working paper
1702

Sample Selection in Quantile Regression: A Survey

Manuel Arellano
Stéphane Bonhomme

January 2017

CENTRO DE ESTUDIOS
MONETARIOS Y FINANCIEROS
Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

CEMFI Working Paper No. 1702
January 2017

Sample Selection in Quantile Regression: A Survey

Abstract

Nonrandom sample selection is a pervasive issue in applied work. In additive models, a number of techniques are available for consistent selection correction. However, progress in the development of non-additive selection corrections has been slower. In this survey we review recent proposals dealing with sample selection in quantile models.

JEL Codes: C13, J31.

Keywords: Quantile regression, sample selection, copula, wage regressions.

Manuel Arellano
CEMFI
arellano@cemfi.es

Stéphane Bonhomme
University of Chicago
sbonhomme@uchicago.edu

Acknowledgement

Chapter prepared for the Handbook of Quantile Regression. We thank Roger Koenker, Blaise Melly, and participants at the December 2015 conference in Cambridge for comments. Arellano acknowledges research funding from the Ministerio de Economía y Competitividad, Grant ECO2016-79848-P.

1 Introduction

Nonrandom sample selection is a pervasive issue in applied work. Selection may arise because of data collection by the analyst. It may also be due to the fact that agents self-select into particular options, the latter form of selection being very common in economics.

A prototypical example of sample selection, due to Gronau (1974) and Heckman (1974), concerns selection into labor market participation. When interest centers on estimating determinants of wage offer functions, standard regression approaches will result in biased estimates if selection into work is not random. In this example, selection may arise due to individuals with low wage potential choosing not to participate to the labor market.

Selection biases show up in a variety of ways in the literature. A situation related to the previous example is the estimation of sector-specific wage offer functions, or more generally alternative-specific payoff functions, when alternatives are chosen in part based on some forecasts of payoffs. Other related settings are missing data problems, and nonrandom attrition in longitudinal data sets. Although references are too many to be mentioned here, recent influential studies where sample selection plays a central role are Mulligan and Rubinstein (2008), Helpman, Melitz and Rubinstein (2008), and Jimenez, Ongena, Peydro and Saurina (2014).

In linear models, Heckman (1976, 1979) proposed a method which has become very popular in empirical work. The assumptions of the Heckman model rely on Gaussianity, while allowing the errors in the outcome and participation equations to be correlated. The Heckman two-step estimator, which we briefly review below, provides a practical alternative to full maximum likelihood.

Building on the control function approach implicit in the Heckman method, a large econometric literature has since then extended the model by relaxing parametric assumptions and proposing semi-parametric estimators. Influential examples include Heckman (1990), Ahn and Powell (1993), Donald (1995), Andrews and Schafgans (1998), Chen and Khan (2003), and Das, Newey and Vella (2003). See also the references in Vella (1998). Additivity of the outcome in observed covariates and unobservables is key in all these approaches.

Much less is known regarding sample selection in nonlinear models. Manski (1994, 2003) derived worst-case bounds on quantiles of potential outcomes; see also Kitagawa (2010). Blundell, Gosling, Ichimura and Meghir (2007) applied the bounds approach to document gender differences in wage inequality in the UK. However, the literature on parametric or

semi-parametric selection corrections in nonlinear models is scarce.

In this chapter we focus on the question of correcting quantile regression estimates for nonrandom sample selection. Quantile regression (Koenker and Bassett, 1978) is a versatile estimation approach which has been extensively studied. However, little work has been done at the intersection of quantile methods and sample selection methods. We review the approach of Arellano and Bonhomme (2016). A central observation is that, in quantile models, even linear ones, quantile curves on the selected sample are generally not linear. However, a correction is available which consists in "rotating" the check function of quantile regression by an amount that is observation-specific and depends on the strength of selection.

Implementing this method requires estimating the degree of sample selection. Formally, the latter is defined as the dependence between the rank error in the equation of interest and the rank error in the selection equation. Working with a parametric copula, Arellano and Bonhomme (2016) derive moment restrictions on the copula parameter. As in linear models, "excluded" covariates affecting participation without entering the potential outcome equation are required for credible identification. The method then consists of three steps: estimation of the propensity score, estimation of the degree of selection (that is, the copula parameter), and computation of quantile estimates through rotated quantile regression.

The sample selection problem we focus on here differs from censoring. Censored quantile regression is a well-studied problem (e.g., Powell, 1986, Chamberlain, 1993, Buchinsky, 1994, Buchinsky and Hahn, 1998, Chernozhukov and Hong, 2002, Portnoy, 2003, Chernozhukov *et al.*, 2015), see also the chapter in this *Handbook* on censoring in survival analysis. It turns out that the Buchinsky and Hahn censoring correction may be interpreted as a selection correction based on a degenerate (Fréchet) copula.

We also review two other approaches to sample selection. Buchinsky (1998, 2001) proposed a control function approach to correct quantile regression estimates for sample selection. This method has been used by Albrecht, van Vuuren and Vroman (2009) and Bollinger, Ziliak and Troske (2011), among others. However, control function methods impose conditions on the data generating process which may be inconsistent with quantile models unless the model is additive and quantile curves are parallel to each other, or selection is random (Huber and Melly, 2015). Lastly, the method in Arellano and Bonhomme (2016) is only one possibility to estimate selection-corrected quantile coefficients, and we briefly review an alternative approach based on maximum likelihood.

A non-quantile regression based approach to selection correction is to parametrically specify both outcome and selection equations, thus providing non-Gaussian extensions to the Heckman model. See Lee (1983), Smith (2003), or the recent application in Van Kerm (2013) for example. Relative to fully parametric approaches, quantile regression provides added flexibility in the modeling of outcome variables.

In the final part of this chapter we revisit the empirical illustration in Huber and Melly (2015), and estimate uncorrected and selection-corrected wage returns to experience and education based on data on female wages and employment status from the 1991 Current Population Survey. Huber and Melly provided evidence of wages being non-additive in covariates and unobservables in this setting. We complement their analysis by providing selection-corrected quantile regression estimates, which remain consistent under non-additivity.

The outline of the chapter is as follows. In Sections 2 and 3 we review the approaches of Heckman (1979) and Arellano and Bonhomme (2016). In Sections 4 and 5 we discuss identification in the absence of parametric assumptions, and review several extensions. Lastly, we present the empirical illustration in Section 6, and conclude in Section 7.

