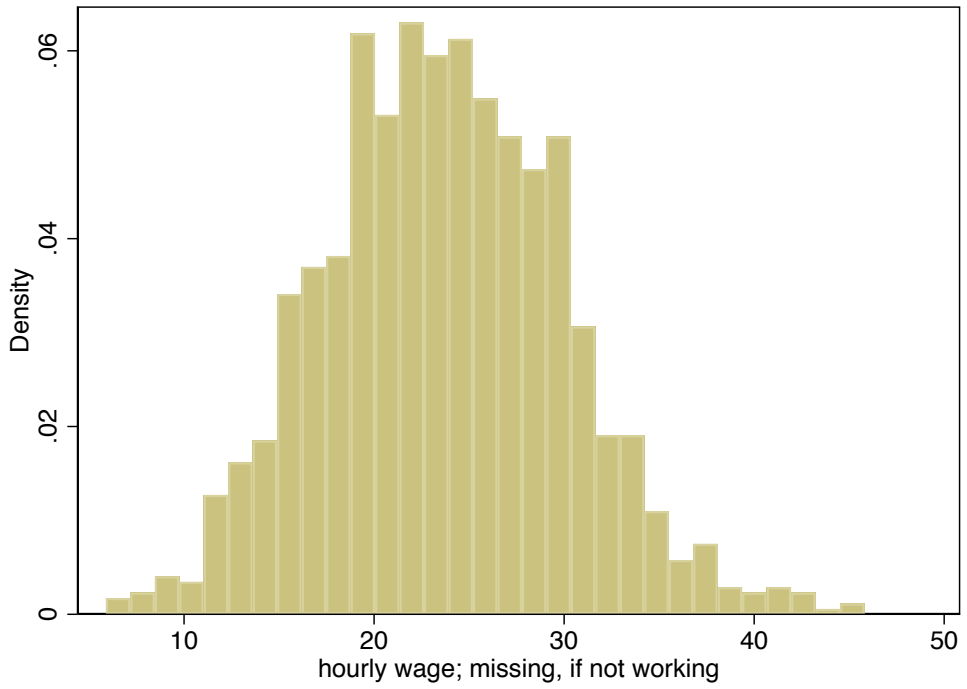

Figure 1: Histogram of wage

First, we estimate a quantile regression over the quantiles 0.1, 0.5, and 0.9 without

corrections for sample selection.

```
. webuse womenwk,clear
. sqreg wage educ age, quantile(.1 .5 .9)
(fitting base model)
Bootstrap replications (20)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....................
Simultaneous quantile regression                  Number of obs =      1,343
  bootstrap(20) SEs                               .10 Pseudo R2 =     0.1068
                                                  .50 Pseudo R2 =     0.1429
                                                  .90 Pseudo R2 =     0.1523
```

wage	Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q10						
education	.8578176	.0651588	13.17	0.000	.7299933	.985642
age	.1234271	.0247599	4.98	0.000	.0748547	.1719995
_cons	.5154006	1.298974	0.40	0.692	-2.032842	3.063644
q50						
education	.9064927	.0734632	12.34	0.000	.7623772	1.050608
age	.160184	.0249218	6.43	0.000	.111294	.2090739
_cons	5.312029	1.235723	4.30	0.000	2.887867	7.73619
q90						
education	.930661	.0998569	9.32	0.000	.7347682	1.126554
age	.1579835	.0331353	4.77	0.000	.0929808	.2229863
_cons	12.20975	1.90783	6.40	0.000	8.467094	15.95241

Next we turn to the estimation of a quantile regression accounting for sample selection by using the command `qregsel` with a Gaussian copula. We specify a grid that goes between -.9 and .9 with steps of length .05.

```
. 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

```
. ereturn list
scalars:
                  e(N) =  1343
               e(rank) =  3
               e(df_r) =  1340
                e(rho) =  -.65
                 e(it) =  14
macros:
             e(copula) : "gaussian"
```

