Universität St.Gallen

Quantile Regression in the Presence of Sample Selection

Martin Huber and Blaise Melly

March 2011 Discussion Paper no. 2011-09

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
School of Economics and Political Science
Department of Economics
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St. Gallen
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25

Electronic Publication: Fax +41 71 224 31 35
http://www.seps.unisg.ch

Quantile Regression in the Presence of Sample Selection[1]

Martin Huber and Blaise Melly

Author's address:

Martin Huber, Ph.D.
SEW, University of St. Gallen
Varnbüelstrasse 14, 9000 St. Gallen
Phone +41 71 224 2299
Fax +41 71 224 2302
Email Martin.Huber@unisg.ch
Website www.sew.unisg.ch

Blaise Melly, Ph.D.
Brown University, Department of Economics
Providence, RI, USA
Phone +1 401 863 6297
Fax +1 401 863 1970
Email Blaise_Melly@brown.edu
Website www.econ.brown.edu/fac/Blaise_Melly

[1] We would like to thank Stefan Hoderlein, Frank Kleibergen, Michael Lechner, and seminar/conference participants at Brown University, University of St. Gallen, the labor market seminar of the University of Zürich in Engelberg, the conference "Inference and Tests in Econometrics" in Marseille, and the COST A23 conference in Paris.

Abstract

Most sample selection models assume that the errors are independent of the regressors. Under this assumption, all quantile and mean functions are parallel, which implies that quantile estimators cannot reveal any (per definition non-existing) heterogeneity. However, quantile estimators are useful for testing the independence assumption, because they are consistent under the null hypothesis. We propose tests for this crucial restriction that are based on the entire conditional quantile regression process after correcting for sample selection bias. Monte Carlo simulations demonstrate that they are powerful and two empirical illustrations indicate that violations of this assumption are likely to be ubiquitous in labor economics.

Keywords

Sample selection, quantile regression, independence, test.

JEL Classification

C12, C13, C14, C21.

1 Introduction

Estimation of economic models is frequently complicated by the problem of sample selection: the variables of interest are only observed for a non-random subsample of the population. A prominent example in labor economics consists in the estimation of the determinants of female wages. Individuals are assumed to offer positive labor supply only if their potential wage exceeds their reservation wage. It is well known that standard procedures will be biased if unobservables jointly affect the decision of working and the potential wage. The ability to consistently estimate econometric models in the presence of non-random sample selection is one of the most important innovations in microeconometrics, as illustrated by the Nobel Prize received by James Heckman.

Gronau (1974) and Heckman (1974, 1976 and 1979) addressed the selectivity bias and proposed fully parametric estimators, assuming that the residuals are independent and jointly normally distributed. This approach yields inconsistent results if the distribution of the error term is misspecified. Therefore, Cosslett (1991), Gallant and Nychka (1987), Powell (1987), and Newey (2009) proposed semiparametric estimators for the sample selection model. They relaxed the distributional assumption but kept the single index structure in both the selection and the outcome equation. In addition, Ahn and Powell (1993) dropped the index structure in the selection equation. More recently, Das, Newey, and Vella (2003) considered fully nonparametric sample selection models. While these estimators have progressively weakened the parametric and distributional assumptions originally made, none of them is robust to the presence of conditional heteroscedasticity or higher order dependence between the residuals and the outcome.[1]

However, dependence in general and heteroscedasticity in particular is a ubiquitous phenomenon in the fields where sample selection models have been used. As suggested by Mincer (1973) in his famous human capital earnings model, residual wage dispersion should increase with experience and education. In line with this finding, the large majority of the applications using quantile regression in the empirical literature find significant heterogeneity in the returns to education and experience. Therefore, the independence assumption cannot be

[1]Mean estimators only require the existence of a conditional moment restriction in the observed sample for consistency. Therefore, a moment condition is sometimes assumed directly without imposing full independence, but having the latter as a potential justification for the assumption. Note, however, that departures from full independence that still satisfy the moment condition are not substantial. E.g., the moment condition allows for heteroscedastic measurement errors affecting the dependent variable but not for heteroscedastic wage functions, see the discussion in Newey and Powell (2003).

taken as granted in most economic applications. Donald (1995) alleviated the independence assumption and proposed a two-step estimator that allows for conditional heteroscedasticity but requires the error terms to be bivariate normally distributed. Chen and Khan (2003) allowed for non-normality and heteroscedasticity. However, we show in Appendix B that proper identification of their model de facto requires a new type of exclusion restriction: the availability of a regressor that affects the variance but not the location of the dependent variable.

Quantile regression has progressively emerged as the method of choice to analyze the effects of variables on the distribution of the outcome. In the absence of selection, Koenker and Bassett (1978) proposed a parametric (linear) estimator for conditional quantile models. Due to its ability to capture heterogeneous effects, its theoretical properties have been studied extensively and it has been used in many empirical studies; see, for example, Powell (1986), Guntenbrunner and Jurečková (1992), Buchinsky (1994), Koenker and Xiao (2002), Angrist, Chernozhukov, and Fernández-Val (2006). Chaudhuri (1991) suggested a nonparametric quantile regression estimator. Buchinsky (1998b), Koenker and Hallock (2001), and Koenker (2005) provide a comprehensive discussion of quantile regression models and recent developments.

Buchinsky (1998a and 2001) was the first to consider the difficult problem of estimating quantile regression in the presence of sample selection.[2] He extended the series estimator of Newey (2009) for the mean to the estimation of quantiles. Even in this approach the independence assumption is required to obtain partially linear representations for the conditional quantile functions in the observed sample. He assumed conditional independence between the error terms and the regressors given the selection probability. This assumption implies that all quantile regression curves are parallel, which limits the usefulness of considering several quantile regressions that by assumption give the same result. In addition, the quantile slope coefficients are identical to the mean slope coefficients.

