The Stata Journal (2021)
21, Number 3, pp. 602–625  DOI: 10.1177/1536867X211045516

arhomme: An implementation of the Arellano and Bonhomme (2017) estimator for quantile regression with selection correction

Martin Biewen
University of Tübingen
Tübingen, Germany
martin.biewen@uni-tuebingen.de

Pascal Erhardt
University of Tübingen
Tübingen, Germany
pascal.erhardt@uni-tuebingen.de

Abstract. Despite constituting a major theoretical breakthrough, the quantile selection model of Arellano and Bonhomme (2017, *Econometrica* 85: 1–28) based on copulas has not found its way into many empirical applications. We introduce the command `arhomme`, which implements different variants of the estimator along with standard errors based on bootstrapping and subsampling. We illustrate the command by replicating parts of the empirical application in the original article and a related application in Arellano and Bonhomme (2018, *Handbook of Quantile Regression*, chap. 13).

Keywords: st0648, arhomme, Arellano and Bonhomme quantile selection model, quantile regression, selection correction, inequality, distribution

1 Introduction

Ever since the contributions by Gronau (1974) and Heckman (1974), economists and researchers from other disciplines have been aware of the possibility that measured relationships may suffer from selection bias. The classic example is the determinants of pay (that is, wages) and the selectivity through participation in employment. If one is interested in measuring how individuals with certain characteristics are paid, one has to deal with the possibility that some of them may actually not take up employment, especially if their potential pay is too low (in comparison with their alternative options). If these individuals differ in terms of unobservables from the general population, omitting them from wage regressions will yield biased estimates of regression coefficients.

Following Heckman (1979), a large literature has studied generalized models of sample selection for regression models, for example, Ahn and Powell (1993); Andrews and Schafgans (1998); Chen and Khan (2003); and Das, Newey, and Vella (2003). This literature initially focused on correcting regressions for the mean outcome (for example, the mean wage). An even more challenging case is to correct *entire outcome distributions* for selection bias. In an influential contribution, Buchinsky (1998, 2001) proposed a control function approach to correcting quantile regressions for selection bias. However, it was later shown by Huber and Melly (2015) that the proposed correction was based on restrictive assumptions that are unlikely to hold in general (conditional independence and additivity). It was not until the contribution by Arellano and Bonhomme (2017)

that the selection problem for entire distributions was solved in some generality. In particular, Arellano and Bonhomme (2017) showed that, in the general case, sample selection corrections may not be additive but nonlinearly "rotate" observed distributional ranks.

Despite representing a theoretical breakthrough, Arellano and Bonhomme's (2017) method has not yet found its way into many empirical applications (recent exceptions include Maasoumi and Wang [2019] and Bollinger et al. [2019]). The purpose of this article is to provide an implementation of their method that is easy to use by practitioners. We also provide some replications of original analyses in Arellano and Bonhomme (2018, 2017). More generally, Arellano and Bonhomme's (2017) contribution is part of an active recent literature that addresses the problem of correcting entire distributions for selection with potential applications in many fields (for example, Albrecht, van Vuuren, and Vroman [2009]; Picchio and Mussida [2011]; Fernández-Val, van Vuuren, and Vella [2018]; D'Haultfoeuille et al. [2020]; and Biewen, Fitzenberger, and Seckler [2020]).

The rest of this article is organized as follows. Section 2 describes the Arellano and Bonhomme (2017) selection model and estimation method. Section 3 introduces and describes the command `arhomme`, which implements this estimation method along with several options. Section 4 presents three empirical examples, two of them being successful replications of original applications in Arellano and Bonhomme (2018, 2017). Section 5 concludes.

2 The Arellano and Bonhomme (2017) method

2.1 Model

Although they consider more general versions in theoretical parts of their analysis, the practical version of the Arellano and Bonhomme (2017) quantile regression model with selection correction takes the form

$$Y^* = \mathbf{X}'\boldsymbol{\beta}(U) \tag{1}$$

$$D = 1\{V \leq p(\mathbf{Z})\} \tag{2}$$

$$Y = Y^* \text{ if } D = 1 \tag{3}$$

where Y^* is the potential outcome, D the selection indicator, and Y the observed outcome (available only for individuals with $D = 1$). The vectors $\mathbf{X}$ and $\mathbf{Z}$ are covariate vectors, where $\mathbf{X}$ is assumed to be a strict subset of $\mathbf{Z}$ (exclusion restriction). The uniformly distributed variable U denotes the rank of the individual in the conditional distribution $Y^*|\mathbf{X}$, while the uniformly distributed V represents a normalized error term describing the resistance toward selection. The propensity score $p(\mathbf{Z}) = P(D = 1|\mathbf{Z})$ describes the selection probability of individuals with characteristics $\mathbf{Z}$. The propensity score is assumed to follow a probit model; that is, $p(\mathbf{Z}) = \Phi(\mathbf{Z}'\boldsymbol{\gamma})$. The main substantive assumption of the model is that (U, V) is jointly statistically independent of $\mathbf{Z}$ given $\mathbf{X}$.

Equation (1) is a linear quantile regression model for the potential outcome Y^* defining the value of Y^* that an individual with rank U would get if he or she was selected (for example, the Uth quantile in a distribution of wage offers for individuals with characteristics $\mathbf{X}$). Equation (2) specifies that, among individuals with characteristics $\mathbf{Z}$, a percentage of $p(\mathbf{Z})$ gets selected, but only those whose resistance toward selection V is low enough. Equation (3) states that outcomes are observed only for selected individuals (for example, an individual would earn some wage Y^* if he or she decided to work, but this wage is observed only if he or she actually decides to do so).

The interest lies in uncovering the coefficients $\boldsymbol{\beta}(U)$ characterizing the conditional distribution $Y^*|\mathbf{X}$, which includes all individuals with characteristics $\mathbf{X}$, although not all of these individuals actually produce observable outcomes Y. For example, the $\boldsymbol{\beta}(U)$ describe the pay structure for women with characteristics $\mathbf{X}$, although not all of these women actually take part in the labor market. To establish a link between the quantiles of the distribution of observable outcomes Y and those of the distribution of potential outcomes Y^*, Arellano and Bonhomme (2017) observed that

$$\begin{aligned} P\left\{Y^* \leq \mathbf{X}'\boldsymbol{\beta}(\tau)|D=1,\mathbf{Z}=\mathbf{z}\right\} &= P\left\{U \leq \tau|V \leq p(\mathbf{z}),\mathbf{Z}=\mathbf{z}\right\} \\ &= \frac{C_{U,V|\mathbf{X}=\mathbf{x}}\{\tau,p(\mathbf{z})\}}{p(\mathbf{z})} := G_{\mathbf{x}}\{\tau,p(\mathbf{z})\} \end{aligned} \tag{4}$$

where $C_{U,V|\mathbf{X}=\mathbf{x}}\{\tau,p(\mathbf{z})\}/p(\mathbf{z}) = G_{\mathbf{x}}\{\tau,p(\mathbf{z})\}$ is the conditional copula of U and V, that is, the probability for an individual with characteristics $\mathbf{Z}=\mathbf{z}$ to have at most ranks (U,V) conditional on having at most rank $p(\mathbf{z})$ in the propensity to get selected. Because the left-hand side of (4) describes outcome quantiles in the selected population, this means that the coefficients $\boldsymbol{\beta}(\tau)$ belonging to the τth quantile in the overall population can be recovered by looking at the $G_{\mathbf{x}}\{\tau,p(\mathbf{z})\}$-quantile observations of the selected population. This establishes the validity of the following "rotated" quantile regression, which uses the observed outcomes Y but applies to them individual specific ranks $G_{\mathbf{x}}\{\tau,p(\mathbf{z})\}$ (instead of the target rank τ).