2 Heckman's parametric selection model

Consider an additive latent response model of the form

$$Y^* = X'\beta + \varepsilon, \tag{1}$$

where ε has mean zero and is identically distributed given X. If a random sample from (Y^*, X) were available, β could be consistently estimated under standard conditions using ordinary least squares (OLS). Sample selection arises as Y^* is only observed when the binary selection indicator D is equal to one (hence the star superscript, which refers to Y^* being a *latent* variable). In turn, D is given by

$$D = \mathbf{1}\{\eta \leq Z'\gamma\}, \tag{2}$$

where X is a subset of Z, and $\mathbf{1}\{\cdot\}$ is an indicator function. The scalar unobservable η is independent of Z, and possibly correlated with ε. Let $Y = DY^*$. A random sample from (Z, D, Y) is available, but Y^* is not observed when $D = 0$.

A textbook example is the following (Heckman, 1974): Y^* are wage offers, X are determinants of wages (such as education and experience), and D denotes labor force participation.

Wage offers are not observed unless they have been accepted. In applications, in addition to X, Z typically contains *excluded* determinants of participation that do not appear in the wage offer equation, such as the number of children, marital status, or potential income when out of work, among others, all of which aim at capturing costs of working unrelated to potential wages. In this example, one expects dependence between η and ε if participation decisions are influenced by unobserved determinants of potential wage offers.

As latent outcomes Y^* are not observed for non-participants, it is not possible to directly estimate an empirical counterpart to $\mathbb{E}(Y^* \,|\, X)$. Instead, the conditional mean for participants, $\mathbb{E}(Y^* \,|\, D=1, Z)$, which is identified from data on participants only, is instrumental in developing a selection correction method. Following Heckman (1979), we have

$$\begin{aligned} \mathbb{E}\left(Y^* \,|\, D=1, Z\right) &= X'\beta + \mathbb{E}\left(\varepsilon \,|\, D=1, Z\right) \\ &= X'\beta + \Lambda(Z), \end{aligned} \tag{3}$$

where $\Lambda(Z) = \mathbb{E}\left(\varepsilon \,|\, \eta \leq Z'\gamma, Z\right)$ is a selection correction factor.

From (3) it follows that an OLS regression of Y on X on participants $D=1$ will generally be inconsistent for β when ε and η are statistically dependent. This case precisely corresponds to nonrandom sample selection. Note that (3) suggests a strategy to consistently estimate β, by regressing Y on a linear function of X and an additive nonlinear function of Z. In the case where distributions are multivariate Gaussian, such a strategy simplifies to the Heckman (1979) method.

Two-step estimation in Gaussian models. Let us now assume that (ε, η) is bivariate Gaussian, independent of Z, with variances σ^2 and 1, respectively, and correlation ρ. In this case

$$\Lambda(Z) = -\rho\sigma\lambda(Z'\gamma), \quad \text{with} \quad \lambda(u) = \frac{\phi(u)}{\Phi(u)},$$

where ϕ and Φ denote the standard Gaussian pdf and cdf, respectively. Note that the propensity score is $p(Z) = \Pr(D=1 \,|\, Z) = \Phi(Z'\gamma)$, so we also have $\Lambda(Z) = -\rho\sigma\lambda\left[\Phi^{-1}\left(p(Z)\right)\right]$.

Heckman (1976, 1979) proposes a two-step estimator. In the first step, γ is estimated by a probit regression of D on Z. Letting $\widehat{\gamma}$ denote the parameter estimate, the selection factor is estimated (up to scale) as $\widehat{\lambda} = \lambda\left(Z'\widehat{\gamma}\right)$. In the second step, β and $\rho\sigma$ are estimated by an OLS regression of Y on X and $\widehat{\lambda}$ in the subsample of participants $D=1$.

Formulas are available to correct the standard errors of the second-step estimator $\widehat{\beta}$ for estimation error in the first step. In the Gaussian model, this two-step "control function"

method provides an alternative to maximum likelihood, albeit at some efficiency cost (e.g., Nelson, 1984). An attractive feature of the method is that it can be extended to allow for semi- or nonparametric specifications, as reviewed in the introduction, provided additivity of (1) in X and ε is maintained. However, non-additive models such as quantile models cannot be studied using those techniques.

3 A quantile generalization

In this section we describe the approach introduced in Arellano and Bonhomme (2016).

3.1 A quantile selection model

Consider now the following linear quantile specification of outcomes

$$Y^* = X'\beta(U), \tag{4}$$

where $\beta(u)$ is increasing in u, and U is uniformly distributed on the unit interval, independent of X. Model (4) is a linear quantile model (Koenker and Bassett, 1978). In particular, $Q(\tau, X) = X'\beta(\tau)$ is the τ-th conditional quantile of Y^* given X. If a random sample from (Y^*, X) were available, one could thus consistently estimate $\beta(\tau)$ for all $\tau \in (0,1)$ by quantile regression, under standard assumptions.

Maintaining the other assumptions of the Heckman Gaussian model, we assume that (2) holds with a Gaussian η independent of Z so that, equivalently,

$$D = \mathbf{1}\{V \leq p(Z)\}, \tag{5}$$

where $p(Z) = \Phi(Z'\gamma)$, and $V = \Phi(\eta)$ is the rank of η, which is uniformly distributed on $(0,1)$ and independent of Z.

Lastly, we assume that (U, V) follows a bivariate Gaussian copula with dependence parameter ρ, independent of Z. We denote as $G(\tau, p; \rho) = C(\tau, p; \rho)/p$ the *conditional copula* of U given V, defined on $(0,1) \times (0,1)$, where $C(\tau, p; \rho)$ denotes the unconditional copula of (U, V).[1] Note that model (4)-(5) simplifies to the Heckman Gaussian model when $X'\beta(U) = X'\beta + \sigma\Phi^{-1}(U)$ is a location-shift Gaussian model. Although we consider a Gaussian copula to fix ideas, any other parametric specification could be used such as the Gumbel, Frank, or Bernstein copulas for example.

