The Stata Journal (2021)
21, Number 4, pp. 952–971 DOI: 10.1177/1536867X211063148

Implementing quantile selection models in Stata

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

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

Abstract. In this article, we describe `qregsel`, a community-contributed command that implements a copula-based sample-selection correction for quantile regression recently proposed by Arellano and Bonhomme (2017, *Econometrica* 85: 1–28). The command allows the user to model selection in quantile regressions by using either a Gaussian or a one-dimensional Frank copula. We illustrate the use of `qregsel` with two examples. First, we apply the method to the fictional dataset used in the *Stata Base Reference Manual* for the `heckman` command. Second, we replicate part of the empirical application of the original article using data for the United Kingdom that cover the period 1978–2000 to compare wages of males and females at different quantiles.

Keywords: st0657, qregsel, sample selection, quantile regression, copula method

1 Introduction

Nonrandom 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 fitting 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 nonrandom sample selection in the context of quantile regression. For example, Arellano and Bonhomme (2018) offer a survey of recently proposed methods with a focus on a copula-based sample-selection model suggested in Arellano and Bonhomme (2017).

As discussed in Arellano and Bonhomme (2018), 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 nonadditive with nonlinear quantile curves on the selected sample (see Huber and Melly [2015]).

In this article, we briefly discuss the copula-based approach proposed by Arellano and Bonhomme (2017) and present a new community-contributed command called `qregsel` that implements it.[1] In addition, we illustrate the method with two empirical examples. First, we fit a quantile regression model with sample selection using the *Stata Base*

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

Reference Manual example for the `heckman` command. Second, we replicate the analysis of wage inequality in the United Kingdom for the period 1978–2000 as in the original article.

This article 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 (2017). The goal is to obtain a consistent estimator when there is sample selection in a nonadditive 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 (c.d.f.) 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:

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

Y^* is the latent outcome variable (for example, wage offers), the function Q is the τth conditional quantile of Y^* given the covariates $\mathbf{X}$ (for example, education and experience), and U is the error term of the outcome equation.

The participation equation is defined as

$$D = \mathbb{I}\{V \leq p(\mathbf{Z})\}$$

where D takes values equal to 1 when the latent variable is observable (for example, employment) and equal to 0 otherwise, $\mathbf{Z}$ contains $\mathbf{X}$ and at least one covariate B that do not appear in the outcome equation (for example, a determinant of employment that does not affect wages directly), $p(\mathbf{Z})$ is a propensity score, and V is an error term of the selection equation. Hence, we observe $(Y, D, \mathbf{Z})$ where $Y = Y^*$ only when $D = 1$.

Under the set of assumptions[2] detailed in Arellano and Bonhomme (2017), we have that the c.d.f. of Y^*, conditional on participation and for all $\tau \in (0,1)$, is

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

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 $\mathbf{X}=\mathbf{x}$) to ranks $G_x\{\tau, p(\mathbf{z})\}$ in the distribution of observed outcomes conditional on participation (given $\mathbf{Z}=\mathbf{z}$). Namely, the conditional $G_x\{\tau, p(\mathbf{z})\}$ 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 $Q(\tau,\mathbf{x})$ from the observed outcomes (that is, 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

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

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's (2017) estimation algorithm can be summarized in three steps: estimation of the propensity score; estimation of the degree of selection via the c.d.f. 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 $\boldsymbol{\gamma}$ by a probit regression:

$$\widehat{\boldsymbol{\gamma}} = \text{argmax}_a \sum_{i=1}^{N} D_i \ln\Phi(\mathbf{Z}_i'\mathbf{a}) + (1-D_i)\ln\Phi(-\mathbf{Z}_i'\mathbf{a})$$

The second step is to estimate ρ by minimizing a method-of-moments objective function, which allows 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

$$\widehat{\rho} = \text{argmin}_c \left\| \sum_{i=1}^{N}\sum_{l=1}^{L} D_i \varphi(\tau_l,\mathbf{Z}_i) \left[\mathbf{1}\left\{ Y_i \leq \mathbf{X}_i'\widehat{\beta}(\tau_l,c) \right\} - G\left\{ \tau_l, \Phi\left(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}\right); c \right\} \right] \right\|$$