The estimator proposed by Buchinsky is nevertheless useful for several reasons. The original motivation for quantile regression was not the estimation of heterogeneous effects on the conditional distribution but the robustness of the estimates in the presence of non-Gaussian errors.[3]

[2]Buchinsky (1998a) was awarded the Richard Stone Prize in Applied Econometrics for the best paper with substantive econometric application that has been published in the 1998 and 1999 volumes of the Journal of Applied Econometrics. It was also included in the virtual issue "Celebrating 25 years of the Journal of Applied Econometrics" as one of the most downloaded and cited articles during the JAE's history.

[3]Ironically, Koenker and Bassett (1978) assume independence in their seminal paper.

A similar result holds in the sample selection model and we illustrate the considerable efficiency gains that can be achieved when the error distribution has fat tails in simulations. The second motivation for quantile regression was to provide robust and powerful tests for heteroscedasticity, as suggested by Koenker and Bassett (1982). Testing the independence assumption is even more acute in the presence of sample selection because, as mentioned above, mean and quantile estimators are inconsistent if this assumption is violated. Under the null hypothesis of independence, the procedure proposed by Buchinsky (1998a) consistently estimates the slope coefficients, which are constant as a function of the quantile. When the independence assumption is violated, the estimated slope coefficients, while inconsistent, will be a nontrivial function of the quantile. Therefore, we suggest testing the independence assumption by testing whether the coefficients vary across quantiles. To the best of our knowledge, this is the first test for this identifying assumption.

We could consider a finite number of quantile regression coefficients and jointly test for their equality but more powerful test statistics can be built using the entire conditional quantile process, see Koenker and Xiao (2002). We therefore suggest a test procedure similar to that proposed by Chernozhukov and Fernandez-Val (2005). The critical values for this test are obtained by resampling the empirical quantile regression processes. Since the computation of the estimates is quite demanding, we follow Chernozhukov and Hansen (2006) and propose score resampling instead of recomputing the whole process. Monte Carlo simulations indicate that size and power properties of the suggested Kolmogorov-Smirnov and Cramer-Von-Mises tests are very satisfactory.

After having provided the technology to detect violations of the independence assumption, we examine whether such violations are an empirically relevant phenomenon by considering two data sets which are representative for the application of sample selection correction procedures. First, we apply the test to the medium-sized data of Martins (2001) and reject the independence assumption at the 5% significance level. Second, using the more recent and considerably larger sample of Mulligan and Rubinstein (2008), we reject the null hypothesis with even higher confidence. We suspect that this problem is not limited to a few cases but is widespread in fields where sample selection models have been used.[4]

[4]The codes for the simulations and applications and the datasets used in this paper can be downloaded at http://www.econ.brown.edu/fac/Blaise_Melly/code_R_selection.html. The interested researchers can, therefore, easily verify whether our claim is true or not in their applications.

What can be done in the case of rejecting the independence assumption? Unfortunately, the parameters of interest are no longer point identified in the absence of the independence (separability) assumption. In our companion paper Melly and Huber (2011), we derive the sharp bounds on the quantile regression parameters when the this assumption is no longer imposed. In this case, point identification can be attained only by an identification at infinity argument or by a parametric assumption. Arellano and Bonhomme (2010) obtain point identification by a clever parametrization of the copula between the error terms in the selection and outcome equations while keeping their marginal distributions nonparametric.

The remainder of this paper is organized as follows. In Section 2 we describe the sample selection model of Buchinsky (1998a) and discuss the implication of the independence assumption in quantile models. Section 3 outlines the test procedure. In Section 4 Monte Carlo simulations document the efficiency and robustness of quantile regression in sample selection models as well as the power and size properties of the proposed test. Section 5 revisits two empirical applications of sample selection models. Section 6 concludes.

2 The Sample Selection Model

In this paper, we consider the same sample selection framework of Buchinsky (1998a), which can be regarded as the quantile version of Newey (2009). As in the seminal work of Heckman (1974, 1976 and 1979), the outcome equation and the latent selection function are linear in the covariates. The error terms in both equations are independent of the covariates (conditional on the selection probability), but in contrast to the model of Heckman their joint distribution is completely unrestricted. At this point, we would like to emphasize that the choice of linear outcome and latent selection equations are made for completeness and to simplify the comparison with important existing estimators. We could relax the assumptions restricting the selection equation and allow for a fully nonparametric selection probability function as in Ahn and Powell (1993). Furthermore, we could also allow for a nonparametric outcome equation as in Das, Newey, and Vella (2003). Therefore, the insights of this paper about the implications and testability of the independence assumption are valid for a much wider set of models than the linear case.

Bearing this in mind, we maintain the following assumption (equation 2 in Buchinsky, 1998a):

$$Y_i^* = c + X_i'\beta + \varepsilon_i, \tag{1}$$

where Y^* denotes a potential outcome of interest, e.g. the potential hourly wage, X denotes a vector of regressors without a constant, β is the vector of slope coefficients and ε_i is the error term.

We do not observe the latent variable Y_i^* but only Y_i, which is defined by

$$Y_i = Y_i^* \text{ if } D_i = 1 \text{ and not observed otherwise.}$$

D is an indicator function that depends on Z, a superset of X.[5] The rest of the paper does not depend on how $\Pr(D = 1|Z)$ is identified but for completeness we make the following assumption:

$$D_i = 1\left(Z_i'\alpha + U_i \geq 0\right). \tag{2}$$

The selection probability is restricted to depend on the linear index $Z'\alpha$. In the implementation of the test we will estimate α using the efficient semiparametric procedure suggested by Klein and Spady (1993). Therefore, we rely on their restrictions and assume that $U \perp Z|Z'\alpha$. This conditional independence assumption for U can be relaxed if $\Pr(D = 1|Z)$ is estimated nonparametrically, as in Ahn and Powell (1993).

