The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–18

Implementing Quantile Selection Models in Stata

Ercio Muñoz
CUNY Graduate Center
New York, USA
emunozsaavedra@gc.cuny.edu

Mariel Siravegna
Georgetown University
Washington DC, USA
mcs92@georgetown.edu

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). The command allows the user to model selection in quantile regressions using either a Gaussian or an one-dimensional Frank copula. We illustrate the use of **qregsel** with two examples. First, we apply the method to the fictional data set employed in the Stata base reference manual for the `heckman` command. Second, we replicate part of the empirical application of the original paper using data for the UK that covers the period 1978-2000 to compare wages of males and females at different quantiles.

Keywords: st0001, sample selection, quantile regression, copula method

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 two empirical examples. First, we estimate a

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

quantile regression model with sample selection using the Stata base reference manual example for the `heckman` command. Second, we replicate the analysis of wage inequality in the UK for the period 1978-2000 as in the original paper.

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 examples, 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, such as quantile regression, which precludes the use of the control function approach. The assumption of additive separability of observables and unobservables in the output equation 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 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. Hence, we observe (*Y,D,Z*) where $Y = Y^*$ only when *D=1*.

Under the set of assumptions[2] detailed in Arellano and Bonhomme (2017b), we have

2. Assumptions: 1) Z is independent of *(U,V)|X* (exclusion restriction), 2) absolutely continuous

that the cdf of Y^* conditional on participation and for all $\tau \in (0,1)$ is:

$$Pr(Y^* \leq x'\beta(\tau)|D=1, Z=z) = Pr(U \leq \tau | V \leq p(z), Z=z) = G_x(\tau, p(z)) \quad (3)$$

where $G_x \equiv C(\tau,p)/p$ is the conditional copula function, which measures the dependence between U and V. Here 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 outcomes (that is, when $D=1$) coincides with the conditional τ-quantile of latent outcomes, which implies that if we are able to estimate the mapping $G_x(\tau, p)$ from latent to observed ranks, we are able to recover $Q(\tau, x)$ from the observed outcomes (i.e. we are able to estimate the τ-quantile correcting for selection).

To implement the method, we assume that the copula function is indexed by a single parameter such that:

$$G_x(\tau,p) \equiv G(\tau,p;\rho) = \frac{C(\tau,p;\rho)}{p} \quad (4)$$

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.

2.2 Estimation

Arellano and Bonhomme (2017b)'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.

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

$$\hat{\gamma} = argmax_a \sum_{i=1}^{N} D_i ln\Phi(Z_i'a) + (1-D_i) ln\Phi(-Z_i'a) \quad (5)$$

The second step is to estimate ρ by minimizing a method-of-moments objective function, 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 is accomplished with a grid search over different values of ρ such that:

$$\hat{\rho} = argmin_c \| \sum_{i=1}^{N} \sum_{l=1}^{L} D_i \varphi(\tau_l, Z_i) [\mathbf{1}\{Y_i \leq X_i'\hat{\beta}(\tau_l, c)\} - G(\tau_l, \Phi(Z_i'\hat{\gamma}), c)] \| \quad (6)$$

bivariate distribution of *(U,V)*|*X*=*x* with standard uniform marginals and rectangular support, 3) continuous outcome, and 4) propensity score, *p(z)*>0 with probability 1.

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:

$$\hat{\beta}_\tau(c) = argmin_{b(\tau)} \sum_{i=1}^{N} D_i[G(\tau, \Phi(Z_i'\hat{\gamma}); c)(Y_i - X_i'b(\tau))^+ + \tag{7}$$

$$(1 - G(\tau, \Phi(Z_i'\hat{\gamma}); c))(Y_i - X_i'b(\tau))^-] \tag{8}$$

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.[3]

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

$$\hat{\beta}(\tau) = argmin_{b(\tau)} \sum_{i=1}^{N} D_i[\hat{G}_{\tau i}(Y_i - X_i'b(\tau))^+ + (1 - \hat{G}_{\tau i})(Y_i - X_i'b(\tau))^-] \tag{9}$$

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

Note that the third step is unnecessary if the quantiles of interest are included in the set $\tau_1 < \tau_2 < \ldots < \tau_L$ used in the second step.