2. Assumptions: 1) $\mathbf{Z}$ is independent of $(U,V)|\mathbf{X}$ (exclusion restriction), 2) absolutely continuous bivariate distribution of $(U,V)|\mathbf{X}=\mathbf{x}$ with standard uniform marginals and rectangular support, 3) continuous outcome, and 4) propensity score, $p(\mathbf{z})>0$ with probability 1.

where $\|\cdot\|$ is the Euclidean norm, $\tau_1 < \tau_2 < \cdots < \tau_L$ is a finite grid on $(0,1)$, and the instrument functions are defined as $\varphi(\tau, \mathbf{Z}_i)$, where the dim $\varphi \leq$ dim ρ, and

$$\widehat{\beta}_\tau(c) = \text{argmin}_{b(\tau)} \sum_{i=1}^{N} D_i \Big(G\left\{\tau, \Phi\left(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}\right); c\right\} \left\{Y_i - \mathbf{X}_i' b(\tau)\right\}^+ + \left[1 - G\left\{\tau, \Phi\left(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}\right); c\right\}\right] \left\{Y_i - \mathbf{X}_i' b(\tau)\right\}^- \Big)$$

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]

Finally, using $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\rho}$ obtained above, the third step consists of computing $\widehat{G}_{\tau i} = G\{\tau, \Phi(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}); \widehat{\rho}\}$ for all i to estimate $\beta(\tau)$ by minimizing a rotated check function of the form

$$\widehat{\beta}(\tau) = \text{argmin}_{b(\tau)} \sum_{i=1}^{N} D_i \left[\widehat{G}_{\tau i} \left\{Y_i - \mathbf{X}_i' b(\tau)\right\}^+ + \left(1 - \widehat{G}_{\tau i}\right) \left\{Y_i - \mathbf{X}_i' b(\tau)\right\}^- \right] \tag{1}$$

where $\widehat{\beta}(\tau)$ will be a consistent estimator of the τth quantile regression coefficient.

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

2.3 Copulas

The Arellano and Bonhomme (2018) 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 consider only the case of a reduced set of one-dimensional copulas. We include the Gaussian and a 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	$-\rho^{-1}\log\left\{1 + \frac{(e^{-\rho U}-1)(e^{-\rho V}-1)}{(e^{-\rho}-1)}\right\}$	$-\infty \leq \rho \leq \infty$

3. In our implementation, we use a grid of nine values $(0.1, 0.2, \ldots, 0.9)$, and $\varphi(\tau_l, \mathbf{Z}_i) = \varphi(\mathbf{Z}_i) = p(\mathbf{Z}_i; \widehat{\rho})$ as in Arellano and Bonhomme's (2017) empirical example.

2.4 Measures of dependence

The parameter ρ, which 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	$-\infty \leq \rho \leq \infty$	$1 + \frac{4}{\rho}\{D_1(\rho) - 1\}$	$1 + \frac{12}{\rho}\{D_2(\rho) - D_1(\rho)\}$

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

2.5 Rotated quantile regression

As previously mentioned, the quantile estimates are obtained by minimizing a rotated check function [see (1)]. The minimization problem can be written as the linear programming problem[4]

$$\text{Min}_{\beta_\tau,u,v} \sum_{i=1}^{N} \widehat{G}_{\tau i} u_i + \left(1 - \widehat{G}_{\tau i}\right) v_i$$

such that

$$\begin{aligned} \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_\tau &= \mathbf{u} - \mathbf{v} \\ \mathbf{u} &\geq \mathbf{0}_n \\ \mathbf{v} &\geq \mathbf{0}_n \end{aligned}$$

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 (2017) 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 [2021b]).
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 (2017) and on Roger Koenker's website.

3 The qregsel command

In this section, we describe the `qregsel` command, which implements 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(`# [# [# ...]]`)` specifies the quantiles to be estimated and should contain numbers between 0 and 1, exclusive. Numbers larger than 1 are interpreted as percentages. `quantile()` is required.

`copula(`*copula*`)` specifies a copula function governing the dependence between the errors in the outcome equation and the 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 by 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 by 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.