2.2 Estimation

Based on an independent and identically distributed sample $(Y_i,D_i,\mathbf{Z}_i)$ (with $i=1,\ldots,N$ and $\mathbf{X}_i \subset \mathbf{Z}_i$), Arellano and Bonhomme's (2017) estimation method proceeds as follows. For practical implementation, one assumes that the true copula $C_{U,V|\mathbf{X}=\mathbf{x}}(u,v)$ belongs to a parametric family with parameter ρ (such as Gaussian or Frank; see below) and that it does not depend on $\mathbf{X}$. The latter restriction can be relaxed by carrying out estimations by subgroup (see below). The resulting conditional copula function is denoted by $G(u,v;\rho)$. For the following, define $a^+ = \max(a,0)$ and $a^- = \max(-a,0)$.

Propensity score estimation (step 1)

$$\widehat{\boldsymbol{\gamma}} = \underset{\mathbf{a}\in\mathcal{A}}{\operatorname{argmin}} \sum_{i=1}^{n} D_i \ln\Phi\left(\mathbf{Z}_i'\mathbf{a}\right) + (1-D_i)\ln\Phi\left(-\mathbf{Z}_i'\mathbf{a}\right) \tag{5}$$

Estimation of the copula parameter (step 2)

$$\widehat{\rho} = \underset{r\in\mathcal{R}}{\operatorname{argmin}} \left\| \sum_{i=1}^{N}\sum_{l=1}^{L} \Big(D_i\, \varphi(\mathbf{Z}_i) \Big[1\Big\{ Y_i < \mathbf{X}_i'\widehat{\boldsymbol{\beta}}(\tau_l, r) \Big\} - G\left\{ \tau_l, \Phi\left(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}\right); r \right\} \Big] \Big) \right\| \tag{6}$$

$$\widehat{\boldsymbol{\beta}}(\tau_l, r) = \underset{\mathbf{b}(\tau)\in\mathcal{B}}{\operatorname{argmin}} \sum_{i=1}^{N} D_i \Big[G\left\{ \tau_l, \Phi\left(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}\right); r \right\} \left\{ Y_i - \mathbf{X}_i'\mathbf{b}(\tau) \right\}^{+} + \left[1 - G\left\{ \tau_l, \Phi\left(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}\right); r \right\} \right] \left\{ Y_i - \mathbf{X}_i'\mathbf{b}(\tau) \right\}^{-} \Big] \tag{7}$$

Rotated quantile regression (step 3)

$$\widehat{\boldsymbol{\beta}}(\tau) = \underset{\mathbf{b}(\tau)\in\mathcal{B}}{\operatorname{argmin}} \sum_{i=1}^{N} D_i \left[\widehat{G}_{\tau,i} \left\{ Y_i - \mathbf{X}_i'\mathbf{b}(\tau) \right\}^{+} + \left(1 - \widehat{G}_{\tau,i} \right) \left\{ Y_i - \mathbf{X}_i'\mathbf{b}(\tau) \right\}^{-} \right] \tag{8}$$

Step 1 estimates the probit parameters of the propensity score. Step 2 is a generalized method of moments estimating equation for the copula parameter ρ, which is identified by the conditional moment condition

$$E\Big[1\{Y_i < \mathbf{X}_i'\boldsymbol{\beta}(\tau)\} - G\left\{ \tau, \Phi\left(\mathbf{Z}_i'\boldsymbol{\gamma}\right); \rho \right\} \Big| D = 1, \mathbf{Z} = \mathbf{z} \Big] = 0$$

[following from (4)]. As an instrumental variable for estimating ρ in (6), one can use a suitable function $\varphi(\mathbf{Z}_i)$ of $\mathbf{Z}_i$, for example, $\varphi(\mathbf{Z}_i) = \Phi(\mathbf{Z}_i'\boldsymbol{\gamma})$. Minimization in (6) is carried out over a grid of candidate values for the copula parameter $r \in \mathcal{R}$. For each candidate value r, the estimated $\widehat{\boldsymbol{\beta}}(\tau_l, r)$ in (6) are obtained by rotated quantile regression over a grid of auxiliary quantiles $\tau_1, \ldots, \tau_L$ [see (7)]. Step 3 estimates, for any desired τ, selection-corrected quantile regressions based on the preestimated copula parameter $\widehat{\rho}$. For this, individual-specific rotated ranks $\widehat{G}_{\tau,i} = G\{\tau, \Phi(\mathbf{Z}_i'\widehat{\boldsymbol{\gamma}}); \widehat{\rho}\}$ are used. This can be seen by comparing (8) with the (infeasible) quantile regression, which would be carried out if potential outcomes Y^* were observed for the whole population (that is, if there was no selection problem),

$$\widetilde{\boldsymbol{\beta}}(\tau) = \underset{\mathbf{b}\in\mathcal{B}}{\operatorname{argmin}} \sum_{i=1}^{N} D_i \left\{ \tau \left(Y_i^* - \mathbf{X}_i'\mathbf{b} \right)^{+} + (1-\tau) \left(Y_i^* - \mathbf{X}_i'\mathbf{b} \right)^{-} \right\} \tag{9}$$

Note that, if one is interested only in $\boldsymbol{\beta}(\tau_1), \ldots, \boldsymbol{\beta}(\tau_L)$, step 3 is not necessary, because these are already estimated in step 2. However, for computational reasons and for reasons of flexibility, it may be useful to separate steps 2 and 3. For example, one may already have obtained individual specific copula estimates $\widehat{G}_{\tau,i}$ (for example, by subgroup estimation) and then carried out the desired rotated quantile regressions of step 3 conditional on these preestimated quantities (also see below).