The model is not point identified without further assumptions. Following Buchinsky (1998a), we assume:

Assumption 1: (U, ε) has a continuous density,
Assumption 2: $f_{U,\varepsilon}(\cdot|Z) = f_{U,\varepsilon}(\cdot|Z'\alpha)$, where $f_{U,\varepsilon}$ denotes the joint density of U and ε.

These assumptions identify the parameter β. By Assumption 2, ε and U are independent of Z given $Z'\alpha$, therefore, for any quantile $0 < \tau < 1$

$$\begin{aligned} Q_\tau(\varepsilon|Z, D = 1) &= Q_\tau(\varepsilon|Z, U \geq Z'\alpha) \\ &= Q_\tau(\varepsilon|Z'\alpha, U \geq Z'\alpha) \\ &= Q_\tau(\varepsilon|Z'\alpha, D = 1). \end{aligned}$$

where $Q_\tau(\varepsilon|Z, D = 1)$ denotes the τ^{th} conditional quantile of ε given Z and $D = 1$. Since X is a subset of Z, it follows that $Q_\tau(\varepsilon(\tau)|X, D = 1)$ depends on X only through the linear index $Z'\alpha$.

[5] For identification, Z has to include at least one continuous variable which is not in X and has a non-zero coefficient in the selection equation.

Thus, for any $0 < \tau < 1$,

$$\begin{aligned} Q_\tau\left(Y^*|X=x, Z=z, D=1\right) &= c + x\beta + Q_\tau(\varepsilon|z'\alpha, D=1) \\ &= x\beta + h_\tau\left(z'\alpha\right), \end{aligned}$$

where $h_\tau\left(z'\alpha\right)$ is an unknown function that depends only on $z'\alpha$ by Assumption 2. In other words, β can be estimated consistently by *any* quantile regression of Y on X and a nonparametric function of $z'\hat{\alpha}$, where $\hat{\alpha}$ is a first stage estimate of α.

It is obvious that these assumptions also have an unwanted consequence. The additivity in ε in equation (1) associated with the conditional independence of ε and X in Assumption 2 implies that all quantile regressions lead to the same slope coefficients. However, in the majority of cases where quantile methods are applied the researcher is particularly interested in the heterogeneity of the coefficients across the distribution. In the sample selection model, this heterogeneity paradoxically points to the violation of (at least) one identifying assumption, since conditional independence implies the homogeneity of the coefficients. In addition, knowing the slope of all quantile regressions is not more informative than knowing the slope of the mean regression.[6] Quantile regression may, however, be preferred for the sake of robustness. We discuss this advantage in Section 4.1.

At the same time, our arguments imply that Assumption 2 can be tested (while maintaining the other assumptions) by testing the equality of the quantile coefficients. If different quantile regressions give different slope coefficients, this has to imply that the independence assumption is violated. We suggest such a test based on the quantile estimator of Buchinsky (1998a) in Section 3. Our test bears great relevance for empirical work, as the independence assumption is a necessary condition for the consistency of the estimators suggested in Heckman (1979), Cosslett (1991), Gallant and Nychka (1987), Powell (1987), and Newey (2009), to name only a few. Even though the importance of this assumption in sample selection models has not remained unnoticed in the literature, see for instance Angrist (1997), we appear to be the first ones who suggest a formal test.

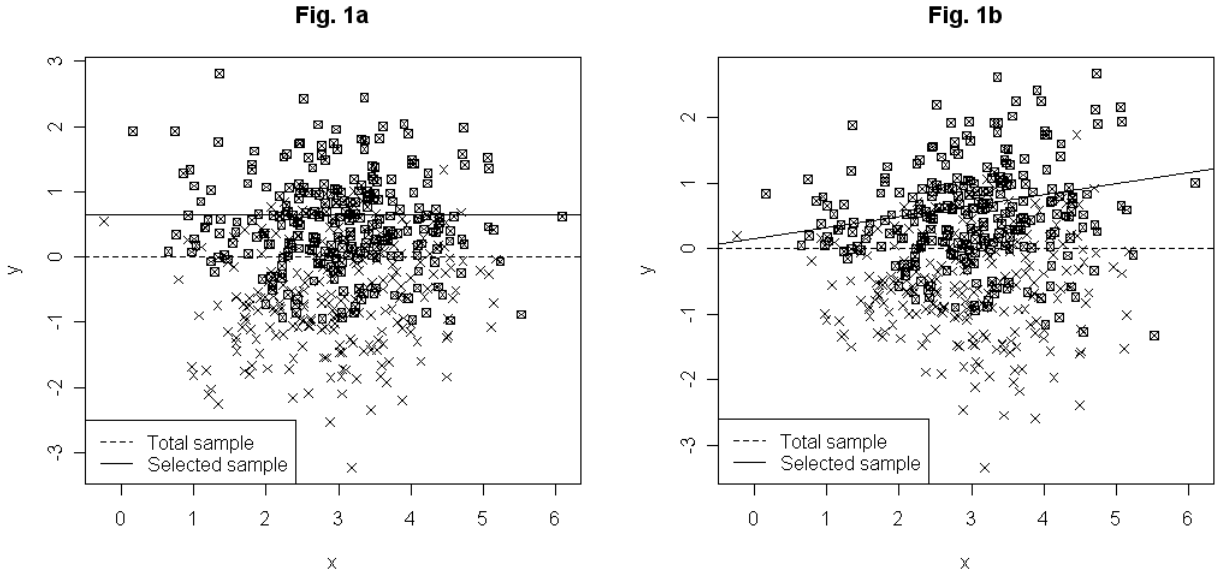
A graphical illustration shall convey the intuition for the necessity of the independence assumption and the possibility to use quantile regression to test its validity. Figure 1 displays 500