[1] Letting $\Phi_2(\cdot, \cdot; \rho)$ denote the bivariate Gaussian cdf with parameter ρ, $C(\tau, p; \rho) = \Phi_2\left(\Phi^{-1}(\tau), \Phi^{-1}(p); \rho\right)$ and $G(\tau, p; \rho) = \Phi_2\left(\Phi^{-1}(\tau), \Phi^{-1}(p); \rho\right)/p$.

In the non-additive model (4)-(5), quantile curves are generally non-additive in the propensity score $p(Z)$ and covariates X. To see this, denote $Z = (X, W)$ (where W are the "excluded" covariates), and note that the conditional cdf of Y^* given $Z = z = (x, w)$ for participants $D = 1$ is, evaluated at $x'\beta(\tau)$ for some τ in the unit interval,

$$\begin{aligned} \Pr\left(Y^* \leq x'\beta(\tau) \,|\, D = 1, Z = z\right) &= \Pr\left(U \leq \tau \,|\, V \leq \Phi\left(z'\gamma\right), Z = z\right), \\ &= G\left(\tau, \Phi\left(z'\gamma\right); \rho\right), \end{aligned} \tag{6}$$

where $G(\cdot, \cdot; \rho)$ is the conditional Gaussian copula with parameter ρ. It follows that the τ-th conditional quantile of Y^* given $D = 1$ and Z is

$$Q^s(\tau, Z) = X'\beta\left(\tau^*(Z)\right), \tag{7}$$

where $\tau^*(Z) = G^{-1}\left(\tau, \Phi\left(Z'\gamma\right); \rho\right)$, and $G^{-1}(\tau, p; \rho)$ denotes the inverse of $G(\tau, p; \rho)$ with respect to its first argument (that is, the conditional quantile function of U given $V \leq p$).[2] The "s" superscript refers to the fact that $Q^s(\tau, Z)$ is conditional on selection.

Non-additivity of quantile curves implies that existing control function strategies cannot be used in the quantile selection model. We next review a method recently proposed by Arellano and Bonhomme (2016) to achieve consistent estimation in this model.

3.2 Estimation

Let us start with the case where γ and ρ are known. Later we will show how these parameters may be consistently estimated. From (6), for every $\tau \in (0, 1)$ the parameter vector $\beta(\tau)$ is then characterized as the solution to the population moment restriction

$$\mathbb{E}\left[\mathbf{1}\left\{Y \leq X'\beta(\tau)\right\} - G\left(\tau, \Phi\left(Z'\gamma\right); \rho\right) \mid D = 1, Z\right] = 0. \tag{8}$$

Hence, using DX as instruments and taking expectations,

$$\mathbb{E}\left[DX\left(\mathbf{1}\left\{Y \leq X'\beta(\tau)\right\} - G\left(\tau, \Phi\left(Z'\gamma\right); \rho\right)\right)\right] = 0. \tag{9}$$

Arellano and Bonhomme (2016) noticed that (9) is the system of first-order conditions in the following optimization

$$\beta(\tau) = \underset{b(\tau)}{\operatorname{argmin}}\ \mathbb{E}\left[D\left(G_{\tau Z}\left(Y - X'b(\tau)\right)^+ + \left(1 - G_{\tau Z}\right)\left(Y - X'b(\tau)\right)^-\right)\right], \tag{10}$$

[2]The assumption that G is strictly monotone in its first argument is not without loss of generality. For example, it is not satisfied by Fréchet copulas; see Section 5.

where $a^+ = \max(a, 0)$, $a^- = \max(-a, 0)$, and $G_{\tau z} = G\left(\tau, \Phi\left(z'\gamma\right); \rho\right)$ denotes the rank of $x'\beta(\tau)$ in the selected sample $D = 1$, conditional on $Z = z$. The function G plays a key role here, as it maps ranks in the latent distribution (that is, τ's) into ranks in the selected distribution (that is, $G_{\tau z}$'s).

It is instructive to compare (10) with the optimization problem which would characterize $\beta(\tau)$ were a sample from (Y^*, X) available, that is

$$\min_{b(\tau)} \mathbb{E}\left[\tau\left(Y^* - X'b(\tau)\right)^+ + (1-\tau)\left(Y^* - X'b(\tau)\right)^-\right]. \tag{11}$$

The function inside the expectation in (11) is the check function. In (10) we see that, in order to account for nonrandom sample selection, one needs to *rotate* the check function. The rotation angle depends on the amount of selection, and it is Z-specific. Such a rotation is needed unless U and V were independent, hence $G_{\tau Z} = \tau$, in which case standard quantile regression in the selected sample would be consistent for $\beta(\tau)$.

Interestingly, like (11), (10) is a linear program, hence in particular convex, and so is its sample counterpart. This implies that, given γ and ρ, one can estimate $\beta(\tau)$ for all τ in a τ-by-τ fashion by solving linear programs.

In practice γ and ρ need to be estimated. In the Gaussian specification for η, γ may be consistently estimated by a probit regression, as in the first step in the Heckman method.

In turn, the copula parameter ρ may be consistently estimated by taking advantage of the fact that (8) implies a number of moment restrictions (in fact, a continuum of such restrictions when covariates are continuously distributed), by using functions of Z as instruments. To describe the method to recover ρ, let us change the notation slightly and explicitly indicate the dependence on ρ and γ in $G_{\tau Z}^{\rho\gamma} = G\left(\tau, \Phi\left(Z'\gamma\right); \rho\right)$. For a given vector of instruments $\varphi(\tau, Z)$, ρ satisfies the following moment restrictions

$$\mathbb{E}\left[D\varphi(\tau, Z)\left(\mathbf{1}\left\{Y \leq X'\overline{\beta}(\tau; \rho, \gamma)\right\} - G_{\tau Z}^{\rho\gamma}\right)\right] = 0, \tag{12}$$

where

$$\overline{\beta}(\tau; \rho, \gamma) = \operatorname*{argmin}_{b(\tau)} \mathbb{E}\left[D\left(G_{\tau Z}^{\rho\gamma}\left(Y - X'b(\tau)\right)^+ + \left(1 - G_{\tau Z}^{\rho\gamma}\right)\left(Y - X'b(\tau)\right)^-\right)\right]. \tag{13}$$

The copula parameter ρ can thus be estimated in (12) for a finite set of τ values, based on the generalized method-of-moments (GMM, Hansen, 1982), by profiling out the $\overline{\beta}(\tau; \rho, \gamma)$ using (13).