```
            e(depvar) : "wage"
          e(indepvars) : "education age _cons"
            e(cmdline) : "qregsel wage  education age, select(married children educ age)"
         e(outcome_eq) : "wage  education age"
       e(selection_eq) : "married children educ age"
                e(cmd) : "qregsel"
            e(predict) : "qregsel_p"
              e(title) : "Quantile selection model"

matrices:
               e(coefs) :  3 x 3
                e(grid) :  37 x 1

functions:
              e(sample)
```

Figure 3 shows the plot we obtain from specifying the option plot. The value of rho that minimizes the criterion function is equal to -0.65, as stored in e(rho).

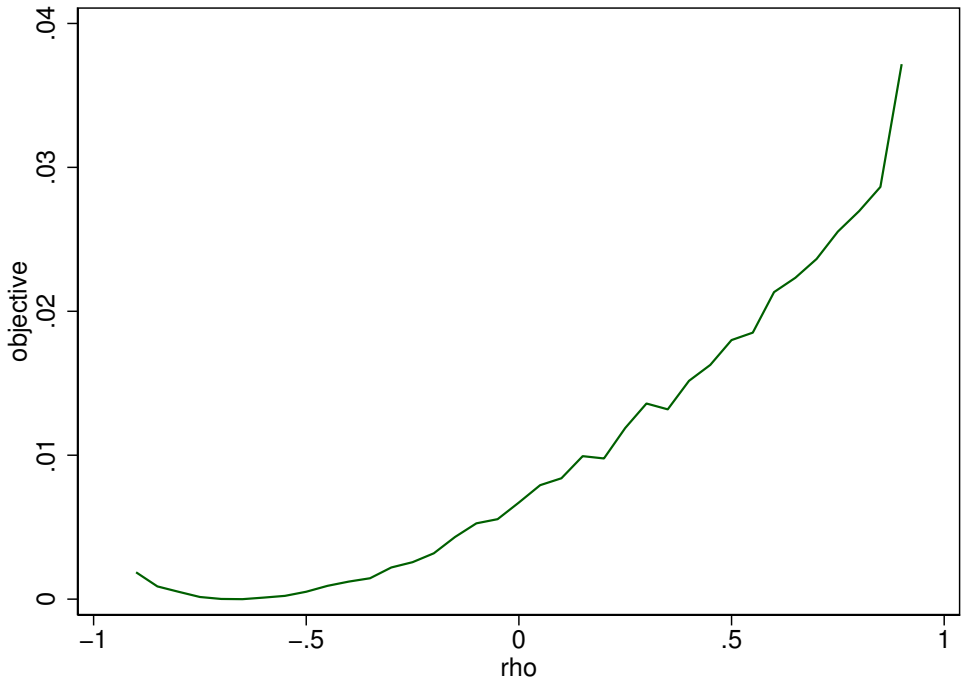

Figure 2: Grid for minimization

After the estimation a counterfactual distribution that is corrected for sample selection may be generated with the post estimation command `predict` as follows.

```
. set seed 1
. predict wage_hat
```

```
. _pctile wage_hat, nq(20)
. mat qs = J(19,3,.)
. forvalues i=1/19{
  2. mat qs[`i´,1] = r(r`i´)
  3. }
. _pctile wage, nq(20)
. forvalues i=1/19{
  2. mat qs[`i´,2] = r(r`i´)
  3. mat qs[`i´,3] = `i´
  4. }
. svmat qs, name(quantiles)
. twoway connected quantiles1 quantiles2 quantiles3, scheme(s1color) ///
>  xtitle("Ventile") ytitle("Wage") legend(order(1 "Corrected" 2 "Uncorrected"))
```

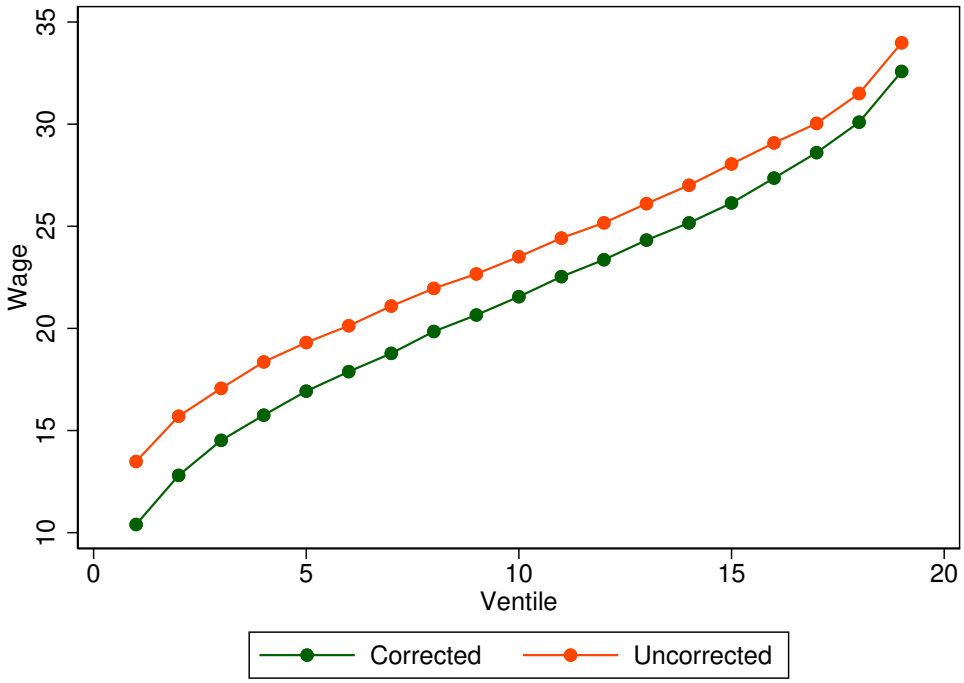

Figure 3: Corrected versus uncorrected quantiles

Finally, we illustrate the use of the `bootstrap` command to construct a confidence interval for the copula parameter ρ using 99 replications. This could be easily extended to the coefficients of the quantile regression or some other statistic computed with the counterfactual distribution.