[6]The constant term of the mean or quantile function is not identified without further assumptions. Buchinsky (1998a) uses an identification at infinity argument to identify $c\left(\tau\right)$. We do not pursue this strategy in this paper and only consider the slope coefficients.

simulated realizations of (X, Y^*) which have exactly the same conditional selection probability $\Pr(D = 1|Z) = 0.69$. In the first case, the errors are independent while they are heteroscedastic in the second case. The true median regression curve (dashed line) is flat in both cases. There is the same positive selection in both cases such that we observe only the realizations with boxes around the crosses. In Figure 1a, the selection induces a shift in the location of the observations but the slope remains the same as without selection. The reason is that the bias induced by the selection is the same for all observations (remember that all observations have the same participation probability) independently of the value of the regressor X. In Figure 1b, the bias is increasing as a function of X because the variance of the errors is increasing with X. Obviously, controlling for the selection probability does not suffice in the absence of full independence between the errors and regressors. The second insight of this figure is that the independence assumption can be tested by comparing the slopes of several quantile regression.

FIGURE 1

MEDIAN REGRESSION SLOPES UNDER INDEPENDENCE (1A) AND HETEROSCEDASTICITY (1B)



Note: Random sample of 500 observations. All observations have Z set such that $\Pr(D = 1|Z) = 0.69$.

3 Test Procedure

Our test procedure can be sketched as follows. We first estimate the selection equation using the Klein and Spady (1993) estimator for binary choice models which is asymptotically efficient in the sense that it attains the semiparametric efficiency bound. We then estimate the conditional quantile regression process by approximating the bias by a series expansion of the inverse Mill's ratio, as suggested by Buchinsky (1998a and 2001). Finally, we test whether the quantile regression slopes are homogenous over the whole conditional outcome distribution. A rejection of this null hypothesis implies a rejection of the independence assumption.

3.1 Estimation

In details, we estimate the selection equation by the semiparametric binary choice estimator suggested in Klein and Spady (1993):

$$\hat{\alpha} \equiv \max_{a \epsilon \Re} \sum \left\{ (1 - D_i) \log[1 - \hat{E}(D|Z, a)] + D_i \log[\hat{E}(D|Z, a)] \right\}, \tag{3}$$

where

$$\hat{E}(D|Z, a) = \frac{\sum_{j \neq i} D_j \kappa((Z_i' a - Z_j' a)/b_n)}{\sum_{j \neq i} \kappa((Z_i' a - Z_j' a)/b_n)}. \tag{4}$$

b_n is a bandwidth that depends on the sample size n and $\kappa(\cdot)$ denotes the kernel function, which is Epanechnikov in our case. We select the bandwidth by the generalized cross validation criterion (GCV) proposed in Craven and Wahba (1979). This estimator attains the semiparametric efficiency bound for this model. Heteroscedasticity is allowed to depend on the regressors only through the linear index. Klein and Spady's Monte Carlo simulations indicate that efficiency losses are only modest compared to probit estimation when the error distribution is standard normal, while being considerably more efficient in finite samples when the errors are non-Gaussian.

In a second step, the function $h_\tau(z'\alpha)$ is approximated by a power series expansion. The exact form of the approximation is asymptotically irrelevant. As suggested by Buchinsky (1998a), we use a power series expansion of the inverse Mill's ratio of the normalized estimated index. Thus, the first order approximation will be sufficient if the error term is normally distributed. In any case the estimator is consistent since the order of the approximation increases with the sample

size. The coefficient estimates $\hat{\beta}(\tau)$ is obtained by solving the following minimization problem:

$$\hat{\beta}\left(\tau\right),\hat{\delta}\left(\tau\right)=\min_{b,\delta}\frac{1}{n}\sum\rho_{\tau}\left(Y_{i}-X_{i}^{\prime}b-\Pi_{J}\left(Z_{i}^{\prime}\hat{\alpha}\right)\delta\right) \tag{5}$$

where $\rho_{\tau}(A)=A(\tau-1\left(A\leq0\right))$ is the check function suggested by Koenker and Bassett (1978) and $1\left(\cdot\right)$ denotes the indicator function. $\Pi_{J}\left(Z_{i}^{\prime}\hat{\alpha}\right)$ is a polynomial vector in the inverse Mill's ratio $\Pi_{J}(Z_{i}^{\prime}\hat{\alpha})=(1,\lambda(Z_{i}^{\prime}\hat{\alpha}),\lambda(Z_{i}^{\prime}\hat{\alpha})^{2},...,\lambda(Z_{i}^{\prime}\hat{\alpha})^{J})$. Again, generalized cross validation is used to determine the optimal order J.

3.2 Testing

Our null hypothesis is: $\beta\left(\tau\right)=\beta$ for $\forall\ \tau\ \epsilon\ \{0,1\}$ where $\beta\left(\tau\right)$ denotes the true τ quantile regression coefficient. Buchinsky (1998a) gives the joint asymptotic distribution of $\hat{\beta}\left(\tau\right)$ for a finite number of τ. Based on his results, we can use a finite number of quantile regressions and apply a χ^{2} test as proposed by Koenker and Bassett (1982) in the absence of sample selection. Even asymptotically, this does not allow for testing at an infinite number of τ and therefore, a χ^{2} test does not have power against all deviations from the null. Using the whole quantile process should generally entail more power and we therefore construct Kolmogorov-Smirnov or Cramer-Von-Mises-Smirnov tests. As suggested by Chernozhukov and Fernandez-Val (2005) we calculate the critical values by resampling. When computing the estimated coefficient is computationally too costly, we use score resampling as suggested by Chernozhukov and Hansen (2006). We approximate the conditional quantile process by a grid of q equidistant quantiles between zero and one, $\tau_{1:q}\ \epsilon\ \mathcal{T}\subset(0,1)$, to test the null hypothesis