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 and an one-parameter Frank. 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	$-\rho^{-1} log\{1 + \frac{(e^{-\rho U} - 1)(e^{-\rho V} - 1)}{(e^{-\rho} - 1)}\}$	$-\infty \leq \rho \leq \infty$

3. In our implementation we use a grid of 9 values (0.1,0.2,...,0.9), and $\varphi(\tau_l, Z_i) = \varphi(Z_i) = p(Z_i; \hat{\rho})$ as in Arellano and Bonhomme (2017b) empirical example.

2.4 Measures of dependence

The parameter ρ that governs the degree of dependence is not directly comparable across copulas (see Hasebe 2013). For this reason, researchers often report Kendall's τ or the Spearman rank correlation coefficient as a measure of the degree of dependence. Both measures take the range of $[-1, 1]$, where a value closer to 1 (-1) indicates a stronger (negative) dependence, and in the case of our copulas can be expressed as closed form in terms of ρ (see Table 2).

Table 2: Copula functions and measures of dependence

Copula name	Range of ρ	Kendall's τ	Spearman's rank correlation
Gaussian	$-1 \leq \rho \leq 1$	$\frac{2}{\pi} sin^{-1}(\rho)$	$\frac{6}{\pi} sin^{-1}(\rho/2)$
Frank	$-\infty \leq \rho \leq \infty$	$1 + \frac{4}{\rho}\{D_1(\rho) - 1\}$	$1 + \frac{12}{\rho}\{D_2(\rho) - D_1(\rho)\}$

Notes: $D_n(\rho)$ is a Debye function, where $D_n(\rho) = \frac{n}{\rho^n} \int_0^\rho \frac{t^n}{e^t - 1} dt$.

2.5 Rotated quantile regression

As previously mentioned, the quantile estimates are obtained by minimizing a rotated check function (see equation 9). The minimization problem can be written as the following linear programming problem:[4]

$$Min_{\beta_\tau, u, v} \sum_{i=1}^{N} \hat{G}_{\tau i} u_i + (1 - \hat{G}_{\tau i}) v_i \tag{10}$$

such that:

$$\mathbf{y} - \mathbf{X}\beta_\tau = \mathbf{u} - \mathbf{v} \tag{11}$$

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

$$\mathbf{v} \geq \mathbf{0}_n \tag{13}$$

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 (2017b) to Mata language.[5]

4. This closely follows the quantile regression example for linear programming available in the Mata reference manual (see example 3 for LinearProgram() in StataCorp (2019a)).
5. The Matlab's routine was originally written by Daniel Morillo and Roger Koenker in Ox, trans-

3 The qregsel command

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

3.1 Syntax

The syntax of the `qregsel` command is:

`qregsel` *depvar* [*indepvars*] [*if*] [*in*] , *select(*[$depvar_S$ =] *varlist*$_S$*)*

`quantile(#)` [`copula(`*copula*`)` `noconstant finergrid coarsergrid rescale nodots`]

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.

`copula(`*copula*`)` specifies a copula function governing the dependence between the errors in the outcome equation and selection equation. *copula* may be *gaussian* or *frank*. The default is *copula(gaussian)*.

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

`finergrid` finds the value of the copula parameter using a grid of 199 values (values such that the Spearman rank correlation is approximately [-0.99,-0.985,..,0.985,0.99]) instead of 100 (values such that the Spearman rank correlation is approximately [-0.99,-0.98,..,0.98,0.99]), as done by default.

`coarsergrid` finds the value of the copula parameter using a grid of 50 values (values such that the Spearman rank correlation is approximately [-0.99,-0.95,..,0.93,0.97]) instead of 100 (values such that the Spearman rank correlation is approximately [-0.99,-0.98,..,0.98,0.99]), as done by default.

`rescale` transforms the independent variables in the outcome equation by subtracting from each its sample mean and dividing each by its standard deviation.

`nodots` suppresses progress dots that indicate status over the grid search.

---

lated 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.

3.3 Returned values

`qregsel` saves the following in e():

Scalars

`e(N)`	Number of observations	`e(N_selected)`	Number of selected observations
`e(rho)`	Copula parameter	`e(kendall)`	Kendall's tau
`e(spearman)`	Spearman's rank correlation		

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(cmd)`	Command name	`e(predict)`	Predict command name
`e(rescale)`	Use of rescale option	`e(title)`	Quantile selection model
`e(properties)`	b		

Matrices

`e(b)`	Coefficient vector	`e(grid)`	Matrix with the values of the objective function for each value of rho, and its respective Spearman rank correlation and Kendall's tau
`e(coefs)`	Coefficient matrix		

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

where the list of new variables must contain two new variable names, the first one for the counterfactual outcome variable, and the second one for a binary indicator of selection, to be generated respectively.