2.3 Inference

Arellano and Bonhomme (2017) showed that the estimators defined in (6), (7), and (8) are asymptotically normal. However, the resulting form of the asymptotic variance matrix is very complex. This makes the use of resampling techniques attractive. In their empirical application, Arellano and Bonhomme (2017) used subsampling (Politis, Romano, and Wolf 1999). An alternative is the bootstrap (for example, Shao and Tu [1995]). The bootstrap draws independent and identically distributed resamples of size N from the original sample and repeats the estimation for several bootstrap replications. The empirical distribution of the bootstrap replications then serves as an estimate of the asymptotic distribution. Subsampling draws subsamples of size $m < N$ without replacement from the original sample and repeats the estimation on the subsamples to obtain an estimate of the asymptotic distribution (after rescaling by m/N). A related method is the m-out-of-n bootstrap, which also draws subsamples of size $m < N$ from the original sample but with replacement. Subsampling and the m-out-of-n bootstrap require that $N \to \infty$ and $m/N \to 0$; that is, the subsamples are required to be small in relation to the sample size N.[1]

Subsampling and the m-out-of-n bootstrap work under more general conditions than the bootstrap. In particular, they do not require that the asymptotic distribution be normal. It suffices that a suitably normalized version of the estimator has a limit distribution (Politis, Romano, and Wolf 1999). The bootstrap is guaranteed to work if the limit distribution is normal (Shao and Tu 1995). In the given case, both methods will work because the limit distribution is known to be normal. Subsampling and the m-out-of-n bootstrap are attractive for computational reasons if the sample size is very large because estimations have to be repeated on smaller portions of the data only. However, subsampling and the m-out-of-n bootstrap have to deal with the difficult issue of determining the subsample size (for example, Politis, Romano, and Wolf [1999]; Chernozhukov and Fernández-Val [2005]; Bickel and Sakov [2008]). Based on Chernozhukov and Fernández-Val (2005), Arellano and Bonhomme (2017) used a subsample size of a constant plus the square root of the sample size, where the constant is chosen such that the subsamples are large enough to ensure a reasonable finite sample performance of the estimator (Arellano and Bonhomme 2017, footnote 19).

Our version of Arellano and Bonhomme's (2017) estimator implements the m-out-of-n bootstrap as well as the conventional bootstrap.

2.4 Algorithms

It is well known that quantile regression problems such as (9) can be solved using linear programming techniques. However, the rotated versions (7) and (8) cannot be handled with standard implementations of quantile regression such as `qreg`, because these do not allow for individual specific ranks $\widehat{G}_{\tau,i}$. An exception are the codes of Morillo, Koenker,

1. As evident from their MATLAB codes, Arellano and Bonhomme (2017) actually use the m-out-of-n bootstrap but call it subsampling. In view of the requirement $m/N \to 0$, the numerical difference between subsampling and m-out-of-n bootstrap is typically small.

and Eilers (available at http://www.econ.uiuc.edu/~roger/research/rq/rq.m), also used by Arellano and Bonhomme (2017). For our implementation of Arellano and Bonhomme (2017), we translated these codes from MATLAB to Mata. The codes are based on an interior point algorithm (Koenker 2005) as opposed to the exterior point algorithm used in the current version of `qreg`. Our experience was that the interior point algorithm converged considerably faster than that used in `qreg` for most of the datasets analyzed by us. Our implementation also includes safeguards against problems related to using copula values too close to the boundary cases of the counter and the comonotonicity copula (in this case, $\widehat{G}_{\tau,i} \to 0$ or $\widehat{G}_{\tau,i} \to 1$; see the ado-file).

Our implementation allows sampling weights (as used in the empirical application of Arellano and Bonhomme [2018], which we replicate in section 4.2). If sampling weights are specified, we premultiply observations $Y_i, \mathbf{X}_i$, and $\varphi(\mathbf{Z}_i)$ with the sampling weight of observation i before carrying out all calculations. This ensures that the sums over $i = 1, \ldots, N$ in (6) to (8) are weighted sums. In addition, we include weights in (5).

2.5 Copula functions

Our implementation allows the user to choose among four copula functions, as shown in table 1 (for an overview of copula functions and their properties, see Joe [2015]). An important feature of all of these copulas is that they contain as limit cases the extreme forms of positive (or negative) dependence described by the comonotonicity (countermonotonicity) copula. Not restricting the strength of dependence between U and V (and therefore the strength of selection) appears to be important to avoid imposing restrictions that are not compatible with the data. Out of the four copulas listed in table 1, the Frank, Gaussian, and Plackett copulas can represent only symmetrical patterns, while the Joe and Ma (2000) copula can also accommodate asymmetrical patterns (Joe 2015).

Table 1. Copula functions $C(u, v; \rho)$

Frank copula	
$-\frac{1}{\rho}\log\left\{1+\frac{(\mathrm{e}^{-\rho u}-1)(\mathrm{e}^{-\rho v}-1)}{\mathrm{e}^{-\rho}-1}\right\}$	$\forall\ \rho \in \mathbb{R}$
Gaussian copula	
$\Phi_2\left\{\Phi^{-1}(u), \Phi^{-1}(v); \rho\right\}$	$\forall\ \rho \in (-1, 1)$
Plackett copula	
$\frac{1}{2(\rho-1)}\Big(1+(\rho-1)(u+v)-\Big[\{1+(\rho-1)(u+v)\}^2 -4(\rho-1)uv\Big]^{\frac{1}{2}}\Big)$	$\forall\ \rho \in (0, \infty)$
Joe and Ma (2000) copula	
$1-F_\Gamma\left(\left[\left\{F_\Gamma^{-1}(1-u;\rho)\right\}^\rho+\left\{F_\Gamma^{-1}(1-v;\rho)\right\}^\rho\right]^{\frac{1}{\rho}};\rho\right)$	$\forall\ \rho \in (0, \infty)$

SOURCE: Joe (2015). $F_\Gamma(\cdot, a)$ is the cumulative Gamma distribution with shape parameter a.

Because the value of the copula parameter ρ typically has no direct interpretation, our estimation command reports standard measures of bivariate concordance as listed in table 2. These represent generalized measures of correlation between U and V, which are a function of the copula and the copula parameter (Joe 2015). The concordance measures describe the association between the rank in the latent outcome distribution U and that in the distribution of resistance toward selection V. For example, if high values of U are associated with low values of V, then individuals who get selected tend to have higher outcomes than those who do not (positive selection). Note that positive (or negative) selection will be represented by negative (or positive) concordance measures because of the definition of V as the *resistance* toward selection.