$$H_{0}:\beta\left(\tau\right)=\beta,\quad\tau\ \epsilon\ \mathcal{T}. \tag{6}$$

We estimate β by the vector of median regression coefficients $\hat{\beta}\left(0.5\right)$. Alternatively, one could use a trimmed mean or the mean coefficients but this last choice requires the existence of at least the first two moments of Y given X. We measure the deviations from the null hypothesis by the Kolmogorov-Smirnov (KS) and the Cramer-Von-Mises-Smirnov (CMS) statistics for the empirical process $\hat{\beta}\left(\tau\right)-\hat{\beta}\left(0.5\right)$:

$$T_{n}^{KS}=\sup_{\tau\epsilon\mathcal{T}}\sqrt{n}||\hat{\beta}\left(\tau\right)-\hat{\beta}\left(0.5\right)||_{\hat{\Lambda}_{\tau}}\qquad\text{and}\qquad T_{n}^{CMS}=n\int_{\mathcal{T}}||\beta\left(\tau\right)-\hat{\beta}\left(0.5\right)||_{\hat{\Lambda}_{\tau}}^{2}d\tau, \tag{7}$$

where $||a||_{\hat{\Lambda}_{\tau}}$ denotes $\sqrt{a^{\prime}\hat{\Lambda}_{\tau}a}$ and $\hat{\Lambda}_{\tau}$ is a positive weighting matrix satisfying $\hat{\Lambda}_{\tau}=\Lambda_{\tau}+o_{p}(1)$, uniformly in τ. Λ_{τ} is positive definite, continuous and symmetric, again uniformly in τ. In the

empirical applications we use the inverse of the variance-covariance-matrix of X as weighting matrix. We renounce to use Anderson-Darling weights because the variance of $\hat{\beta}(\tau) - \hat{\beta}(0.5)$ converges to 0 as $\tau \to 0.5$.

Inference requires the knowledge of the asymptotic distributions of T_n^{KS}, T_n^{CMS}. Chernozhukov and Fernandez-Val (2005) show that asymptotically valid critical values can be obtained by resampling the recentered test statistics. To this end, B samples of block size m (with $m \leq n$) are drawn from the original sample with replacement to compute the inference process

$$\hat{\beta}_{m,j}(\tau) - \hat{\beta}_{m,j}(0.5), \tag{8}$$

where $1 \leq j \leq B$ and $\hat{\beta}_{m,j}(\tau)$ are the quantile slope coefficient estimates for draw j and block size m. Note that there is no statistical reason for using $m < n$ when $m = n$ is computationally feasible. The corresponding KS and CMS statistics of the recentered resampled process are

$$\begin{aligned} T_{n,m,j}^{KS} &= \sup_{\tau \epsilon \mathcal{T}} \sqrt{m} || \hat{\beta}_{m,j}(\tau) - \hat{\beta}_{m,j}(0.5) - (\hat{\beta}(\tau) - \hat{\beta}(0.5)) ||_{\hat{\Lambda}_\tau} \quad \text{and} \\ T_{n,m,j}^{CMS} &= m \int_{\mathcal{T}} || \hat{\beta}_{m,j}(\tau) - \hat{\beta}_{m,j}(0.5) - (\hat{\beta}(\tau) - \hat{\beta}(0.5)) ||_{\hat{\Lambda}_\tau}^2 d\tau. \end{aligned} \tag{9}$$

The p-value for the respective test statistic is computed as the share of $T_{n,m,j}$ being larger than T_n: $1/B \sum_{j=1}^{B} 1(T_{n,m,j} > T_n)$.

The repeated computation of the coefficients for each bootstrap sample can be quite costly, especially when the sample sizes are large. For this reason, we follow Chernozhukov and Hansen (2006) and use score resampling based on the linear approximation of the empirical processes instead, which is considerably less burdensome. In Appendix A we derive the following asymptotic linear representation

$$\sqrt{n}(\hat{\beta}(\tau) - \hat{\beta}(0.5)) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i(\tau) + o_p(1). \tag{10}$$

$s_i(\tau)$ denotes the score contribution of the ith observation at quantile τ. Again, B samples of score estimates with block size m are drawn. Let Υ_j denote a specific (sub)sample of scores, $1 \leq j \leq B$. The KS and CMS statistics for the j^{th} iteration are

$$T_{n,m,j}^{KS} \equiv \sup_{\tau \epsilon \mathcal{T}} \sqrt{m} || 1/m \sum_{i \epsilon \Upsilon_j} \hat{s}_i(\tau) ||_{\hat{\Lambda}_\tau} \qquad \textit{and} \qquad T_{n,m,j}^{CMS} \equiv m \int_{\mathcal{T}} || 1/m \sum_{i \epsilon \Upsilon_j} \hat{s}_i(\tau) ||_{\hat{\Lambda}_\tau}^2 d\tau.$$

p-values are computed analogously as outlined above.

4 Monte Carlo Simulations

In this section, we present the results of simulations. We consider a very simple data generating process:

$$\begin{aligned} D_i &= I\{X_i + W_i + U_i > 0\}, \\ Y_i &= X_i + (1 + X_i\gamma) \cdot \varepsilon_i \text{ if } D_i = 1, \\ X &\sim N(0,1), \quad W \sim N(0,1) \end{aligned} \tag{11}$$

The parameter γ controls the amount of heteroscedasticity. When $\gamma = 0$ there is independence and both mean and quantile estimators are consistent. We consider this case in the first subsection and we examine the relative efficiency of the mean and quantile estimators under different joint distributions for U and ε. In the second subsection, we analyze the size and power properties of the procedures proposed in Section 3 to test the independence assumption when $\gamma = 0$, 0.25 and 0.5.