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).

The selection indicator is generated by randomly drawing values of the error in the selection equation V from the conditional distribution of V given $U{=}u$, derived from the chosen copula using the estimated copula parameter and the values of U randomly generated to create the counterfactual outcome variable in the previous paragraph. This approach follows the empirical exercise performed in Arellano and Bonhomme (2017b).

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 (see Politis et al. 1999) as done in Arellano and Bonhomme (2017b) due to the computational advantage when using large sample sizes.

In our first empirical application we illustrate how to use bootstrap to create a confidence interval for the estimated coefficients of the quantile regression and the copula parameter.

4 Empirical Examples

In this section we illustrate the use of the command with two empirical examples. First, we use the classic example of wages of women in which we use the data available from the Stata manual example for the command `heckman`. Second, we replicate part of an exercise presented in Arellano and Bonhomme (2017b) with data from the UK.

4.1 Wages of women

In this application 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. 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.

First, we estimate a quantile regression over the quantiles 0.1, 0.5, and 0.9 without corrections for sample selection as a benchmark.

```
. 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

------------------------------------------------------------------------------
             |               Bootstrap
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q10          |
```

```
   education |   .8578176   .0822727    10.43   0.000     .6964203    1.019215
         age |   .1234271   .0206434     5.98   0.000     .0829302    .1639239
       _cons |   .5154006   1.256476     0.41   0.682    -1.949473    2.980274
-------------+----------------------------------------------------------------
q50          |
   education |   .9064927   .0638967    14.19   0.000     .7811443    1.031841
         age |    .160184   .0313763     5.11   0.000      .098632    .2217359
       _cons |   5.312029   1.007443     5.27   0.000     3.335692    7.288366
-------------+----------------------------------------------------------------
q90          |
   education |    .930661   .0856044    10.87   0.000     .7627278    1.098594
         age |   .1579835   .0462329     3.42   0.001     .0672868    .2486803
       _cons |   12.20975    1.55745     7.84   0.000     9.154448    15.26506
------------------------------------------------------------------------------
```

Next we turn to the estimation of a quantile regression accounting for sample selection by using the command `qregsel` with a Gaussian copula. In addition, we plot the value of the objective function over the minimization grid (see Figure 1). The value of rho that minimizes the criterion function is approximately equal to -0.65, as stored in e(rho). The interpretation of this estimated value is that women with higher wages (higher U) tend to participate more (lower V).

```
. 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
------------------------------------------------------------------------------
. ereturn list
scalars:
                  e(N) =  2000
         e(N_selected) =  1343
                e(rho) =  -.647834836
```

```
            e(kendall) =  -.43389025
           e(spearman) =  -.63
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(select_eq) : "married children educ age"
                e(cmd) : "qregsel"
            e(predict) : "qregsel_p"
            e(rescale) : "non-rescaled"
              e(title) : "Quantile selection model"
         e(properties) : "b"
matrices:
                 e(b) :  1 x 9
              e(grid) :  100 x 4
             e(coefs) :  3 x 3
functions:
             e(sample)
. svmat e(grid), name(col)
. qui gen lvalue = log10(value)
. twoway connected lvalue spearman
```