Table 2. Bivariate concordance measures

Spearman's rank correlation
$\rho_{\mathrm{S}} = 12 \int_0^1 \int_0^1 uv \, dC(u, v; \rho) - 3$
Kendall's tau
$\tau_{\mathrm{K}} = P\{(U_1 - U_1')(V_2 - V_2') > 0\} - P\{(U_1 - U_1')(V_2 - V_2') < 0\}$ $= \int_0^1 \int_0^1 C(u, v; \rho) \; dC(u, v; \rho)$
Blomqvist's beta
$\beta_{\mathrm{Bl}} = P\left\{\left(U - \frac{1}{2}\right)\left(V - \frac{1}{2}\right) > 0\right\} - P\left\{\left(U - \frac{1}{2}\right)\left(V - \frac{1}{2}\right) < 0\right\}$ $= 4C\left(\frac{1}{2}, \frac{1}{2}; \rho\right) - 1$

SOURCE: Joe (2015).

The interpretation of the different concordance measures is as follows. Spearman's rank correlation measures the (ordinary) correlation between the ranks U and V. Kendall's tau is positive if it is more likely that ranks go into the same rather than into opposite directions, and it is negative otherwise. Blomqvist's beta is positive if it is more likely that both ranks U and V lie on the same side of the median rank (which is one half) than on opposite sides.

3 The arhomme command

3.1 Syntax

`arhomme` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] ,
`select(`[*depvar*$_s$] [`=`] *varlist*$_s$`)` [`rhopoints(#) taupoints(#) meshsize(#) centergrid(#) frank gaussian plackett joema nostderrors subsample(#) repetitions(#) fillfraction(#) instrument(`*varname*`) copulaparameter(`*varname*`) quantiles(# [ # [ # ... ] ]) graph output(`[`normal`] [`bootstrap`]`)`]

`pweight`s are allowed; see [U] **11.1.6 weight**.

`arhomme` is byable.

3.2 Options

3.2.1 Selection

`select(`[*depvar*$_s$] [`=`] *varlist*$_s$`)` specifies the variables and options for the selection equation. It is an integral part of specifying the Arellano and Bonhomme (2017) model and is required. The selection equation must contain at least one variable that is not in the outcome equation.

If *depvar*$_s$ is specified, it should be coded as 0 or 1, with 0 indicating an observation not selected and 1 indicating a selected observation. If *depvar*$_s$ is not specified, observations for which *depvar* is not missing are assumed selected and those for which *depvar* is missing are assumed not selected.

3.2.2 Grid tuning

`rhopoints(#)` determines the number of candidate points for the copula parameter grid search. The default is `rhopoints(19)`. When the option `frank` is chosen, the copula candidate values are constructed as follows. First, the unit interval is divided into $(\# + 1)$ equidistant intervals. Then, the ith candidate is defined as the ith quantile of a Cauchy distribution with scale `meshsize()` and shift `centergrid()`. With the option `gaussian`, the quantiles of a sinus density with emphasis `centergrid()` and range `meshsize()` $\times (1 - |$`centergrid()`$|)$ are built. The grid for the copula options `plackett` and `joema` is designed as the square root of the ith unit interval point divided by 1 minus this point. This method ensures that the resulting grid is denser around `centergrid()`. The user can shift the focus of the grid search by specifying the desired `centergrid()`, by reducing (or increasing) `meshsize()`, or by increasing (or reducing) `rhopoints()`. Note that the default `rhopoints(19)` is likely to be too small for many applications.

`taupoints(#)` specifies the number of quantiles for which the moment restriction is supposed to hold (step 2 in Arellano and Bonhomme [2017]). We recommend using this option with `graph`. The resulting scatterplot should suggest a smooth objective function (at least around the gravity center of search; the objective function may look erratic toward outer values no matter how many `taupoints()` are used). Increase `taupoints()` to further smooth the objective function. The default `taupoints(3)` is a good start in many applications, but many `taupoints()` are recommended for more reliable estimates.

`meshsize(#)` scales the grid search interval up (or down). For large #, the resulting grid becomes less dense but searches a wider range. # is restricted to strictly positive real values for the options `frank`, `plackett`, and `joema` and is restricted to $(0, 1]$ when using `gaussian`. The default `meshsize(1)` tends to be a good start.

`centergrid(#)` sets the gravity center of the grid. If you already suspect the optimal copula parameter to be a specific value, this option helps shift the emphasis of your search. # is restricted to $(-1, 1)$ with `gaussian` and to $(0, \infty)$ for `plackett` and `joema`, and it is unrestricted with `frank`. By default, the grid will always be symmetric about the independence copula, that is, `centergrid(0)` for `frank` and `gaussian`, and `centergrid(1)` for `plackett` and `joema`.

`frank` specifies the Frank copula to model individually rotated quantiles. The copula parameter is $\rho \in \mathbb{R}$, with $\rho \to -\infty$ corresponding to the lower Fréchet–Hoeffding bound, $\rho = 0$ to the independence copula, and $\rho \to \infty$ to the upper Fréchet–Hoeffding bound.

`gaussian` specifies the Gaussian copula used to model individually rotated quantiles. The copula parameter is $\rho \in (-1, 1)$, with $\rho \to -1$ corresponding to the lower Fréchet–Hoeffding bound, $\rho = 0$ to the independence copula, and $\rho \to 1$ to the upper Fréchet–Hoeffding bound.

`plackett` specifies the Plackett copula used to model individually rotated quantiles. The copula parameter is $\rho \in (0, \infty)$, with $\rho \to 0$ corresponding to the lower Fréchet–Hoeffding bound, $\rho = 1$ to the independence copula, and $\rho \to \infty$ to the upper Fréchet–Hoeffding bound. If standard errors are computed, the copula parameter is tested for $\rho = 1$ instead of $\rho = 0$. The p-value is reported accordingly.

`joema` specifies the Joe and Ma (2000) copula to model individually rotated quantiles. The copula parameter is $\rho \in (0, \infty)$, with $\rho \to 0$ corresponding to the lower Fréchet–Hoeffding bound, $\rho = 1$ to the independence copula, and $\rho \to \infty$ to the upper Fréchet–Hoeffding bound. If standard errors are computed, the copula parameter is tested for $\rho = 1$ instead of $\rho = 0$. The p-value is reported accordingly.

3.2.3 Standard errors/subsampling

`nostderrors` disables the computation of standard errors. This option precludes the use of `subsample(#)` and `repetitions(#)`.

`subsample(#)` draws samples of size # with replacement from the marked dataset. Standard errors are computed by the m-out-of-n bootstrap method. If # is greater than or equal to the effective size of the entire dataset, the conventional bootstrap is executed.

`repetitions(#)` specifies the number of bootstrap replications to be used to obtain an estimate of the variance–covariance matrix of the estimators. The default is `repetitions(100)`, which is likely to be too small in many applications.

`fillfraction(#)` determines up to which fraction of overall bootstrap repetitions the program replaces subsamples in case of failed convergence. If this limit is reached, further failed subsamples are dropped without being replaced. The default is `fillfraction(.3)`.

3.2.4 Instrument/copula parameter

`instrument(`*varname*`)` lets the user define a variable in the dataset that serves as the instrument to estimate the copula parameter [see (6)]. The instrument has to be a function of $varlist_s$. The default is the propensity score.

