Implementing Quantile Selection Models in Stata*

Ercio Muñoz† Mariel Siravegna‡

September 10, 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). 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: *sample selection, quantile regression, copula method*

*We thank Jim Albrecht, Wim Vijverberg, and the participants of the 2020 Virtual Stata Conference 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 two empirical examples. First, we estimate a 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

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

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

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

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:

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

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. Assumptions: 1) Z is independent of *(U,V)*$|X$ (exclusion restriction), 2) absolutely continuous bivariate distribution of *(U,V)*, 3) continuous outcome, and 4) propensity score, *p(z)* >0.

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:

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

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:

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

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.

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.

2.3 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	$-\rho^{-1} log\{1 + \frac{(e^{-\rho U}-1)(e^{-\rho V}-1)}{(e^{-\rho}-1)}\}$	$-\infty \leq \rho \leq \infty$

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]

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` `rescale`]

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.

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, 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.

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

`rescale` transform the regressors in the outcome equation by substracting from each its sample mean and dividing each by its standard deviation.

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(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

Matrices

`e(coefs)`	Coefficient matrix	`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

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

In our second 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.

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 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
```

wage	Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q10						
education	.8578176	.0752264	11.40	0.000	.7102433	1.005392
age	.1234271	.0230344	5.36	0.000	.0782397	.1686144
_cons	.5154006	1.596242	0.32	0.747	-2.616004	3.646805
q50						
education	.9064927	.066965	13.54	0.000	.7751251	1.03786
age	.160184	.0276063	5.80	0.000	.1060278	.2143402
_cons	5.312029	1.104635	4.81	0.000	3.145027	7.47903
q90						
education	.930661	.0818078	11.38	0.000	.7701757	1.091146
age	.1579835	.0308184	5.13	0.000	.0975259	.2184412
_cons	12.20975	1.356139	9.00	0.000	9.549367	14.87014

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)
Quantile selection model                            Number of obs      =      1343
```