In sum, given an i.i.d. sample (Y_i, Z_i, D_i), $i = 1, ..., N$ (where $Z_i = (X_i, W_i)$), Arellano and Bonhomme (2016)'s three-step estimation algorithm is as follows. A code written in Matlab is provided in the appendix.

Algorithm 1

1. *Estimate γ by a probit regression,*

$$\widehat{\gamma} = \underset{a}{\operatorname{argmax}} \sum_{i=1}^{N} D_i \ln \Phi(Z_i'a) + (1 - D_i) \ln \Phi(-Z_i'a).$$

2. *Estimate ρ by profiled GMM,*

$$\widehat{\rho} = \underset{c}{\operatorname{argmin}} \left\| \sum_{i=1}^{N} \sum_{\ell=1}^{L} D_i \varphi(\tau_\ell, Z_i) \left[\mathbf{1}\left\{ Y_i \leq X_i'\widehat{\beta}(\tau_\ell, c) \right\} - G\left(\tau_\ell, \Phi(Z_i'\widehat{\gamma}); c\right) \right] \right\|, \quad (14)$$

where $\| \cdot \|$ is the Euclidean norm, $\tau_1 < \tau_2 < ... < \tau_L$ is a finite grid on $(0, 1)$, $\varphi(\tau, Z_i)$ are instrument functions with $\dim \varphi \geq \dim \rho$, and

$$\begin{aligned} \widehat{\beta}(\tau, c) \quad = \quad & \underset{b(\tau)}{\operatorname{argmin}} \sum_{i=1}^{N} D_i \Big[G\left(\tau, \Phi(Z_i'\widehat{\gamma}); c\right) \left(Y_i - X_i'b(\tau)\right)^+ \\ & + \left(1 - G\left(\tau, \Phi(Z_i'\widehat{\gamma}); c\right)\right) \left(Y_i - X_i'b(\tau)\right)^- \Big]. \end{aligned} \quad (15)$$

3. *For any desired $\tau \in (0, 1)$, compute $\widehat{G}_{\tau i} = G\left(\tau, \Phi(Z_i'\widehat{\gamma}); \widehat{\rho}\right)$ for all i, and estimate $\beta(\tau)$ by rotated quantile regression,*

$$\widehat{\beta}(\tau) \quad = \quad \underset{b(\tau)}{\operatorname{argmin}} \sum_{i=1}^{N} D_i \Big[\widehat{G}_{\tau i} \left(Y_i - X_i'b(\tau)\right)^+ + \left(1 - \widehat{G}_{\tau i}\right) \left(Y_i - X_i'b(\tau)\right)^- \Big]. \quad (16)$$

Note that Step 3 is not needed when the researcher is only interested in $\beta(\tau_1), ..., \beta(\tau_L)$, in which case Steps 1 and 2 suffice.

The main computational cost of this algorithm is in Step 2. The objective function in (14) is neither continuous nor convex, because it features indicator functions. When modeling selection through a Gaussian copula, one can rely on grid-search for computation. Evaluating the objective function is usually fast and straightforward, because (15) is a linear program. In addition, using many percentile values τ_ℓ in (14) may smooth the objective function, hence aid computation.

The grid of τ values on the unit interval, and the instrument function $\varphi(\tau, Z)$, are to be chosen by the researcher. Although large grids slow down computation, it seems desirable to exploit a large number of restrictions to increase precision. Regarding the instruments, with a scalar ρ a possibility is to take φ to be the propensity score, or the propensity score multiplied by a function of τ. Optimal instruments may be constructed given a finite grid of τ's. However, characterizing efficiency properties in quantile selection models would require working under a continuum of moment restrictions.

Steps 1 and 2 in the above algorithm amount to profiled GMM estimation of a finite number of parameters: γ, ρ, and $\beta(\tau_1), ..., \beta(\tau_L)$. This is a well-understood estimation problem based on non-smooth moment functions. For example, the methods described in Newey and McFadden (1994) can be used to show root-N consistency and asymptotic normality, and characterize and estimate asymptotic variances. The asymptotic distribution of $\widehat{\beta}(\tau)$ in (16) may be derived using the same techniques. See Arellano and Bonhomme (2016) for a derivation of asymptotic variances. In fact, inference on the $\beta(\tau)$ process would also follow from standard arguments (Koenker and Xiao, 2002). Non-analytical methods, such as subsampling (Chernozhukov and Fernandez-Val, 2005), may also be used for inference. An important condition for estimator consistency is identification, which we discuss in the next section in a nonparametric setting.

Lastly, given estimates of conditional quantile functions, unconditional quantiles of latent outcomes and counterfactual distributions may be constructed using standard methods (Machado and Mata, 2005, Chernozhukov, Fernández-Val and Melly, 2013).

4 Identification

The methods described in the previous section rely on parametric assumptions on the propensity score and the copula, in addition to the assumed linear quantile specification for outcomes. It is possible to formulate and analyze a nonparametric quantile selection model where these assumptions are relaxed.

To proceed, let us replace (4) by a general quantile representation $Y^* = Q(\tau, X)$, where Q is increasing in its first argument, and let us allow for a nonparametric propensity score $p(Z)$ in (5). Lastly, let us assume that (U, V) is conditionally independent of Z given X, and denote the conditional copula of U given V and $X = x$ as $G_x(\tau, p)$. Here this function is also nonparametric.

Using similar arguments as in Section 3, one may derive the following set of restrictions

$$\Pr\left(Y \leq Q\left(\tau, x\right) \mid D = 1, Z = z\right) \quad = \quad G_x\left(\tau, p(z)\right). \tag{17}$$

The aim is to recover Q and G from (17). The propensity score $p(Z)$ is clearly identified based on data on participation and covariates. However, the quantile function Q and the conditional copula G consistent with (17) are not unique in general. This reflects the fact that the nonparametric quantile selection model is generally *set-identified*.

Arellano and Bonhomme (2016) emphasize two situations where the model is nonparametrically identified. A first case where $Q(\cdot, x)$ and G_x are identified is when $p(Z) = 1$ with positive probability conditional on $X = x$.[3] This case corresponds to "identification at infinity" (Chamberlain, 1986, Heckman, 1990). A second situation where identification holds is when the conditional copula G_x is real analytic. This case could be called "identification by extrapolation".