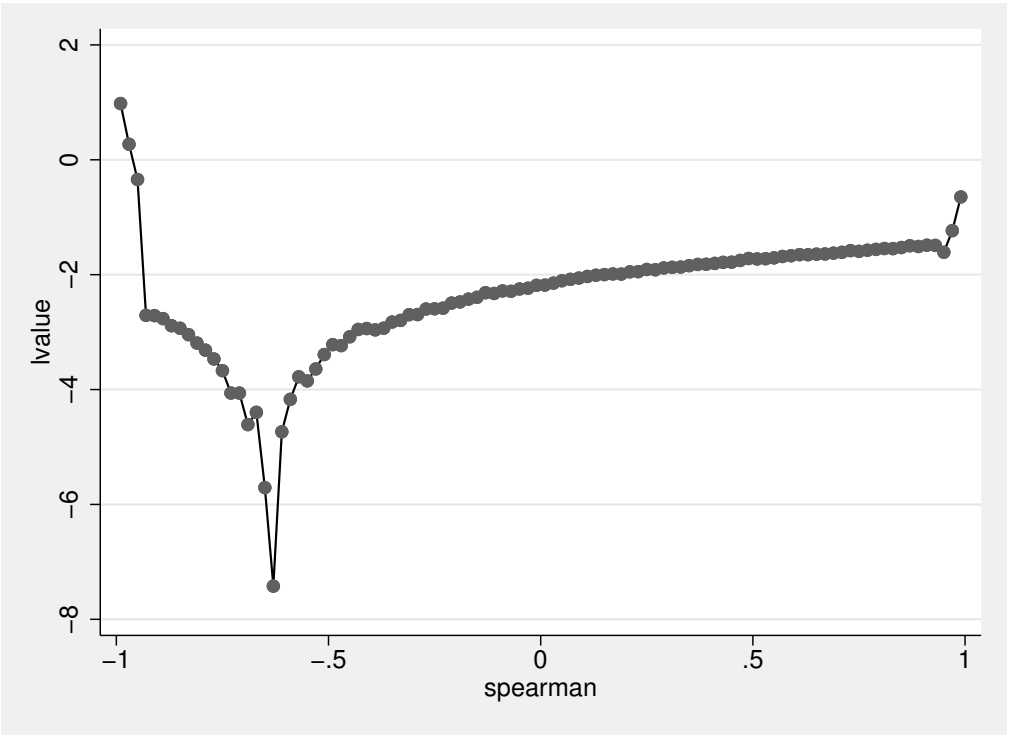
After the estimation a counterfactual distribution that is corrected for sample selection may be generated with the post estimation command `predict` as follows. Figure 2 displays the ventiles of the distribution corrected for sample selection versus the uncorrected one. We can see how wages are lower after correcting for selection at each ventile of the distribution.

```
. set seed 1
. predict wage_hat participation
. _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, ///
>  xtitle("Ventile") ytitle("Wage") legend(order(1 "Corrected" 2 "Uncorrected"))
```

Finally, we illustrate the use of the `bootstrap` command to construct a confidence interval for the coefficients associated to three different quantiles and the copula parameter ρ using 100 replications.

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

Figure 1: Grid for minimization

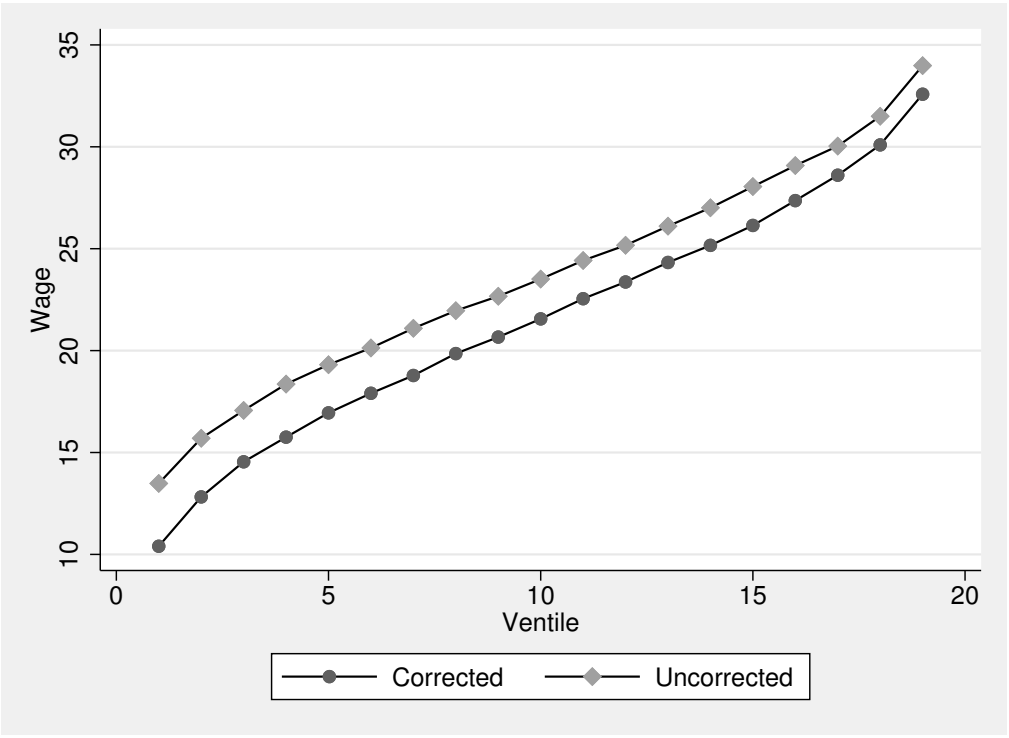



```
Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
Bootstrap results                                 Number of obs     =      2,000
                                                  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               Bootstrap
             |      Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
q10          |
```