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

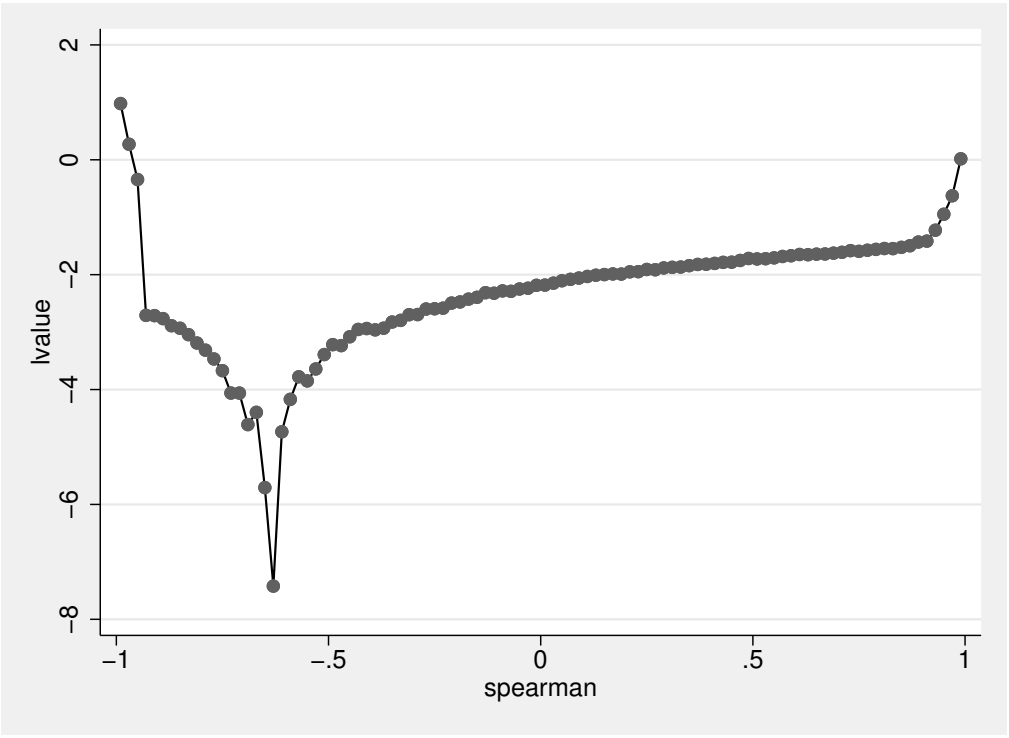
```
. ereturn list
scalars:
                  e(N) =  1343
               e(rank) =  3
               e(df_r) =  1340
                e(rho) =  -.647834836
            e(kendall) =  -.43389025
           e(spearman) =  -.63
macros:
             e(copula) : "gaussian"
             e(depvar) : "wage"
          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"
matrices:
              e(coefs) :  3 x 3
               e(grid) :  100 x 4
functions:
             e(sample)
. svmat e(grid), name(col)
. gen lvalue = log10(value)
(1,900 missing values generated)
. 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 param-

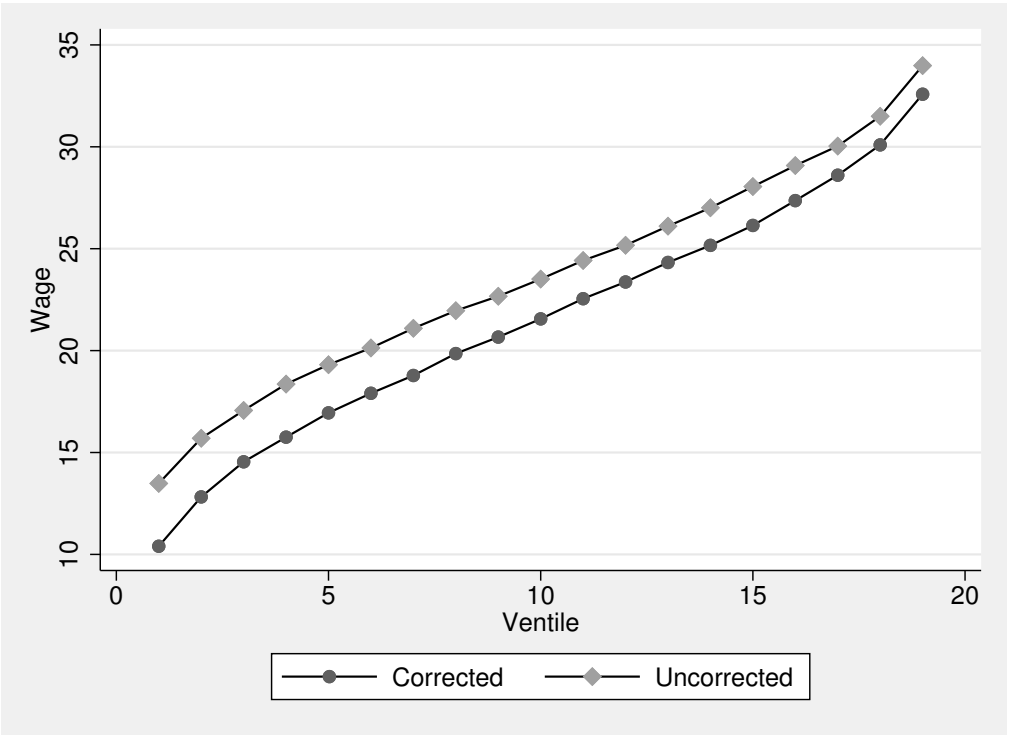
Figure 1: Grid for minimization



eter ρ using 99 replications.

```
. set seed 2
. 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(.1 .5 .9)
  4.         local colnames : colfullnames e(coefs)
  5.         local rownames : rowfullnames e(coefs)
  6.         foreach lname1 of local colnames   {
  7.                 foreach lname2 of local rownames   {
  8.                         local names = "`names´ `lname1´:`lname2´"
  9.                 }
 10.         }
 11.         mata: st_matrix("`bb´", vec(st_matrix("e(coefs)"))´)
```

Figure 2: Corrected versus uncorrected quantiles