4.1 Efficiency and robustness under independence

The need for robust statistical procedures has been stressed by many authors both in the statistical and econometric literature. This was the first motivation for considering quantile regression in Koenker and Bassett (1978). One way to measure robustness is to require that the estimators have bounded influence. Ronchetti and Trojani (2001) show that GMM are (locally) robust if and only if the orthogonality conditions are bounded. Interpreting the estimator proposed by Buchinsky (1998a) as a GMM estimator, we see that the scores given in Appendix A are bounded in the direction of ε (or Y) because Y is inside the indicator function and in the direction of U (or D) because a probability is necessarily bounded between 0 and 1. However, it is not bounded in the Z direction, which is well known for quantile regression. We consider this to be a limited problem in many applications because the support of the covariates is often bounded and outliers in Z are easier to identify. If this was not the case, a trimming function in Z could be added to the quantile objective function to obtain a fully robust estimator.

In their seminal paper on quantile regression, Koenker and Bassett (1978) provide Monte Carlo evidence on the precision of mean and quantile regression for several error distributions. They conclude that in the presence of Gaussian errors, the median estimator makes only small efficiency sacrifices compared to the mean estimator. It is, however, considerably more accurate

when errors have a non-Gaussian distribution, such as Laplace, Cauchy or contaminated Gaussian. Thus, even when errors are independent of the regressors, quantile regression can be preferable for the sake of robustness. To illustrate that such efficiency and robustness considerations also apply to sample selection models, we conduct Monte Carlo simulations. Define the following location and scale parameters

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \nu = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}. \tag{12}$$

We consider six different distributions for (U, ε). (i) Gaussian distribution: $(U, \varepsilon) \sim N(\mu, \nu)$ (ii) t-distribution with three degrees of freedom, location vector μ and scale matrix ν (iii) Cauchy distribution with location vector μ and scale matrix ν, (iv) contaminated normal errors (Gaussian mixture):

$$\begin{aligned} C &= Ber\,(0.95)\,,\ (U_1, \varepsilon_1) \sim N(\mu, \nu),\ (U_2, \varepsilon_2) \sim N(\mu, \nu) \\ U &= C \cdot U_1 + (1 - C) \cdot U_2 \\ \varepsilon &= C \cdot \varepsilon_1 + (1 - C) \cdot 10 \cdot \varepsilon_2, \end{aligned}$$

(v) contaminated data: $(U, \varepsilon) \sim N(\mu, \nu)$ but the regressor is contaminated by the factor 10 with 5% probability:

$$C = Ber\,(0.95)\,,\ X^* \sim N(0, 1),\ X = (C + (1 - C) \cdot 10) \cdot X^*.$$

For each model, 1000 Monte Carlo replications are conducted with $n = 400$ and 1600 observations. The bandwidth for the Klein and Spady (1993) estimator and the order of the power series approximation of the selection bias term are determined by GCV.

Table 1

Coefficient estimates and variances of mean and median estimators

n=400	Median estimator			Mean estimator		
Distributions	Mean	Variance	MSE	Mean	Variance	MSE
(i) Gaussian	1.009	0.013	0.013	1.003	0.009	0.009
(ii) Student's t (df=3)	1.021	0.024	0.025	1.009	0.035	0.036
(iii) Cauchy	1.044	0.089	0.091	1.356	1430.522	1430.649
(iv) Contaminated Gaussian error	0.997	0.014	0.014	0.984	0.062	0.062
(v) Contaminated data	1.105	0.016	0.027	1.909	0.176	1.002
n=1600	Median estimator			Mean estimator		
Distributions	Mean	Variance	MSE	Mean	Variance	MSE
(i) Gaussian	1.002	0.003	0.003	1.000	0.002	0.002
(ii) Student's t (df=3)	1.007	0.004	0.005	1.002	0.006	0.006
(iii) Cauchy	1.015	0.009	0.009	1.199	434.655	434.695
(iv) Contaminated Gaussian error	1.003	0.003	0.003	1.000	0.014	0.014
(v) Contaminated data	1.096	0.004	0.013	1.901	0.040	0.851

The mean, variance and mean squared errors (MSEs) of the median and mean[7] coefficients are reported in Table 1. Apart from specification (i), where the error terms are jointly normally distributed, the median estimator is more precise and has a smaller MSE than the mean estimator. When the error terms are Cauchy distributed, the moments of the mean estimator do not exist whereas the bias and the variance of the median estimator are relatively well-behaved. In presence of contaminated date, the mean estimator is severely upward biased and quite noisy, whereas the median estimator is only slightly biased and very precise.

4.2 Power and size properties of the independence tests

In this section, we present Monte Carlo evidence about the size and power properties of the independence test that we have proposed in Section 3. We use the same data generating process as above, which is defined in display (11). We consider three distributions for (U, ε): Gaussian, $t\,(3)$ and $t\,(1)$. The location and scale of these distributions are set to μ and ν defined in (12). We consider three values for the critical parameter γ. Under he null hypothesis (independence)

[7] Analogous to the estimation of the quantile coefficients, the mean coefficients are estimated following the two step procedure of Newey (2009).

the regressor X has a pure location shift effect; this corresponds to $\gamma = 0$. This case allows us to analyze the empirical size of our tests. We also evaluate the power of our tests in two location scale shift models ($\gamma = 0.2,\ 0.5$).