Given identification of a nonparametric Q for a parametric or analytic G, nonparametric rotated quantile regression methods based on kernel or series versions of (16) may then be used for estimation. Such methods may be combined with a flexible specification for G, based on Bernstein copulas, for example.

In other situations, the quantile function and conditional copula are generally partially identified. That is, a set of such functions is consistent with the population distribution. In a quantile model, failure of point-identification affects the entire quantile curve. This contrasts with semi-parametric linear models, where arguments such as "identification at infinity" are only needed to point-identify intercept parameters (Andrews and Schafgans, 1998, Das, Newey and Vella, 2003). Bounds on these functions may be constructed following Manski (1994, 2003). In practice, estimating such bounds may help assess the impact of parametric forms on the results, as described in Arellano and Bonhomme (2016).

5 Other approaches

In this section we review several related approaches. We start with an alternative approach to Arellano and Bonhomme (2016) based on maximum likelihood. We then review control

[3] To see why this is the case, take a $z = (x, w)$ such that $p(z) = 1$. As $G_x(\tau, 1) = \tau$, evaluating (17) at $Z = z$ shows that $Q(\tau, x)$ is identified as the τ-th conditional quantile of Y given $D = 1$ and $Z = z$. This is intuitive, as conditioning on the propensity score being 1 removes the sample selection problem.

function approaches to selection-correction in quantile models. Lastly, we clarify the link between selection correction and censoring correction in quantile regression.

A likelihood approach. The approach outlined in Section 3 is only one of several estimation possibilities in quantile selection models. As an example, a principled alternative would be to estimate the parameters of interest using a maximum likelihood approach. To see how this would work, note that evaluating (6) at $\tau = F(y \,|\, x)$ yields

$$\Pr\left(Y^* \leq y \,|\, D = 1, Z = z\right) \;\; = \;\; G\left(F(y \,|\, x; \beta(\cdot)), \Phi\left(z'\gamma\right); \rho\right),$$

where we have indicated the dependence of $F(y \,|\, x)$ on the quantile process $\beta(\tau)$. A joint (semi-parametric) maximum likelihood estimator would thus maximize

$$\sum_{i=1}^{N} D_i \ln \Phi(Z_i'a) + (1 - D_i) \ln \Phi(-Z_i'a) + \sum_{i=1}^{N} D_i \ln f(Y_i \,|\, X_i; b(\cdot)) \\ + \sum_{i=1}^{N} D_i \ln \nabla G\left(F(Y_i \,|\, X_i; b(\cdot)), \Phi\left(Z_i'a\right); c\right)$$

with respect to a, c, and all $b(\tau)$ for $\tau \in (0, 1)$, where f is the conditional pdf of Y^* given X, and ∇G denotes the derivative of G with respect to its first argument. An intermediate approach would be to profile out the $\beta(\tau)$'s using (15), and estimate ρ based on the profiled likelihood.

In contrast with a likelihood-based approach, the estimator described in Section 3 exploits the τ-by-τ separability of the rotated quantile regression problems, as well as the convexity of the rotated quantile regression objective functions. On the other hand, such sequential estimators are not asymptotically efficient in general.

Control function approaches. Buchinsky (1998, 2001) introduced a control function method to correct for sample selection in quantile regression models. The method consists in controlling for functions of the propensity score in the quantile regression. This approach delivers consistent estimates in additive models with independent errors, as in this case quantile functions of selected outcomes are indeed additive in covariates X and propensity score $p(Z)$. As an example, in the Gaussian Heckman model, (7) becomes

$$Q^s(\tau, Z) = X'\beta + \sigma\Phi^{-1}\left[G^{-1}\left(\tau, p(Z); \rho\right)\right], \tag{18}$$

where $p(Z) = \Phi(Z'\gamma)$, and $G^{-1}(\cdot, \cdot; \rho)$ is the inverse of the conditional Gaussian copula with respect to its first argument.

However, as (7) shows, in non-additive models such as quantile selection models, quantile curves are generally not additive in X and $p(Z)$. As a result, as pointed out by Huber and Melly (2015), additive control function methods will not be consistent in general. Huber and Melly use this observation to develop tests of the additivity assumption (which they call "conditional independence") that make use of the Buchinsky control function estimator.

Link to censoring corrections. The selection correction problem reviewed in this chapter is different from other censoring corrections that have been extensively studied in the quantile regression literature. To see the link between these two problems consider an outcome variable modeled as in (4), observed only when $Y^* \leq \mu$, where μ is a known constant. That is, Y^* is censored above μ. In this case, denoting censoring as $D = 1$, the censoring rule takes the form in (5), with $Z = X$, $p(X) = F(\mu \,|\, X)$ (where F is the conditional cdf of Y^* given X), and $V = U$. The threshold μ need not be known. Moreover, it could be a (known or unknown) function $\mu(X)$ of covariates.

The conditional copula of (U, V) is thus known in this case, however it is degenerate. It coincides with the conditional upper Fréchet bound (Fréchet, 1951), whose expression is

$$G^+(\tau, p) = \min\left\{\frac{\tau}{p}, 1\right\}.$$

Note that G^+ is not strictly monotone in its first argument. Analogously as in (10) one may base estimation of $\beta(\tau)$, for any given τ, on the following optimization problem, focusing on the subpopulation with $p(X) > \tau$ where the τth conditional quantile is identified,

$$\min_{b(\tau)} \mathbb{E}\left[D\mathbf{1}\{p(X) > \tau\}\left(\frac{\tau}{p(X)}\left(Y - X'b(\tau)\right)^+ + \left(1 - \frac{\tau}{p(X)}\right)\left(Y - X'b(\tau)\right)^-\right)\right]. \tag{19}$$

Equation (19) is the basis for the censored quantile regression estimator of Buchinsky and Hahn (1998). A slight difference with the analysis in Buchinsky and Hahn (1998) is that they consider a case where outcomes are censored from below, so the relevant conditional copula in their case is the lower Fréchet bound $G^-(\tau, p) = \max\left\{\frac{\tau+p-1}{p}, 0\right\}$. Buchinsky and Hahn propose to nonparametrically estimate the propensity score. Their estimator solves a convex problem, like the sample selection estimator reviewed in Section 3. This contrasts with the estimator of Powell (1986), which is based on a non-convex objective function.