```
. set seed 12345
. webuse womenwk,clear
```

```
. global wage_eqn wage educ age
. global seleqn married children educ age
. capture program drop myqregsel
. program myqregsel, eclass
  1.      version 16
  2.      tempname bb
  3.      quietly qregsel $wage_eqn, select($seleqn) quantile(.5) ///
>           grid_min(-.9) grid_max(.9) grid_length(.05)
  4.      matrix `bb´=e(rho)
  5.      ereturn post `bb´
  6.      ereturn local cmd="bootstrap"
  7. end
. bootstrap _b, reps(99) nowarn: myqregsel
(running myqregsel on estimation sample)
Bootstrap replications (99)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................    50
.....................................................
Bootstrap results                                 Number of obs     =      2,000
                                                  Replications      =         99
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c1 |       -.65   .0768337    -8.46   0.000    -.8005913   -.4994087
------------------------------------------------------------------------------
```

5 Concluding remarks

In this article, we introduce a new Stata module called `qregsel`, which implements a copula-based method proposed in Arellano and Bonhomme (2017b) to correct for sample selection in quantile regressions.

Two recent applications of the econometric method here introduced include the analysis of the gender gap between earnings distributions in Maasoumi and Wang (2019), and the analysis of earnings inequality correcting for non-response in Bollinger et al. (2019).

6 Acknowledgments

We thank Jim Albrecht and Wim Vijverberg for useful comments and suggestions.

7 References

Arellano, M., and S. Bonhomme. 2017a. Sample Selection in Quantile Regression: A Survey. In *Handbook of Quantile Regression*, ed. R. Koenker, V. Chernozhukov, X. He, and L. Peng, 1st ed., chap. 13, 463. Chapman and Hall/CRC.

———. 2017b. Quantile Selection Models With an Application to Understanding Changes in Wage Inequality. *Econometrica* 85(1): 1–28.

Bollinger, C. R., B. Hirsch, C. Hokayem, and J. P. 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): 2143–2185.

Hasebe, T. 2013. Copula-based Maximum-Likelihood Estimation of Sample-Selection Models. *The Stata Journal* 13: 547–573.

Heckman, J. J. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47(1): 153–161.

Huber, M., and B. Melly. 2015. A Test of the Conditional Independence Assumption in Sample Selection Models. *Journal of Applied Econometrics* 30(7): 1144–1168.

Jann, B. 2005. moremata: Stata module (Mata) to provide various functions. *Statistical Software Components S455001, Department of Economics, Boston College* . http://ideas.repec.org/c/boc/bocode/s455001.html.

Koenker, R., and G. Bassett. 1978. Regression Quantiles. *Econometrica* 46(1): 33–50.

Maasoumi, E., and L. Wang. 2019. The Gender Gap between Earnings Distributions. *Journal of Political Economy* 127(5): 2438–2504.

Machado, J. A. F., and J. Mata. 2005. Counterfactual Decomposition of Changes in Wage Distributions using Quantile Regression. *Journal of Applied Econometrics* 20: 445–465.

Politis, D., J. Romano, and M. Wolf. 1999. *Subsampling*. Springer Series in Statistics.

Portnoy, S., and R. Koenker. 1997. The Gaussian Hare and the Laplacian Tortoise : Computability of Squared-Error versus Absolute-Error Estimators. *Statistical Papers* 12(4): 279–300.

StataCorp. 2019a. *Mata Reference Manual*. College Station, TX: Stata Press.

———. 2019b. Stata 16 Base Reference Manual. College Station, TX: Stata Press.

Vella, F. 1998. Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources* 33(1): 127–169.

About the authors

Ercio Munoz is Ph.D. candidate in Economics at CUNY Graduate Center.

Mariel Siravegna is Ph.D. candidate in Economics at Georgetown University.