As above, the bandwidth for the Klein and Spady (1993) estimator and the order of the power series approximation of the selection bias term are determined by GCV. We present the results of the score bootstrap tests based on the Kolmogorov-Smirnov (KS) and Cramer-Von-Mises-Smirnov (CMS) statistics. In order to construct the test statistics, the coefficients $\hat{\beta}(\tau)$ are estimated at equidistant quantiles with step size 0.01 and compared to the median estimate $\hat{\beta}(0.5)$. Results are presented for three different regions of τ over which the quantile coefficients are estimated: $[0.05, 0.95]$, $[0.1, 0.9]$, and $[0.2, 0.8]$.

In the simulations, we consider five sample sizes from $n = 100$ to $n = 3200$. We run 1000 Monte Carlo replications and draw 250 bootstrap samples within each replication. The theoretical level of significance is set at 5%. For the sake of brevity, we only report the rejection frequencies for the bootstrap, i.e., for the block size $m = n$. The results for subsampling (i.e., for some m smaller than n) are comparable and available from the authors upon request.

Table 2

Empirical rejection frequencies for 5% bootstrap tests

Normal distribution, 1000 replications, 250 bootstrap draws

	Kolmogorov-Smirnov statistics								
	$\gamma = 0$			$\gamma = 0.2$			$\gamma = 0.5$		
$\tau \in$	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
$n = 100$	0.003	0.000	0.000	0.005	0.001	0.001	0.021	0.004	0.001
$n = 400$	0.022	0.006	0.002	0.273	0.145	0.054	0.895	0.780	0.434
$n = 800$	0.024	0.019	0.009	0.636	0.482	0.253	0.996	0.994	0.952
$n = 1600$	0.038	0.033	0.017	0.934	0.885	0.705	1.000	1.000	1.000
$n = 3200$	0.042	0.029	0.027	0.999	1.000	0.975	1.000	1.000	1.000
	Cramer-Von-Mises-Smirnov statistics								
	$\gamma = 0$			$\gamma = 0.2$			$\gamma = 0.5$		
$\tau \in$	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
$n = 100$	0.001	0.000	0.000	0.001	0.000	0.000	0.006	0.005	0.003
$n = 400$	0.011	0.006	0.002	0.192	0.112	0.045	0.924	0.838	0.498
$n = 800$	0.017	0.009	0.010	0.617	0.463	0.252	0.999	0.998	0.957
$n = 1600$	0.026	0.020	0.014	0.958	0.911	0.735	1.000	1.000	1.000
$n = 3200$	0.026	0.024	0.021	1.000	0.999	0.978	1.000	1.000	1.000

The empirical rejection frequencies reported in Table 2 suggest that the bootstrap score tests have good size and power properties with normally distributed error terms. In the presence of independent errors ($\gamma = 0$), both the KS and CMS tests are conservative, at least for the sample sizes considered. However, the empirical size slowly converges to the theoretical size of 5% as the sample size increases. The KS test does so at a faster pace than the CMS test. Under heteroscedastic errors, the rejection probabilities correctly converge to 100% as n becomes larger. As expected, this happens at a faster pace for $\gamma = 0.5$ than for $\gamma = 0.2$. The power properties of the KS and CMS tests are rather similar, albeit the latter become relatively more powerful in larger samples and/or for a higher γ. The empirical power increases as the range of quantiles considered increases and this holds true for both test statistics and both values of γ. Summing up, the KS and CMS tests seem to perform well in finite samples with Gaussian errors. Under sample sizes of several thousand observations, they are powerful in any scenario considered.

Table 3 reports the rejection frequencies for $t\,(3)$-distributed error terms: $(U, \varepsilon) \sim t\,(3, \mu, \nu)$. As one would expect, deviations from the null hypothesis are harder to detect because of the fatter

tails. The KS test statistic suffers particularly because it is more likely to confound a single outlier with a deviation from H_0. Accordingly, rejection frequencies of the KS test 'overshoot' in small samples when the range of quantiles used is too large. Even in this case, the empirical size converges eventually to the true value. The CMS test performs better as it stays on the 'safe side' for all ranges of quantiles and is more conservative than the than the theoretical rate of 5%. Furthermore, the CMS rejection frequencies converge faster to 100% when the errors are heteroscedastic.

TABLE 3

EMPIRICAL REJECTION FREQUENCIES FOR 5% RESAMPLING TESTS

$t\,(3)$ DISTRIBUTION, 250 BOOTSTRAP DRAWS, 1000 REPLICATIONS

	Kolmogorov-Smirnov statistics								
	$\gamma = 0$			$\gamma = 0.2$			$\gamma = 0.5$		
τ ϵ	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
$n = 100$	0.037	0.006	0.000	0.041	0.005	0.001	0.064	0.006	0.001
$n = 400$	0.070	0.032	0.009	0.252	0.153	0.043	0.801	0.758	0.415
$n = 800$	0.103	0.068	0.013	0.447	0.414	0.253	0.966	0.986	0.943
$n = 1600$	0.066	0.061	0.026	0.642	0.743	0.632	0.999	1.000	1.000
$n = 3200$	0.068	0.055	0.045	0.909	0.969	0.945	1.000	1.000	1.000
	Cramer-Von-Mises-Smirnov statistics								
	$\gamma = 0$			$\gamma = 0.2$			$\gamma = 0.5$		
τ ϵ	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
$n = 100$	0.000	0.000	0.000	0.001	0.000	0.000	0.014	0.003	0.001
$n = 400$	0.020	0.007	0.001	0.205	0.104	0.019	0.890	0.800	0.462
$n = 800$	0.053	0.017	0.004	0.581	0.462	0.230	1.000	0.998	0.960
$n = 1600$	0.050	0.034	0.020	0.896	0.862	0.665	1.000	1.000	1.000
$n = 3200$	0.049	0.036	0.036	0.995	0.998	0.976	1.000	1.000	1.000