Figure 2: Corrected versus uncorrected quantiles



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

4.2 Wage inequality in UK

In this example we apply the model to measure market-level changes in wage inequality in the UK. We compare wages of males and females at different quantiles of the wage distribution, correcting for selection into work. We replicate Arellano and Bonhomme (2017b) using the data set provided by the authors, which originally comes from the Family Expenditure Survey (FES) from 1978 to 2000.[6]

We model log-hourly wages Y and employment status D. The controls X include linear, quadratic, and cubic time trends, four cohort dummies (born in 1919-1934, 1935-1944, 1955-1964, and 1965-1977, omitting 1945-1954), two education dummies (end of schooling at 17 or 18, and end of schooling after 18), 11 regional dummies, marital status, and the number of kids split by age categories (six dummies, from 1 year old to 17-18 years old).

The excluded regressor follows Blundell et al. (2003) and corresponds to their measure of potential out-of-work (welfare) income, interacted with marital status. This variable was constructed for each individual in the sample using the Institute of Fiscal Studies tax and welfare-benefit simulation model.

Arellano and Bonhomme (2017b) estimate the sample selection model independently by gender and marital status. We replicate (see code below) the exercise reported in the paper using a Frank copula and find that the copula parameter in the case of married individuals is -1.548 for males and -1.035 for females (the associated rank correlations are -0.250 and -0.170, respectively). For single individuals is -7.638 for males and -0.421 for females (the respective rank correlations are -0.790 and -0.070). After the estimation using each sub-sample, we use `predict` to generate counterfactual outcomes, which are then used to plot quantiles by gender with and without correction for sample selection over time. We are able to replicate the empirical facts documented in the original paper (see Figure 3). We see that correcting for sample selection makes an important difference at the bottom of the wage distribution for males while the difference seems to be less important in the case of women.

```
. ** Female and single
. set seed 3
. use data_2 if married==0,clear
. global wage_eqn lw ed17 ed18 trend1 trend2 trend3 c1919_34 c1935_44 c1955_64 ///
>  c1965_77 reg_d1 reg_d2 reg_d3 reg_d4 reg_d5 reg_d6 reg_d7 reg_d8 reg_d9 ///
>  reg_d10 reg_d11 kids_d1 kids_d2 kids_d3 kids_d4 kids_d5 kids_d6
. global seleqn s_zero ed17 ed18 trend1 trend2 trend3 c1919_34 c1935_44 ///
>  c1955_64 c1965_77 reg_d1 reg_d2 reg_d3 reg_d4 reg_d5 reg_d6 reg_d7 reg_d8 ///
>  reg_d9 reg_d10 reg_d11 kids_d1 kids_d2 kids_d3 kids_d4 kids_d5 kids_d6
. qregsel $wage_eqn, select($seleqn) rescale quantile(50) copula(frank) finergrid
Grid for the copula parameter (199)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.................................................
.................................................
.................................................
.................................................
```