3.3 Stored results

`qregsel` stores 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 τ
`e(spearman)`	Spearman's rank correlation

Macros

`e(cmd)`	`qregsel`
`e(cmdline)`	command as typed
`e(depvar)`	name of dependent variable
`e(indepvars)`	independent variables
`e(title)`	title in estimation output
`e(copula)`	specified `copula()`
`e(outcome_eq)`	outcome equation
`e(select_eq)`	selection equation
`e(rescale)`	use of the `rescale` option
`e(properties)`	`b`
`e(predict)`	program used to implement `predict`

Matrices

`e(b)`	coefficient vector
`e(grid)`	matrix with the values of the objective function for each value of ρ and its respective Spearman rank correlation and Kendall's τ
`e(coefs)`	coefficient matrix; each column corresponds to the coefficients for a quantile

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. The syntax is

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

where *newvarlist* must contain the names for two new variables: the first one for the counterfactual outcome variable and the second one for a binary indicator of selection.

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 (2017).

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, Romano, and Wolf [1999]) as done in Arellano and Bonhomme (2017) because of 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 (2017) with data from the United Kingdom.

4.1 Wages of women

In this application, we use the fictional dataset used in the documentation of the Heckman selection model in the *Stata Base Reference Manual* (see StataCorp [2021a]) 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 dataset already has a bell-shaped histogram. In addition, we follow the example in the Stata 17 *Base Reference Manual* by not including squared age because 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
. 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	Coefficient	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. In addition, we plot the value of the objective function over the minimization grid (see figure 1). The value of ρ 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	Coefficient
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 childr.."
         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)
. quietly generate lvalue = log10(value)
```

```
. twoway connected lvalue spearman
```

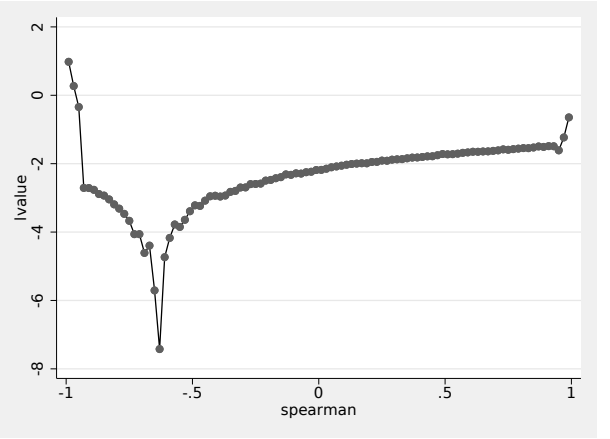


Figure 1. Grid for minimization

After the estimation, a counterfactual distribution that is corrected for sample selection may be generated with the postestimation 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)
. matrix 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"))
```

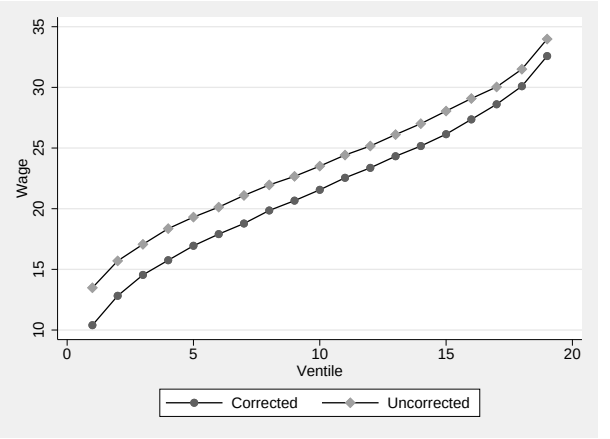


Figure 2. Corrected versus uncorrected quantiles

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)

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 coefficient	Bias	Bootstrap std. err.	[95% conf.	interval]	
q10						
education	1.1128663	-.0367987	.14715321	.7483546	1.322367	(P)
age	.20436202	-.0064956	.04906217	.0912168	.2998732	(P)
_cons	-8.4985072	.7398983	2.4903276	-11.27083	-2.926636	(P)
q50						
education	1.0170248	.0091963	.0704033	.9073696	1.155043	(P)
age	.20289786	.0008098	.02794786	.1479627	.2588321	(P)
_cons	.58280893	-.1816018	1.3878812	-1.880296	2.965075	(P)
q90						
education	.88888792	.0150792	.06247062	.7735702	1.034392	(P)
age	.22720039	-.0033785	.02609233	.1670902	.2715747	(P)
_cons	8.9149942	-.102337	1.1223042	6.964433	10.89201	(P)
_eq4						
rho	-.64783484	-.0218298	.07382065	-.8230287	-.5277461	(P)