`copulaparameter(`*varname*`)` indicates that the copula parameter has already been estimated by the user (for example, separately by sample subgroups in a first stage) and stored per observation in the variable *varname*. In this case, only step 3 of Arellano and Bonhomme (2017) is performed (estimation of the selection-corrected quantile coefficients). The values in *varname* are restricted to $(-1, 1)$ with the option `gaussian` and to the positive real line for `plackett` and `joema`. They are unrestricted for `frank`. This option must be used in connection with `nostderrors` and precludes the use of `rhopoints()`, `taupoints()`, `meshsize()`, `centergrid()`, `subsample()`, `repetitions()`, `instrument()`, and `graph`. The reason is that the user will have to code his or her own bootstrap procedure, including all the different stages of his or her estimations (for example, using `bootstrap`). It is only in this way that the sampling variability of the preestimated copula parameters is accounted for.

3.2.5 Reporting

`quantiles(`# [# [# ...]]`)` specifies the quantiles to be estimated. Valid inputs range from 0 to 1, exclusively, and in ascending order. The default is `quantiles(0.1(0.1)0.9)` corresponding to all deciles.

`graph` specifies that a scatterplot of all objective function values be automatically generated after estimation.

`output(`[`normal`] [`bootstrap`]`)` defines whether the output table generated is based on the asymptotic, that is, normal, or the bootstrap distribution. If both are specified, two separate output tables are produced. The first stage (probit) standard errors in the output are always asymptotic (coming from the default `probit` command). `repetitions()` should always be set to at least 500 when choosing

`output(bootstrap)`. If both `normal` and `bootstrap` are specified, then results based on the normal distribution are reported first.

4 Empirical examples

4.1 Comparison with Heckman selection model

Our first empirical illustration uses the example in the Stata manual for the command `heckman`, which fits the Heckman (1979) selection model for the mean outcome.

The result of using `heckman` is

```
. webuse womenwk
. global X educ age
. global B married children
. heckman wage $X, select($X $B)
Iteration 0:   log likelihood = -5178.7009
Iteration 1:   log likelihood = -5178.3049
Iteration 2:   log likelihood = -5178.3045
Heckman selection model                         Number of obs     =      2,000
(regression model with sample selection)              Selected    =      1,343
                                                      Nonselected =        657
                                                Wald chi2(2)      =     508.44
Log likelihood = -5178.304                      Prob > chi2       =     0.0000
```

wage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
wage						
education	.9899537	.0532565	18.59	0.000	.8855729	1.094334
age	.2131294	.0206031	10.34	0.000	.1727481	.2535108
_cons	.4857752	1.077037	0.45	0.652	-1.625179	2.59673
select						
education	.0557318	.0107349	5.19	0.000	.0346917	.0767718
age	.0365098	.0041533	8.79	0.000	.0283694	.0446502
married	.4451721	.0673954	6.61	0.000	.3130794	.5772647
children	.4387068	.0277828	15.79	0.000	.3842534	.4931601
_cons	-2.491015	.1893402	-13.16	0.000	-2.862115	-2.119915
/athrho	.8742086	.1014225	8.62	0.000	.6754241	1.072993
/lnsigma	1.792559	.027598	64.95	0.000	1.738468	1.84665
rho	.7035061	.0512264			.5885365	.7905862
sigma	6.004797	.1657202			5.68862	6.338548
lambda	4.224412	.3992265			3.441942	5.006881

```
LR test of indep. eqns. (rho = 0): chi2(1) = 61.20        Prob > chi2 = 0.0000
```

We then use `arhomme` to fit a selectivity-corrected regression model for the median. For doing this, we also illustrate a useful stepwise procedure to arrive at a reasonable choice for the grid used to estimate the copula parameter.

A first step is

```
. arhomme wage $X, select($X $B) nostderrors gaussian quantiles(.5)
First step estimation (probit model) successfully completed.
Second step (gaussian copula parameter estimation) successfully completed.
Found objective function minimum 1.705e-05 for rho = -0.5903
Third step (minimization of rotated check function) successfully completed.

Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)

                                                Number of obs.   =      2,000
                                                Num. of selected =      1,343
                                                Rho points       =         19
                                                Tau points       =          3
                                                Meshsize         =     1.0000
                                                Spearman's rho   =    -0.5723
                                                Kendall's tau    =    -0.4020
                                                Blomqvist's beta =    -0.4020
                                                Minimum Fval     =  1.705e-05
------------------------------------------------------------------------------
        wage | Coefficient
-------------+----------------------------------------------------------------
select       |
   education |   .0583645
         age |   .0347211
     married |   .4308575
    children |   .4473249
       _cons |  -2.467365
-------------+----------------------------------------------------------------
.5_quantile  |
       _cons |   1.487906
   education |    .992114
         age |   .1923601
-------------+----------------------------------------------------------------
_anc         |
         rho |  -.5903345
------------------------------------------------------------------------------
note: parameter estimates based on Gaussian copula model
. local c = e(rho)
```

We then use a refined grid:

```
. arhomme wage $X, select($X $B) nostderrors gaussian quantiles(.5)
> taupoints(7) rhopoints(25) centergrid(`c') meshsize(.2)
First step estimation (probit model) successfully completed.
  (output omitted)
------------------------------------------------------------------------------
Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)
------------------------------------------------------------------------------
                                                Number of obs.   =      2,000
                                                Num. of selected =      1,343
                                                Rho points       =         25
                                                Tau points       =          7
                                                Meshsize         =     0.2000
                                                Spearman's rho   =    -0.6339
                                                Kendall's tau    =    -0.4519
                                                Blomqvist's beta =    -0.4519
                                                Minimum Fval     =  8.993e-07
------------------------------------------------------------------------------
                 |
  (output omitted)
-----------------+------------------------------------------------------------
.5_quantile      |
           _cons |   .5695862
       education |   1.016767
             age |    .203274
-----------------+------------------------------------------------------------
_anc             |
             rho |  -.6516752
------------------------------------------------------------------------------
note: parameter estimates based on Gaussian copula model
. local c = e(rho)
```