One difference between this model and the bivariate sample selection model is that, in the censoring model, the propensity score $p(X) = F(\mu \mid X)$ depends on the unknown distribution of latent outcomes. This explains the need for nonparametric estimation of $p(X)$, in contrast with the quantile selection model where the propensity score may be parametrically specified while preserving statistical coherency.

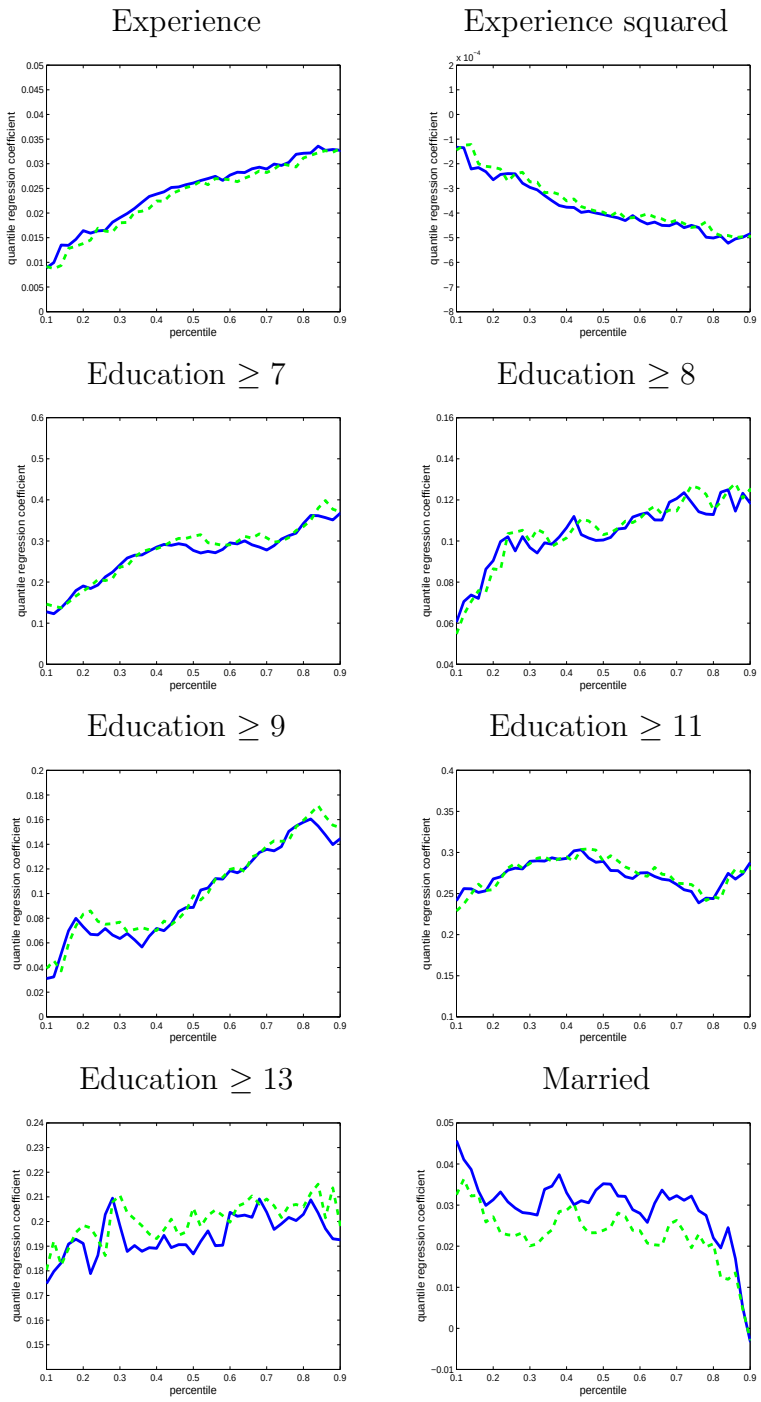
6 Empirical illustration

In this last section we revisit the empirical illustration in Huber and Melly (2015), who study the returns to education and experience for women in the US. We take their sample from the 2011 Merged Outgoing Rotation Groups of the Current Population Survey (CPS). The sample consists of 44,562 White women, 20,055 of whom are working outside of self-employment, the military, agriculture, and the public sector. The only difference with the sample in Huber and Melly (2015) is that we drop working women with missing wage values (1.6% of observations). Working is defined as having worked more than 35 hours in the week preceding the survey.

Huber and Melly test, and reject, the assumption that quantile functions are additive in X and τ on these data. Using the methods described in Section 3, here we estimate quantile regression specifications that account for the presence of sample selection. The dependent variable Y is the log-hourly wage, covariates X contain general labor market experience (that is, age minus the number of years of schooling) and its square, five education indicators (more than 7, 8, 9, 11, and 13 years of education), interactions of experience and its square with years of schooling, and indicators for marital status and region of residence (4 regions). As determinants of participation assumed not to enter the wage equation, we take the number of children in 3 age ranges and their interactions with marital status. We use CPS sample weights in all the computations. Finally, our specification is based on a probit propensity score and a Gaussian copula, we take $\varphi(\tau, Z) = \sqrt{\tau(1-\tau)}\widehat{p}(Z)$, with $\widehat{p}(Z)$ the estimated propensity score, and we take a grid τ_ℓ of deciles when estimating the copula parameter ρ. Computation of $\widehat{\rho}$ in (14) is based on grid search.

Figure 1 presents the estimates for a selection of covariates. Quantiles corrected for sample selection are shown in dashed, while uncorrected ones are shown in solid lines. Similarly as in Huber and Melly (2015), excluded covariates (that is, number of children and interactions with marital status) are strongly significant in the participation equation, estimates

Figure 1: Female wages (CPS, 1991), quantile regression curves



Notes: CPS, 1991. Sample comprising 44,562 women, 20,055 of whom are working. Quantile regression estimates. Dashed line: corrected for selection. Solid line: uncorrected.