Key: P: Percentile

4.2 Wage inequality in the United Kingdom

In this example, we apply the model to measure market-level changes in wage inequality in the United Kingdom. We compare wages of males and females at different quantiles of the wage distribution, correcting for selection into work. We replicate Arellano and Bonhomme (2017) using the dataset provided by the authors, which originally comes from the Family Expenditure Survey 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, 4 cohort dummies (born in 1919–1934, 1935–1944, 1955–1964, and 1965–1977, omitting 1945–1954), 2 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 (6 dummies, from 1 year old to 17–18 years old).

6. The data and replication codes can be found here: https://www.econometricsociety.org/publications/econometrica/2017/01/01/quantile-selection-models-application-understanding-changes.

The excluded regressor follows Blundell, Reed, and Stoker (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 by using the Institute of Fiscal Studies tax and welfare-benefit simulation model.

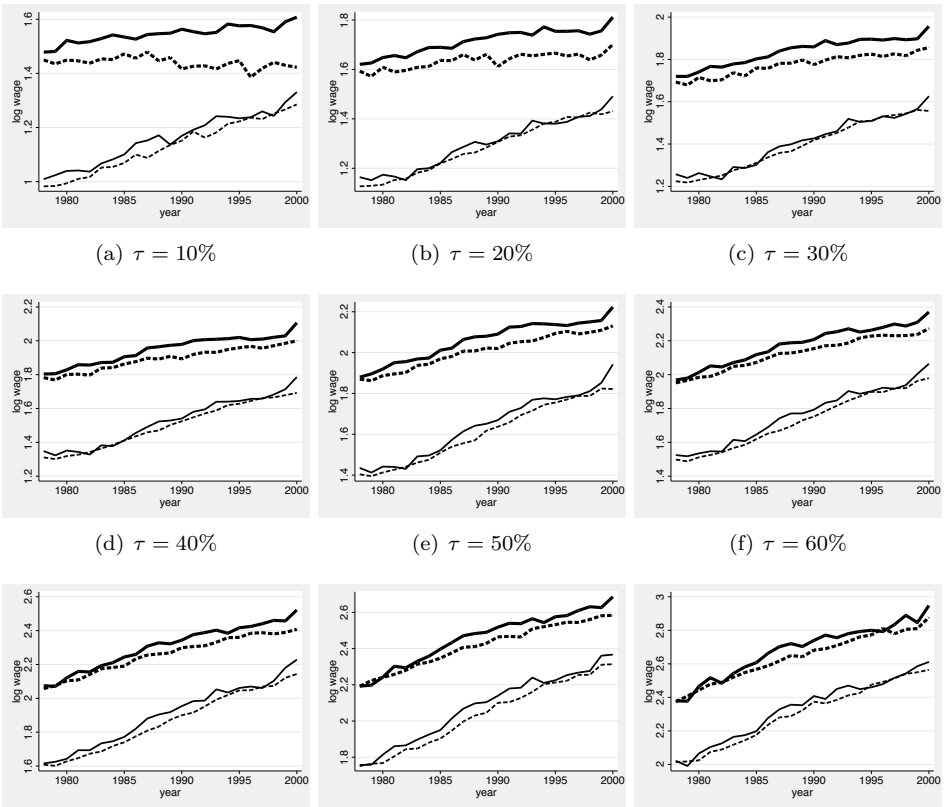
Arellano and Bonhomme (2017) fit the sample-selection model independently by gender and marital status. We replicate (see code below) the exercise reported in the article using a Frank copula and find that the copula parameter in the case of married individuals is -1.548 for males and -1.035 for females (the associated rank correlations are -0.250 and -0.170, respectively). For single individuals, the copula parameter 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 subsample, 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 article (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 for females.

```
. ** 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
.................................................
.................................................
.................................................
.................................................
Quantile selection model                        Number of obs      =    23583
                                                Selected           =    15185
                                                Nonselected        =     8398
Copula parameter (frank):    -0.42
------------------------------------------------------------------------------
          lw | Coefficient
-------------+----------------------------------------------------------------
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

. quietly 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
```

```
. quietly: 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
. quietly 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
. quietly: 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
. quietly 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
. quietly: 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
. quietly 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. matrix 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. matrix 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.dta'
. forvalues i=78(1)100 {
  2. _pctile yhat if year==`i', p(10 20 30 40 50 60 70 80 90)
  3. matrix 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. matrix qs = 4,`i',r(r1),r(r2),r(r3),r(r4),r(r5),r(r6),r(r7),r(r8),r(r9)\qs
  4. }
. matrix 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
-----------------------------------------------------------------------------
. replace year=1900+year
(23 real changes made)
. 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. }
```