So far, we have used the option `nostd` to save computing time. In the next step, we also include the computation of standard errors:

```
. set seed 1337
. arhomme wage $X, select($X $B) gaussian quantiles(.5) taupoints(7)
> centergrid(`c') repetitions(250)
option subsample left unspecified: subsample automatically set to 2000
(bootstrap)
use option nostderrors to disable estimation of covariance matrix
First step estimation (probit model) successfully completed.
Second step (gaussian copula parameter estimation) successfully completed.
Found objective function minimum 8.993e-07 for rho = -0.6517
Third step (minimization of rotated check function) successfully completed.
Initialising standard error estimation by 2000 out of 2000 bootstrap method:
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................     50
..................................................    100
  (output omitted)
..................................................    250

------------------------------------------------------------------------------
Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)
------------------------------------------------------------------------------
                                                Number of obs.   =      2,000
                                                Num. of selected =      1,343
                                                Rho points       =         19
                                                Tau points       =          7
                                                Meshsize         =     1.0000
                                                Spearman's rho   =    -0.6339
                                                Kendall's tau    =    -0.4519
                                                Blomqvist's beta =    -0.4519
                                                Minimum Fval     =  8.993e-07
                                                Replications     =        250
                                                Subsample Size   =      2,000
------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
select       |
   education |   .0583645   .0111586     5.23   0.000      .036494    .0802351
  (output omitted)
-------------+----------------------------------------------------------------
.5_quantile  |
       _cons |   .5695862   1.387861     0.41   0.682    -2.150571    3.289744
   education |   1.016767   .0756374    13.44   0.000     .8685207    1.165014
         age |    .203274   .0259474     7.83   0.000     .1524181    .2541299
-------------+----------------------------------------------------------------
_anc         |
         rho |  -.6516752   .0764418    -8.53   0.000    -.8014983    -.501852
------------------------------------------------------------------------------
note: parameter estimates based on Gaussian copula model
```

Both `heckman` and `arhomme` find substantial positive selection (recall that a negative copula parameter in `arhomme` represents positive selection). The coefficients for `education` and `age` in `arhomme` for the median wage are quite similar to those for the

mean wage in `heckman`. This will be the likely outcome if the conditional distributions are symmetric (implying that the mean is equal to the median).

Finally, we illustrate the postestimation features of `arhomme`:

```
. test [.5_quantile]education = [.5_quantile]age
 ( 1)  [.5_quantile]education - [.5_quantile]age = 0
           chi2(  1) =   93.17
         Prob > chi2 =    0.0000
. predict medpred, equation(.5_quantile)
```

4.2 Replication of Arellano and Bonhomme (2018)

This section replicates the empirical example in Arellano and Bonhomme (2018). The data are from Huber and Melly (2015) and can be downloaded from the *Journal of Applied Econometrics* data archive (http://qed.econ.queensu.ca/jae/2015-v30.7/huber-melly/). The application refers to the returns to education and experience for women in the United States using data from the 2011 Current Population Survey. The sample covers white non-Hispanic women aged between 25 and 54 years. Individuals who are self-employed or work for the military, public, or agricultural sector are excluded. Working is defined as having worked for more than 35 hours in the week preceding the survey. The application uses the Current Population Survey sampling weights.

We first load the data and modify some of the variable names to make them conform to those used in Arellano and Bonhomme (2018):

```
. use application, clear
. rename hsg educ_7
. rename some_college educ_8
. rename associate educ_9
. rename college educ_11
. rename advanced educ_13
. rename mw midwest
. rename So south
. rename We west
```

We now apply `arhomme`, following as closely as possible the specification in Arellano and Bonhomme (2018):

```
. global X educ_7 educ_8 educ_9 educ_11 educ_13 exp exp2 exp_edu exp2_edu
> midwest south west married
. global B child02 child35 child613 child02_m child35_m child613_m
. set seed 1337
. arhomme lwage $X [pw=wgt], select(ft = $X $B) taupoints(4) rhopoints(39)
> gaussian subsample(1000) repetitions(500) quantiles(.25 .5 .75)
> centergrid(-.0989229)
```

(*output omitted*)

```
Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)
```

Number of obs.	44,562
Num. of selected	20,055
Rho points	39
Tau points	4
Meshsize	1.0000
Spearman's rho	-0.0945
Kendall's tau	-0.0631
Blomqvist's beta	-0.0631
Minimum Fval	1.473e-08
Replications	500
Subsample Size	1,000

lwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ft						
educ_7	.5869417	.0428666	13.69	0.000	.5029247	.6709586
educ_8	.073392	.0226713	3.24	0.001	.0289571	.1178269
educ_9	.2325266	.0261318	8.90	0.000	.1813092	.2837441
educ_11	.0598427	.0287012	2.09	0.037	.0035894	.1160959
educ_13	.1910608	.0310857	6.15	0.000	.130134	.2519876
exp	.0036565	.0044806	0.82	0.414	-.0051252	.0124382
exp2	-.0003162	.0001053	-3.00	0.003	-.0005225	-.0001098
exp_edu	-.0020776	.0008418	-2.47	0.014	-.0037275	-.0004278
exp2_edu	.0000259	.0000201	1.29	0.197	-.0000134	.0000652
midwest	.0948917	.0208117	4.56	0.000	.0541015	.135682
south	.0542725	.0199717	2.72	0.007	.0151286	.0934163
west	-.0596079	.0214657	-2.78	0.005	-.1016799	-.017536
married	-.1490132	.0186358	-8.00	0.000	-.1855387	-.1124877
child02	-.3361118	.0471503	-7.13	0.000	-.4285247	-.2436989
child35	-.1825115	.0374979	-4.87	0.000	-.256006	-.109017
child613	-.1016576	.0209263	-4.86	0.000	-.1426724	-.0606428
child02_m	-.0714047	.0510075	-1.40	0.162	-.1713775	.0285681
child35_m	-.0990061	.0411064	-2.41	0.016	-.1795732	-.0184391
child613_m	-.101258	.0232153	-4.36	0.000	-.1467592	-.0557568
_cons	-.5040045	.0760898	-6.62	0.000	-.6531377	-.3548713