6. The data and replication codes can be found here.

```
Quantile selection model                        Number of obs      =    23583
                                                Selected           =    15185
                                                Nonselected        =     8398
Copula parameter (frank):    -0.42
```

lw	Coef.
q50	
ed17	.1107013
ed18	.2078859
trend1	-.0541206
trend2	.4185438
trend3	-.2659457
c1919_34	-.0203966
c1935_44	-.0127007
c1955_64	-.0211737
c1965_77	-.064329
reg_d1	.007508
reg_d2	.0145522
reg_d3	.02818
reg_d4	.0140872
reg_d5	.0236211
reg_d6	.0070201
reg_d7	.1256261
reg_d8	.0708555
reg_d9	.0187373
reg_d10	.0041181
reg_d11	.032367
kids_d1	-.0102305
kids_d2	-.0126629
kids_d3	-.0342705
kids_d4	-.0577489
kids_d5	-.0541355
kids_d6	-.0115029
_cons	1.76145

```
. matlist e(rho)
```

	c1
r1	-.421

```
. predict yhat participation
. keep yhat lw year
. tempfile data_2_single
. qui save `data_2_single´
.
. ** Female and married
. use data_2 if married==1,clear
. global seleqn m_zero ed17 ed18 trend1 trend2 trend3 c1919_34 c1935_44 ///
>  c1955_64 c1965_77 reg_d1 reg_d2 reg_d3 reg_d4 reg_d5 reg_d6 reg_d7 reg_d8 ///
>  reg_d9 reg_d10 reg_d11 kids_d1 kids_d2 kids_d3 kids_d4 kids_d5 kids_d6
. qui: qregsel $wage_eqn, select($seleqn) rescale quantile(50) copula(frank) finergrid
. matlist e(rho)
```

	c1
r1	-1.035

```
. predict yhat participation
```

```
. keep yhat lw year
. tempfile data_2_married
. qui save `data_2_married´
.
. ** Male and single
. use data_1 if married==0,clear
. global seleqn s_zero ed17 ed18 trend1 trend2 trend3 c1919_34 c1935_44 ///
>  c1955_64 c1965_77 reg_d1 reg_d2 reg_d3 reg_d4 reg_d5 reg_d6 reg_d7 reg_d8 ///
>  reg_d9 reg_d10 reg_d11 kids_d1 kids_d2 kids_d3 kids_d4 kids_d5 kids_d6
. qui: qregsel $wage_eqn, select($seleqn) rescale quantile(50) copula(frank) finergrid
. matlist e(rho)
             |        c1
-------------+----------
          r1 |    -7.638
. predict yhat participation
. keep yhat lw year
. tempfile data_1_single
. qui save `data_1_single´
.
. ** Male and married
. use data_1 if married==1,clear
. global seleqn m_zero ed17 ed18 trend1 trend2 trend3 c1919_34 c1935_44 ///
>  c1955_64 c1965_77 reg_d1 reg_d2 reg_d3 reg_d4 reg_d5 reg_d6 reg_d7 reg_d8 ///
>  reg_d9 reg_d10 reg_d11 kids_d1 kids_d2 kids_d3 kids_d4 kids_d5 kids_d6
. qui: qregsel $wage_eqn, select($seleqn) rescale quantile(50) copula(frank) finergrid
. matlist e(rho)
             |        c1
-------------+----------
          r1 |    -1.548
. predict yhat participation
. keep yhat lw year
. tempfile data_1_married
. qui save `data_1_married´
.
. ** Plotting quantiles
. use `data_2_married´,clear
. append using `data_2_single´
.
. forvalues i=78(1)100 {
  2. _pctile yhat if year==`i´, p(10 20 30 40 50 60 70 80 90)
  3. mat qs = 1,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\nullmat(qs)
  4. }
. forvalues i=78(1)100 {
  2. _pctile lw if year==`i´, p(10 20 30 40 50 60 70 80 90)
  3. mat qs = 2,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\qs
  4. }
.
. use `data_1_married´,clear
. append using `data_1_single´
.
. forvalues i=78(1)100 {
```