(a) $\tau = 10\%$ (b) $\tau = 20\%$ (c) $\tau = 30\%$

(d) $\tau = 40\%$ (e) $\tau = 50\%$ (f) $\tau = 60\%$

(g) $\tau = 70\%$ (h) $\tau = 80\%$ (i) $\tau = 90\%$

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.

5 Concluding remarks

In this article, we introduced a new community-contributed command called `qregsel`, which implements a copula-based method proposed in Arellano and Bonhomme (2017) to correct for sample selection in quantile regressions. The use of the command was illustrated with two empirical examples.

Additional empirical applications of the econometric method here implemented included the analysis of the gender gap between earnings distributions in Maasoumi and Wang (2019) and the analysis of earnings inequality correcting for nonresponse 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 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 21-4
. net install st0657     (to install program files, if available)
. net get st0657         (to install ancillary files, if available)
```

8 References

Arellano, M., and S. Bonhomme. 2017. Quantile selection models with an application to understanding changes in wage inequality. *Econometrica* 85: 1–28. https://doi.org/10.3982/ECTA14030.

———. 2018. Sample selection in quantile regression: A survey. In *Handbook of Quantile Regression*, ed. R. Koenker, V. Chernozhukov, X. He, and L. Peng, chap. 13. Handbooks of Modern Statistical Methods, Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9781315120256-13.

Blundell, R., H. Reed, and T. M. Stoker. 2003. Interpreting aggregate wage growth: The role of labor market participation. *American Economic Review* 93: 1114–1131. https://doi.org/10.1257/000282803769206223.

Bollinger, C. R., B. T. Hirsch, C. M. Hokayem, and J. P. Ziliak. 2019. Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch. *Journal of Political Economy* 127: 2143–2185. https://doi.org/10.1086/701807.

Hasebe, T. 2013. Copula-based maximum-likelihood estimation of sample-selection models. *Stata Journal* 13: 547–573. https://doi.org/10.1177/1536867X1301300307.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161. https://doi.org/10.2307/1912352.

Huber, M., and B. Melly. 2015. A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics* 30: 1144–1168. https://doi.org/10.1002/jae.2431.

Koenker, R., and G. Bassett, Jr. 1978. Regression quantiles. *Econometrica* 46: 33–50. https://doi.org/10.2307/1913643.

Maasoumi, E., and L. Wang. 2019. The gender gap between earnings distributions. *Journal of Political Economy* 127: 2438–2504. https://doi.org/10.1086/701788.

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. https://doi.org/10.1002/jae.788.

Politis, D. N., J. P. Romano, and M. Wolf. 1999. *Subsampling*. New York: Springer.

Portnoy, S., and R. Koenker. 1997. The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science* 12: 279–300. https://doi.org/10.1214/ss/1030037960.

StataCorp. 2021a. *Stata 17 Base Reference Manual*. College Station, TX.

———. 2021b. *Stata 17 Mata Reference Manual*. College Station, TX.

Vella, F. 1998. Estimating models with sample selection bias: A survey. *Journal of Human Resources* 33: 127–169. https://doi.org/10.2307/146317.

About the authors

Ercio Muñoz is a PhD candidate in economics at CUNY Graduate Center.

Mariel Siravegna is a PhD candidate in economics at Georgetown University.