```
 12.     matrix `bb´ = `bb´,e(rho)
 13.         mat colnames `bb´ = `names´ rho:rho
 14.     ereturn post `bb´
 15.         ereturn local cmd="bootstrap"
 16. 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]
-------------+----------------------------------------------------------------
q10          |
   education |   1.112866   .1470708     7.57   0.000     .8246128     1.40112
         age |    .204362   .0494404     4.13   0.000     .1074606    .3012634
------------------------------------------------------------------------------
```

_cons	-8.498507	2.496796	-3.40	0.001	-13.39214	-3.604876
q50						
education	1.017025	.0708082	14.36	0.000	.8782432	1.155806
age	.2028979	.0279242	7.27	0.000	.1481674	.2576283
_cons	.5828089	1.384067	0.42	0.674	-2.129912	3.29553
q90						
education	.8888879	.0627717	14.16	0.000	.7658577	1.011918
age	.2272004	.0262191	8.67	0.000	.1758118	.2785889
_cons	8.914994	1.128363	7.90	0.000	6.703443	11.12655
rho						
rho	-.6478348	.0731618	-8.85	0.000	-.7912294	-.5044403

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.

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 are able to replicate (see code below) the estimates reported in the paper using a Frank copula and find that the copula parameter in the case of married individuals is -1.417 for males and -1.035 for females (the associated rank correlations are -0.230 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).[6] 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 again 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

6. The small difference in the values of the copula parameter for married males, married females, and single females is due to a difference in the grid, given that we use 100 values instead of 199.

difference seems to be less important in the case of women.

```
. #delimit ;
delimiter now ;
. use "data_2.dta" 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(10 50 90) copula(frank);

Quantile selection model                        Number of obs      =    15185
```

	q10	q50	q90
ed17	.0665847	.1107013	.0988773
ed18	.1454937	.2078859	.2108566
trend1	.0057028	-.0541205	.2040515
trend2	.1476044	.4185437	-.0039498
trend3	-.0999532	-.2659457	-.0449968
c1919_34	-.0154949	-.0203966	3.24e-06
c1935_44	-.0187101	-.0127007	.000309
c1955_64	-.0050833	-.0211737	-.0458773
c1965_77	-.0250789	-.064329	-.1073678
reg_d1	-.0002014	.007508	.0033621
reg_d2	.0138982	.0145522	-.0091101
reg_d3	.0248777	.02818	.004356
reg_d4	.0170612	.0140871	-.0039688
reg_d5	.0191299	.0236211	.0020279
reg_d6	.0187921	.0070201	.0023608
reg_d7	.1027003	.1256261	.0998025
reg_d8	.0640108	.0708555	.0445042
reg_d9	.0075992	.0187373	-.0069425
reg_d10	-.0032203	.0041181	-.0077449
reg_d11	.0319565	.032367	-.0028479
kids_d1	-.0167587	-.0102305	-.0071818
kids_d2	-.026656	-.0126629	-.0176153
kids_d3	-.0372376	-.0342705	-.0149425
kids_d4	-.0675987	-.0577489	-.0445404
kids_d5	-.0586155	-.0541355	-.0560959
kids_d6	-.0021103	-.0115029	-.0062805
_cons	1.280272	1.76145	2.268481

```
. predict yhat participation;
. matlist e(rho);
```

	c1
r1	-.421

```
. keep yhat lw year;
. qui save "data_2_single",replace;
. use "data_2.dta" 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(10 50 90) copula(frank);
```