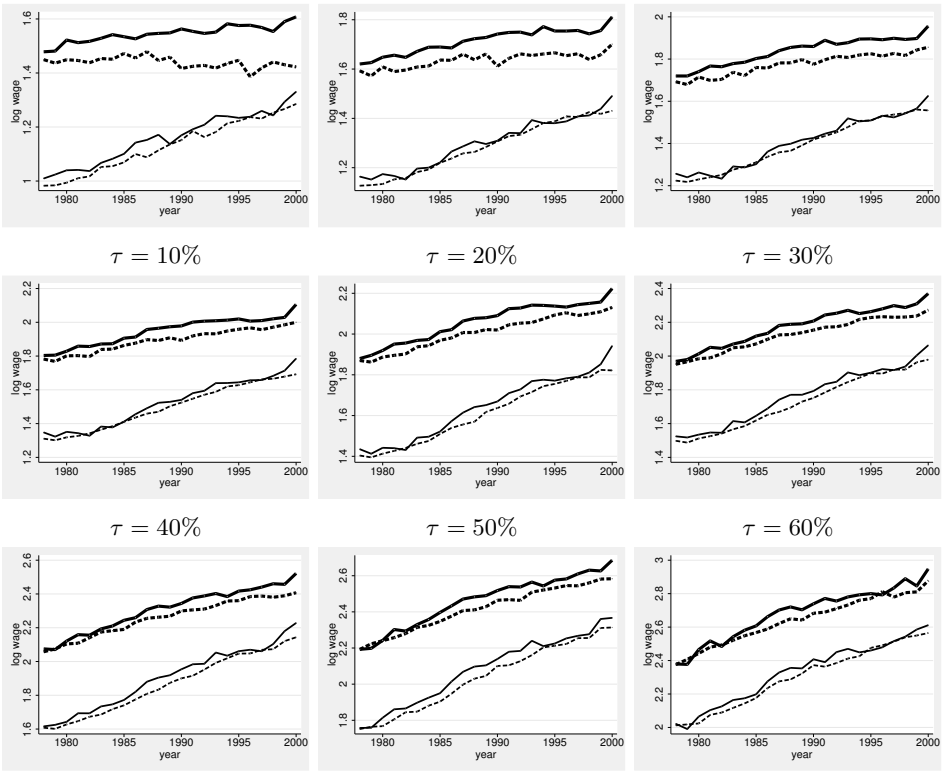
```
  2. _pctile yhat if year==`i´, p(10 20 30 40 50 60 70 80 90)
  3. mat qs = 3,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\qs
  4. }
. forvalues i=78(1)100 {
  2. _pctile lw if year==`i´, p(10 20 30 40 50 60 70 80 90)
  3. mat qs = 4,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\qs
  4. }
. mat colnames qs = serie year q10 q20 q30 q40 q50 q60 q70 q80 q90
. clear
. svmat qs, name(col)
number of observations will be reset to 92
Press any key to continue, or Break to abort
number of observations (_N) was 0, now 92
. reshape wide q*, i(year) j(serie)
(note: j = 1 2 3 4)
Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       92   ->      23
Number of variables                  11   ->      37
j variable (4 values)             serie   ->   (dropped)
xij variables:
                                    q10   ->   q101 q102 ... q104
                                    q20   ->   q201 q202 ... q204
                                    q30   ->   q301 q302 ... q304
                                    q40   ->   q401 q402 ... q404
                                    q50   ->   q501 q502 ... q504
                                    q60   ->   q601 q602 ... q604
                                    q70   ->   q701 q702 ... q704
                                    q80   ->   q801 q802 ... q804
                                    q90   ->   q901 q902 ... q904
-----------------------------------------------------------------------------
. qui replace year=1900+year
.
. local k=10
. while `k´<=90{
  2. twoway scatter q`k´3 q`k´4 q`k´1 q`k´2 year, c(l l l l) ms(p p p p) ///
>  lwidth(vthick vthick thick thick) lpattern(dash solid dash solid) ///
>  legend(off) xtitle("year",size(large)) ytitle("log wage",size(large)) ///
>  xlabel(,labsize(large)) ylabel(,labsize(large)) name(q`k´,replace)
  3. qui graph export "q`k´.eps", replace
  4. local k=`k´+10
  5. }
```

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. The use of the command is illustrated with two empirical examples.

Additional empirical applications of the econometric method here implemented include the analysis of the gender gap between earnings distributions in Maasoumi and

Figure 3: Wage quantiles, by gender



Notes: Quantiles of log-hourly wages, conditional on employment (solid lines) and corrected for selection (dashed). Male wages are plotted in thick lines, while female wages are in thin lines.

Wang (2019), and the analysis of earnings inequality correcting for non-response in Bollinger et al. (2019).

6 Acknowledgments

We thank Jim Albrecht, Wim Vijverberg, and the participants of the 2020 Virtual Stata Conference 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.

Blundell, R., H. Reed, and T. M. Stoker. 2003. Interpreting Aggregate Wage Growth: The Role of Labor Market Participation. *American Economic Review* 93(4): 1114–1131.

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): 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.

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.