Table 1: Female wages (CPS, 1991), parameter estimates

	Uncorrected estimates			
	Regression	Quantile regression		
		$\tau = .25$	$\tau = .50$	$\tau = .75$
Experience	.022 (.0024)	.016 (.0026)	.026 (.0024)	.029 (.0026)
Experience squared	-.00035 (.000061)	-.00024 (.000057)	-.00041 (.000062)	-.00043 (.000057)
Education ≥ 7	.30 (.028)	.20 (.031)	.28 (.029)	.31 (.031)
Education ≥ 8	.10 (.014)	.10 (.015)	.10 (.014)	.12 (.015)
Education ≥ 9	.10 (.015)	.067 (.016)	.089 (.015)	.14 (.017)
Education ≥ 11	.26 (.017)	.28 (.018)	.29 (.017)	.24 (.019)
Education ≥ 13	.19 (.018)	.20 (.019)	.19 (.017)	.20 (.019)
Experience $\times$ education	.0020 (.00053)	.0013 (.00063)	.0024 (.00058)	.0034 (.00064)
Experience squared $\times$ education	-.000044 (.000013)	-.000024 (.000015)	-.000045 (.000014)	-.000075 (.000015)
Married	.027 (.0091)	.032 (.0092)	.035 (.0085)	.030 (.0093)
	Selection-corrected estimates			
	Regression	Quantile regression		
		$\tau = .25$	$\tau = .50$	$\tau = .75$
Experience	.022 (.0026)	.016 (.0041)	.026 (.0030)	.030 (.0034)
Experience squared	-.00034 (.000061)	-.00024 (.00010)	-.00040 (.000071)	-.00046 (.000079)
Education ≥ 7	.32 (.037)	.20 (.15)	.31 (.089)	.30 (.037)
Education ≥ 8	.11 (.014)	.10 (.026)	.10 (.014)	.13 (.017)
Education ≥ 9	.10 (.017)	.075 (.026)	.098 (.019)	.14 (.021)
Education ≥ 11	.27 (.018)	.28 (.029)	.30 (.020)	.25 (.025)
Education ≥ 13	.19 (.021)	.19 (.029)	.21 (.022)	.20 (.024)
Experience $\times$ education	.0020 (.00063)	.0011 (.0012)	.0020 (.00076)	.0033 (.00089)
Experience squared $\times$ education	-.000042 (.000016)	-.000021 (.000033)	-.000041 (.000019)	-.000070 (.000022)
Married	.022 (.011)	.023 (.017)	.024 (.013)	.021 (.013)

Notes: CPS, 1991. Sample comprising 44,562 women. Regression and quantile regression estimates, corrected for selection and uncorrected. Covariates also include regional indicators and an intercept (not included). Robust standard errors in parentheses. Standard errors in bottom panel based on subsampling (subsample size 1000*,* 500 *replications).*

being omitted here for brevity. We see that quantile regression estimates vary substantially along the distribution. This is in line with Huber and Melly's finding that an additive model is not appropriate for this data. At the same time, correcting for sample selection tends to make small differences on the results. Although the coefficients for marital status differ by some margin, most of the uncorrected and corrected estimates are close to each other. This is so in spite of an estimated correlation $\widehat{\rho} = -.10$ (standard error .064), which reflects some positive selection of women into participation.[4] The variation along the distribution, and the similarity between uncorrected and selection-corrected estimates, are confirmed by the parameter estimates reported in Table 1.

We performed a number of robustness checks. We experimented with grids of τ's of different sizes (equidistant grids with 2 to 50 knots), a different choice for the instrument function ($\varphi(\tau, Z) = \widehat{p}(Z)$), and a different choice for G (based on the Frank copula). These choices all lead to very similar results as in Figure 1 and Table 1. Moreover, the minimum of the objective function in ρ was easy to identify in all experiments, although the objective function tended to be erratic for large values of $|\rho|$.

It is to be noted, however, that leaving the functional form of the copula fully unrestricted in this application would most likely lead to lack of point-identification. Due to the restricted support of the propensity score (99% of the estimated propensity score being below .73 in the sample), nonparametric bounds based on a worst-case analysis would be wide. If desired, the method described in Arellano and Bonhomme (2016) may be used to compute estimates of worst-case bounds in a given application.

Overall, on this data correcting for sample selection confirms the findings from standard quantile regression. It would be of interest to estimate similar specifications on other periods, especially since Mulligan and Rubinstein (2008) argue that the direction and intensity of female's selection into employment has changed since the 1970's in the US.

7 Conclusion

Sample selection correction methods for linear models remain very popular in applied work. Since James Heckman's pioneering work, the use of these methods has led to uncovering important empirical regularities. In this chapter we have reviewed recently proposed selection-

[4]Note that ρ is the correlation between U in (4) and V in (5). Hence, a negative ρ means that high-U women have a higher propensity to participate to the labor market.

correction approaches in nonlinear quantile models. The hope is that these methods could help documenting distributional effects in a variety of empirical settings where nonrandom sample selection arises.

Existing work on nonlinear selection models is scarce, however, and there remains a lot to be done. A particular issue is the reliance on parametric functional forms. In general, relaxing these assumptions results in lack of point-identification. This is an area where recently developed methods allowing for uniform inference in the presence of partial identification (e.g., Tamer, 2010) might prove particularly useful.

References

[1] Ahn, H., and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

[2] 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(4), 383–396.

[3] Andrews, D. W., and M. M. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65(3), 497–517.

[4] Arellano, M., and S. Bonhomme (2016): "Quantile Selection Models," to appear in *Econometrica*.

[5] Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75, 323–364.

[6] Bollinger, C., J. P. Ziliak, and K. R. Troske (2011): "Down from the Mountain: Skill Upgrading and Wages in Appalachia," *Journal of Labor Economics*, 29(4), 819–857.

[7] Buchinsky, M. (1994): "Changes in the U.S. Wage Structure 1963 to 1987; An Application of Quantile Regressions," *Econometrica*, 62, 405–458.

[8] 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.

[9] Buchinsky, M. (2001): "Quantile Regression with Sample Selection: Estimating Women's Return to Education in the US," *Empirical Economics*, 26, 87–113.

[10] Buchinsky, M., and J. Hahn (1998): "An Alternative Estimator for the Censored Regression Model," *Econometrica*, 66, 653–671.

[11] Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

[12] Chamberlain, G. (1993): "Quantile Regressions, Censoring and the Structure of Wages," in C. Sims (ed.), *Advances in Econometrics: Proceedings of the 6th World Congress in Barcelona*, Vol. I (Cambridge: Cambridge University Press).

[13] Chen, S., and S. Khan (2003): "Semiparametric Estimation of a Heteroskedastic Sample Selection Model," *Econometric Theory*, 19(6), 1040–1064.