```
. predict yhat participation;
. matlist e(rho);
             |        c1
-------------+------------
          r1 |     -1.035
. keep yhat lw year;
. qui save "data_2_married",replace;
. use "data_1.dta" 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(10 50 90) copula(frank);
. predict yhat participation;
. matlist e(rho);
             |        c1
-------------+------------
          r1 |     -7.638
. keep yhat lw year;
. qui save "data_1_single",replace;
. use "data_1.dta" 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(10 50 90) copula(frank);
. predict yhat participation;
. matlist e(rho);
             |        c1
-------------+------------
          r1 |     -1.417
. keep yhat lw year;
. qui save "data_1_married",replace;
. use "data_2_married.dta",clear;
. append using "data_2_single.dta";
. mat quantiles = J(1,11,.);
. forvalues i=78(1)100 {;
  2. _pctile yhat if year==`i´, p(10 20 30 40 50 60 70 80 90);
  3. mat quantiles = 1,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\quantiles;
  4.        };
. forvalues i=78(1)100 {;
  2. _pctile lw if year==`i´, p(10 20 30 40 50 60 70 80 90);
  3. mat quantiles = 2,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\quantiles;
  4.        };
. use "data_1_married.dta",clear;
. append using "data_1_single.dta";
. forvalues i=78(1)100 {;
  2. _pctile yhat if year==`i´, p(10 20 30 40 50 60 70 80 90);
  3. mat quantiles = 3,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\quantiles;
  4.         };
. forvalues i=78(1)100 {;
  2. _pctile lw if year==`i´, p(10 20 30 40 50 60 70 80 90);
```

```
  3. mat quantiles = 4,`i´,r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\quantiles;
  4.       };
. mat colnames quantiles = serie year q10 q20 q30 q40 q50 q60 q70 q80 q90;
. clear;
. qui svmat quantiles, name(col);
. qui drop if serie==.;
. qui reshape wide q*, i(year) j(serie);
. 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. graph export "q`k´.eps", replace;
  4. local k=`k´+10;
  5.     };
.  #delimit cr
delimiter now cr
```

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.

Recent empirical 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 References

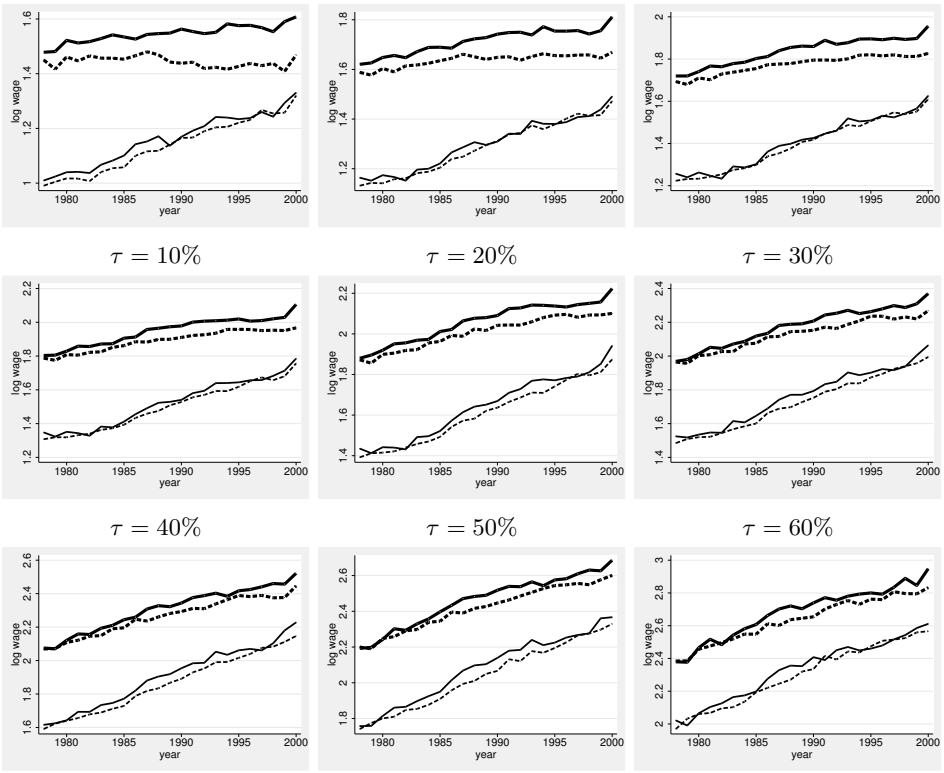
Arellano, M., and S. Bonhomme. 2017a. Sample Selection in Quantile Regression: A Survey. In *Handbook of Quantile Regression*, ed. R. Koenker, V. Chernozhukov, X. He, and L. Peng, 1st ed., chap. 13, 463. Chapman and Hall/CRC.

———. 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.

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.

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.