.25_quantile						
_cons	1.95851	.0885078	22.13	0.000	1.785038	2.131982
educ_7	.2042428	.0712498	2.87	0.004	.0645957	.3438898
educ_8	.1047957	.0181724	5.77	0.000	.0691785	.140413
educ_9	.0759379	.0212158	3.58	0.000	.0343558	.11752
educ_11	.2806021	.0230952	12.15	0.000	.2353364	.3258678
educ_13	.1891199	.0260612	7.26	0.000	.1380409	.240199
exp	.0163292	.0031733	5.15	0.000	.0101096	.0225488
exp2	-.0002454	.0000769	-3.19	0.001	-.0003962	-.0000946
exp_edu	.0011344	.0008901	1.27	0.203	-.0006102	.0028789
exp2_edu	-.000021	.0000233	-0.90	0.368	-.0000666	.0000246
midwest	-.0753589	.0185544	-4.06	0.000	-.1117249	-.038993
south	-.1066912	.0180001	-5.93	0.000	-.1419708	-.0714117
west	-.0272446	.0190795	-1.43	0.153	-.0646398	.0101505
married	.0232687	.0130662	1.78	0.075	-.0023407	.048878
.5_quantile						
_cons	2.018633	.053574	37.68	0.000	1.91363	2.123636
educ_7	.3126759	.0308411	10.14	0.000	.2522284	.3731234
educ_8	.1042804	.0150589	6.92	0.000	.0747656	.1337953
educ_9	.098068	.0191286	5.13	0.000	.0605766	.1355595
educ_11	.2915985	.0205943	14.16	0.000	.2512344	.3319625
educ_13	.2039454	.0225455	9.05	0.000	.1597571	.2481337
exp	.0255916	.0027972	9.15	0.000	.0201092	.031074
exp2	-.0003969	.0000667	-5.95	0.000	-.0005276	-.0002663
exp_edu	.0019973	.0006724	2.97	0.003	.0006794	.0033153
exp2_edu	-.0000407	.000017	-2.40	0.016	-.000074	-7.44e-06
midwest	-.0907382	.0144737	-6.27	0.000	-.1191062	-.0623702
south	-.1120291	.0146407	-7.65	0.000	-.1407244	-.0833338
west	-.0219371	.0160317	-1.37	0.171	-.0533587	.0094845
married	.0235552	.0109308	2.15	0.031	.0021313	.0449791
.75_quantile						
_cons	2.249392	.0619358	36.32	0.000	2.128	2.370784
educ_7	.2992705	.0356045	8.41	0.000	.229487	.369054
educ_8	.1287676	.017044	7.55	0.000	.0953619	.1621733
educ_9	.142979	.0199721	7.16	0.000	.1038345	.1821236
educ_11	.2513494	.022425	11.21	0.000	.2073973	.2953015
educ_13	.2046976	.0235484	8.69	0.000	.1585436	.2508516
exp	.0301231	.0032344	9.31	0.000	.0237838	.0364624
exp2	-.0004632	.0000762	-6.08	0.000	-.0006126	-.0003138
exp_edu	.0032962	.000767	4.30	0.000	.0017928	.0047995
exp2_edu	-.00007	.0000195	-3.59	0.000	-.0001083	-.0000318
midwest	-.1216595	.0159755	-7.62	0.000	-.152971	-.0903481
south	-.0978834	.0164134	-5.96	0.000	-.1300529	-.0657138
west	-.0157402	.0168385	-0.93	0.350	-.048743	.0172627
married	.0210093	.0123409	1.70	0.089	-.0031784	.045197
_anc						
rho	-.0989229	.0536844	-1.84	0.065	-.2041424	.0062966

note: parameter estimates based on Gaussian copula model

The results for the selection-corrected quantile regression coefficients are almost identical to those reported by Arellano and Bonhomme (2018, table 1). Standard-error estimates are also very similar. The point estimate for the copula parameter and its standard error also come close to those reported by Arellano and Bonhomme (2018)

($\widehat{\rho} = -0.10$ with standard error 0.054). Small differences between our results and those of Arellano and Bonhomme (2018) are to be expected because standard errors are the result of a random process (subsampling), because of numerical software differences, and because we do not know the exact choices of Arellano and Bonhomme (2018) for grid search, number of supporting quantiles, etc.

4.3 Partial replication of Arellano and Bonhomme (2017)

Our last empirical example replicates selected results in Arellano and Bonhomme (2017). `wagedata.dta` can be downloaded from Stéphane Bonhomme's webpage (https://sites.google.com/site/stephanebonhommeresearch/) and refers to selectivity-corrected wage distributions for the UK for the period 1978–2000. This section illustrates the use of the `graph` option, which displays the grid search optimization for the copula parameter.

To replicate a selected quantile regression for the subgroup of single females, we first run some preparatory steps (taken from the file `sample_construction.do`, which can also be downloaded from the above webpage; details are available on request).

```
. use wagedata, clear
. * now data preparation as in sample_construction.do downloadable
. * from https://sites.google.com/site/stephanebonhommeresearch/
 (commands omitted)
. * keep females
. keep if sex==2
(90,731 observations deleted)
. * keep singles
. keep if married==0
(66,851 observations deleted)
 (commands omitted)
```

We then start with a first crude estimation attempt:

```
. global X trend1 trend2 trend3 c1919_34 c1935_44 c1955_64 c1965_77 ed17 ed18
> 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
. arhomme lw $X, select(work = $X s_zero) frank graph rhopoints(49) taupoints(4)
> quantiles(.5) nostderrors
  (output omitted)
. local c = e(rho)
. graph save "female single objective function.gph"
file female single objective function.gph saved
```

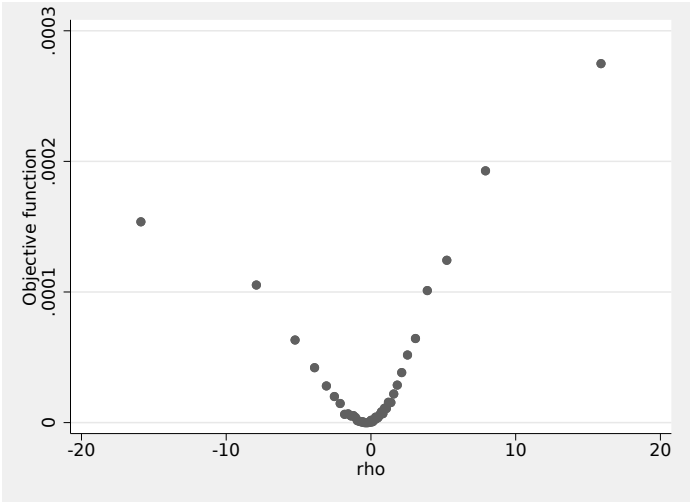

Figure 1. Plot of objective function over crude grid

Refine grid:

```
. arhomme lw $X, sel(work = $X s_zero) frank graph rhopoints(19) taupoints(7)
> quantiles(.5) centergrid(`c') meshsize(0.1) nostderrors
```

(*output omitted*)

```
. local c = e(rho)

. graph save "female single objective function magnified.gph"
file female single objective function magnified.gph saved
```

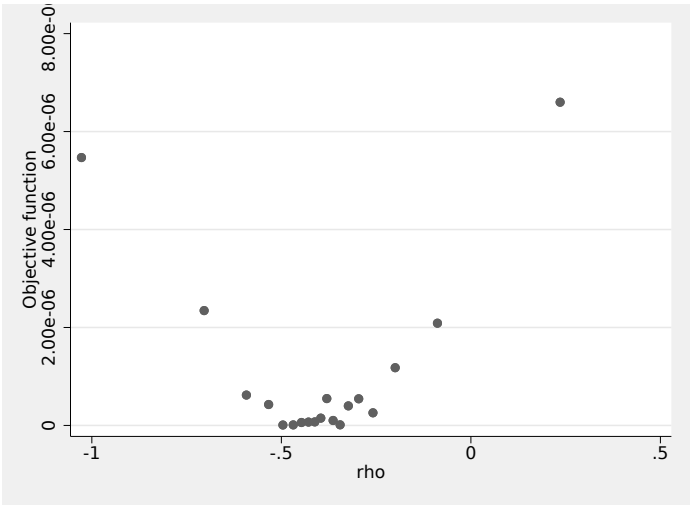

Figure 2. Plot of objective function over refined grid

Finally, we estimate the preferred specification with standard errors based on subsampling (using the same subsample size as in Arellano and Bonhomme [2017]):

```
. local s = 1000 + ceil(sqrt(_N))
. set seed 1337
. arhomme lw $X, select(work = $X s_zero) frank rhopoints(39) taupoints(7)
> quantiles(.5) centergrid(`c') repetitions(250) subsample(`s')
  (output omitted)
Initialising standard error estimation by 1154 out of 23583 bootstrap method:
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................     50
..................................................    100
..................................................    150
..................................................    200
..................................................    250
```

Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)

Number of obs.	23,583
Num. of selected	15,185
Rho points	39
Tau points	7
Meshsize	1.0000
Spearman's rho	-0.0824
Kendall's tau	-0.0550
Blomqvist's beta	-0.0618
Minimum Fval	9.002e-09
Replications	250
Subsample Size	1,154

lw	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
work						
trend1	-.1179342	.0177487	-6.64	0.000	-.1527209	-.0831474
(output omitted)						
_cons	2.300847	.1125997	20.43	0.000	2.080156	2.521539
.5_quantile						
_cons	1.405257	.0351858	39.94	0.000	1.336294	1.47422
(output omitted)						
kids_d6	-.0778446	.0321618	-2.42	0.016	-.1408805	-.0148087
_anc						
rho	-.495928	.4760549	-1.04	0.298	-1.428978	.4371224

note: parameter estimates based on Frank copula model

Arellano and Bonhomme (2017) do not document their estimated selectivity-corrected regression coefficients (they are used only in their later calculations), but they do report estimated copula parameters for different population subgroups. For the group of single females, they report an estimated Spearman rank correlation of -0.080 (Arellano and Bonhomme 2017, 16) and an estimated parameter for the Frank copula of -0.4820 (documented in replication file `Graphs_Frank_copula.m` downloadable from Stéphane

Bonhomme's webpage; see above). Up to numerical differences, this is very close to the results we obtain above.

5 Conclusion

In this article, we described a command, `arhomme`, implementing the Arellano and Bonhomme (2017) method of sample selection correction for quantile regressions along with standard errors based on bootstrapping and subsampling. `arhomme` is fast and potentially applicable in many fields in which there is a need to correct estimates of conditional distributions for sample selection. If one is interested in obtaining unconditional distributions corrected for sample selection, the resulting conditional distributions may be aggregated up as described in Albrecht, van Vuuren, and Vroman (2009) or Chernozhukov, Fernández-Val, and Melly (2013).

6 Acknowledgments

We thank Michael Wolf (University of Zürich) for discussions and advice on subsampling. We also thank Ben Jann, Blaise Melly, Philippe Van Kerm, and the participants of the Swiss Stata Conference 2020 for many helpful comments and suggestions. Financial support through the DFG Priority Program 1764 and DFG project BI 767/3-1 is gratefully acknowledged.

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 21-3
. net install st0648     (to install program files, if available)
. net get st0648         (to install ancillary files, if available)
```

8 References

Ahn, H., and J. L. Powell. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58: 3–29. https://doi.org/10.1016/0304-4076(93)90111-H.

Albrecht, J., A. van Vuuren, and S. Vroman. 2009. Counterfactual distributions with sample selection adjustments: Econometric theory and an application to the Netherlands. *Labour Economics* 16: 383–396. https://doi.org/10.1016/j.labeco.2009.01.002.

Andrews, D. W. K., and M. M. A. Schafgans. 1998. Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies* 65: 497–517. https://doi.org/10.1111/1467-937X.00055.

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.

Bickel, P. J., and A. Sakov. 2008. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* 18: 967–985.

Biewen, M., B. Fitzenberger, and M. Seckler. 2020. Counterfactual quantile decompositions with selection correction taking into account Huber/Melly (2015): An application to the German gender wage gap. *Labour Economics* 67: 101927. https://doi.org/10.1016/j.labeco.2020.101927.

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.

Buchinsky, M. 1998. The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. *Journal of Applied Econometrics* 13: 1–30. https://doi.org/10.1002/(SICI)1099-1255(199801/02)13:1<1::AID-JAE474>3.0.CO;2-A.

———. 2001. Quantile regression with sample selection: Estimating women's return to education in the U.S. *Empirical Economics* 26: 87–113. https://doi.org/10.1007/s001810000061.

Chen, S., and S. Khan. 2003. Semiparametric estimation of a heteroskedastic sample selection model. *Econometric Theory* 19: 1040–1064. https://doi.org/10.1017/S0266466603196077.

Chernozhukov, V., and I. Fernández-Val. 2005. Subsampling inference on quantile regression processes. *Sankhyā* 67: 253–276.

Chernozhukov, V., I. Fernández-Val, and B. Melly. 2013. Inference on counterfactual distributions. *Econometrica* 81: 2205–2268. https://doi.org/10.3982/ECTA10582.

Das, M., W. K. Newey, and F. Vella. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* 70: 33–58. https://doi.org/10.1111/1467-937X.00236.

D'Haultfoeuille, X., A. Maurel, X. Qiu, and Y. Zhang. 2020. Estimating selection models without an instrument with Stata. *Stata Journal* 20: 297–308. https://doi.org/10.1177/1536867X20930998.

Fernández-Val, I., A. van Vuuren, and F. Vella. 2018. Nonseparable sample selection models with censored selection rules: An application to wage decompositions. IZA Discussion Paper No. 11294, Institute of Labor Economics (IZA). http://ftp.iza.org/dp11294.pdf.

Gronau, R. 1974. Wage comparisons–A selectivity bias. *Journal of Political Economy* 82: 1119–1143. https://doi.org/10.1086/260267.

Heckman, J. J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42: 679–694. https://doi.org/10.2307/1913937.

———. 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.

Joe, H. 2015. *Dependence Modeling with Copulas*. Boca Raton, FL: CRC Press.

Joe, H., and C. Ma. 2000. Multivariate survival functions with a min-stable property. *Journal of Multivariate Analysis* 75: 13–35. https://doi.org/10.1006/jmva.1999.1891.

Koenker, R. 2005. *Quantile Regression*. New York: Cambridge University Press.

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.

Picchio, M., and C. Mussida. 2011. Gender wage gap: A semi-parametric approach with sample selection correction. *Labour Economics* 18: 564–578. https://doi.org/10.1016/j.labeco.2011.05.003.

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

Shao, J., and D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

About the authors

Martin Biewen is a professor of econometrics in the School of Business and Economics at the University of Tübingen.

Pascal Erhardt is a lecturer of statistics in the School of Business and Economics at the University of Tübingen.