[14] Chernozhukov, V., and I. Fernández-Val (2005): "Subsampling Inference on Quantile Regression Processes," *Sankhya*, 67(2), 253–276.

[15] Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski (2015): "Quantile Regression with Censoring and Endogeneity," *Journal of Econometrics*, 186(1), 201–221.

[16] Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81(6), 2205-68.

[17] Chernozhukov, V., and H. Hong (2002): "Three-Step Censored Quantile Regression and Extramarital Affairs", *Journal of the American Statistical Association*, 97(459), 872–882.

[18] Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

[19] Donald, S. G. (1995): "Two-Step Estimation of Heteroskedastic Sample Selection Models," *Journal of Econometrics*, 65(2), 347–380.

[20] Fréchet, M. (1951): "Sur les Tableaux de Corrélation dont les Marges sont Données," *Ann. Univ. Lyon Sér. 3*, 14, 53–77.

[21] Gronau, R. (1974): "Wage Comparison - A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143.

[22] Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1029–1054.

[23] Heckman, J. J. (1974): "Shadow Prices, Market Wages and Labour Supply," *Econometrica*, 42, 679–694.

[24] Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," in *Annals of Economic and Social Measurement*, 5(4), 475–492.

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

[26] Heckman, J. J. (1990): "Varieties of Selection Bias," *The American Economic Review*, 80, 313–318. (Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association.)

[27] Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating Trade Flows: Trading Partners and Trading Volumes," *Quarterly Journal of Economics*, 123(2), 441–487.

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

[29] Jimenez, G., S. Ongena, J. L. Peydró, and J. Saurina (2014): "Hazardous Times for Monetary Policy: What Do Twenty-Three Million Bank Loans Say About the Effects of Monetary Policy on Credit Risk-Taking?" *Econometrica*, 82(2), 463–505.

[30] Kitagawa, T. (2010): "Testing for Instrument Independence in the Selection Model," Unpublished Manuscript.

[31] Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

[32] Koenker, R., and Z. Xiao (2002): "Inference on the quantile regression process," *Econometrica*, 70, 1583–1612.

[33] Lee, L.-F. (1983): "Generalized Econometric Models with Selectivity," *Econometrica*, 51(2), 507–512.

[34] 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.

[35] Manski, C. F. (1994): "The Selection Problem" in *Advances in Econometrics*, Sixth World Congress, Vol 1, ed. by C. Sims. Cambridge, U.K.: Cambridge University Press, 143–170.

[36] Manski, C. F. (2003): *Partial Identification of Probability Distributions*, Springer Series in Statistics, Vol. 12. Berlin: Springer Verlag.

[37] Mulligan, C., and Y. Rubinstein (2008): "Selection, Investment, and Women's Relative Wages Over Time," *Quarterly Journal of Economics*, 123(3), 1061–1110.

[38] Nelson, F. D. (1984): "Efficiency of the Two-Step Estimator for Models with Endogenous Sample Selection," *Journal of Econometrics*, 24, 181–196.

[39] Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245.

[40] Portnoy, S. (2003): "Censored Regression Quantiles," *Journal of the American Statistical Association*, 98(464), 1001–1012.

[41] Powell, J. L. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

[42] Smith, M. D. (2003): "Modelling Sample Selection Using Archimedean Copulas," *The Econometrics Journal*, 6, 99–123.

[43] Tamer, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

[44] Van Kerm, P. (2013): "Generalized Measures of Wage Differentials," *Empirical Economics*, 45(1), 465–482.

[45] Vella, F. (1998): "Estimating Models with Sample Selection Bias: A Survey," 33(1), 127–169.

APPENDIX: CODE

```
%%% Input:
%%% participation indicator D (binary)
%%% outcome variable Y (note: values of Y for D=0 are arbitrary, they may even be missing)
%%% matrix of covariates X, not including the constant
%%% matrix of excluded covariates B, not including the constant
%%% Z=(X,B) is constructed below

%%% Specification in this version:
%%% propensity score: probit
%%% copula: Gaussian
%%% instrument function: varphi(Z) = propensity score p(Z)
%%% grid of tau's for tau=.2,...,.8
%%% estimates of beta(tau) for tau=.05,...,.95
%%% grid search for rho with 99 equidistant values

%%% The code uses the routine rq.m from Morillo, Koenker and Eilers
%%% available at: http://www.econ.uiuc.edu/~roger/research/rq/rq.m


%%% Estimate propensity score
Z=[X B];
gamma=glmfit(Z,D,'binomial','link','probit');
pZ=normcdf(gamma(1)+Z*gamma(2:end));

%%% Instrument function
varphi=pZ;

%%% Select participants
Y1=Y(D==1);
pZ1=pZ(D==1);
X1=X(D==1,:);
varphi1=varphi(D==1);
[N1 colx]=size(X1);

%%% Percentile values to estimate rho
vectau=(.20:.20:.80)';
[t n]=size(vectau);

%%% Values of rho in the grid
vecrhoa=(-.98:.02:.98)';
[s n]=size(vecrhoa);

%%% Objective function to be minimized
object=zeros(s,1);
```

```
%%% Start grid search

for j=1:s
rhoa=vecrhoa(j);
obj=0;

for k=1:t
tau=vectau(k);

%%% Gaussian copula
G=copulacdf('Gaussian',[tau*ones(N1,1) pZ1],rhoa)./pZ1;

%%% Rotated quantile regression
beta=rq([ones(N1,1) X1],Y1,G);
obj=obj+(mean(varphi1.*((Y1<=beta(1)+X1*beta(2:colx+1))-G)));
end

object(j)=abs(obj);
end

%%% Minimize the objective function
[C I]=min(object);
rho=vecrhoa(I);

%%% Estimate selection-corrected quantile parameters beta(tau)
beta=zeros(colx+1,19);

for tau=.05:.05:.95

%%% Gaussian copula
G=copulacdf('Gaussian',[tau*ones(N1,1) pZ1],rho)./pZ1;

%%% Rotated quantile regression
beta(:,round(20*tau))=rq([ones(N1,1) X1],Y1,G);
end

%%% Output
%%% the first column in beta corresponds to the intercept
%%% different columns are tau=.05 to tau=.95
display('copula parameter')
rho
display('quantile coefficients')
beta
```