Table 4 displays the rejection rates for Cauchy distributed error terms. Since the tails are extremely fat, larger sample sizes are required to obtain satisfactory results. The KS is, as expected, more sensitive to outliers and performs less well than the CMS test statistics. In contrast to Gaussian errors, $\mathcal{T} = [0.05, 0.95]$ is generally not the best choice with worse size and power properties. The narrowest range, $\mathcal{T} = [0.2, 0.8]$ yields the best results because it does not use the uninformative tails of the Cauchy. The differences between the results for different

distributions show that none of the ranges or test statistics is uniformly more powerful. For well-behaved distributions, using a wide range of quantiles and the KS statistic yields better results. For fatter tailed distributions, CVM applied to a narrower range of quantiles is preferable. This suggests that an applied researcher should choose regions $\mathcal{T}$ that are not too close to the boundaries if she suspects the error distribution to have fat tails. Given the uncertainty about the shape of the distribution, it may also be beneficial to report the results of several tests.

TABLE 4

EMPIRICAL REJECTION FREQUENCIES FOR 5% RESAMPLING TESTS

CAUCHY DISTRIBUTION, 250 BOOTSTRAP DRAWS, 1000 REPLICATIONS

	Kolmogorov-Smirnov statistics								
	$\gamma = 0$			$\gamma = 0.2$			$\gamma = 0.5$		
τ ϵ	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
$n = 100$	0.088	0.018	0.004	0.088	0.025	0.003	0.113	0.030	0.006
$n = 400$	0.091	0.056	0.018	0.103	0.084	0.030	0.285	0.333	0.270
$n = 800$	0.086	0.063	0.036	0.097	0.076	0.072	0.433	0.553	0.661
$n = 1600$	0.069	0.031	0.023	0.132	0.144	0.220	0.704	0.881	0.963
$n = 3200$	0.079	0.041	0.035	0.279	0.325	0.472	0.926	0.989	0.999
	Cramer-Von-Mises-Smirnov statistics								
	$\gamma = 0$			$\gamma = 0.2$			$\gamma = 0.5$		
τ ϵ	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
$n = 100$	0.053	0.005	0.000	0.057	0.007	0.000	0.066	0.013	0.001
$n = 400$	0.072	0.036	0.004	0.090	0.066	0.013	0.395	0.441	0.254
$n = 800$	0.083	0.040	0.016	0.110	0.092	0.063	0.625	0.773	0.812
$n = 1600$	0.052	0.033	0.014	0.174	0.245	0.292	0.888	0.980	0.993
$n = 3200$	0.061	0.031	0.025	0.385	0.525	0.668	0.994	1.000	1.000

Before ending this section, we investigate whether the violation of the null hypothesis actually biases the estimators. If this was not the case, one would not be too worried about the rejection of the independence assumption. Table 5 reports the mean, variance and MSE of the median and mean regression estimator for $n = 1600$ and for the different scenarios. Biases are not negligible under heteroscedasticity ($\gamma = 0.2$, 0.5) and MSEs are largely driven by these biases, such that these rejections have to be taken seriously at least for the DGPs considered.

Table 5

Coefficient estimates and variances of mean and median estimators

n=1600	Median estimator			Mean estimator		
Distributions	Mean	Variance	MSE	Mean	Variance	MSE
Normal, $\gamma = 0$	1.002	0.003	0.003	1.000	0.002	0.002
Normal, $\gamma = 0.2$	1.076	0.003	0.009	1.067	0.003	0.007
Normal, $\gamma = 0.5$	1.208	0.004	0.047	1.174	0.004	0.035
Student's t (df=3), $\gamma = 0$	1.007	0.004	0.005	1.002	0.006	0.006
Student's t (df=3), $\gamma = 0.2$	1.101	0.005	0.015	1.123	0.007	0.023
Student's t (df=3), $\gamma = 0.5$	1.258	0.005	0.072	1.305	0.011	0.104
Cauchy, $\gamma = 0$	1.015	0.009	0.009	1.199	434.655	434.695
Cauchy, $\gamma = 0.2$	1.144	0.010	0.030	2.353	503.527	505.357
Cauchy, $\gamma = 0.5$	1.340	0.009	0.124	4.082	673.510	683.008

Note: 1000 Monte Carlo replications. The true value is 1.

5 Labor Market Applications

5.1 Female wage distribution in Portugal

In this section we apply our tests proposed in Section 3 to empirical data. The first one uses female labor market data from Portugal previously analyzed by Martins (2001), who compared parametric and semiparametric estimators. The sample stems from the 1991 wave of the Portuguese Employment Survey and consists of 2,339 married women aged below 60 whose husbands earned labor income in 1991. The data contain information on the wages and hours worked for 428 women with positive labor supply along with a set of regressors for the whole sample. We observe the outcome hourly wage only for those 1,400 women who participate in the labor market, whereas explanatory variables are observed for the entire sample. In the test procedures, we use the same model specification as in Martins (2001). The regressors (X) in the wage equation include (potential experience)/10, (potential experience)^2/100, and the interactions of both terms with the number of children. The variables (Z) characterizing labor market participation contain age/10, age^2/100, years of education, the number of children under 18, the number of children under 3, and the log of the husband's monthly wage.

Table 6

Labor market application I: p-values

$\tau \in$	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
KS	0.044	0.005	0.011
CMS	0.015	0.032	0.306

Note: 10,000 bootstrap draws.

Table 6 reports the p-values of the KS and CMS tests for $B = 10,000$ bootstrap replications. We consider three ranges of quantiles for τ, $[0.05, 0.95]$, $[0.1, 0.9]$ and $[0.2, 0.8]$, with steps of 0.01. GCV is applied to determine the optimal bandwidth b_n^{opt} in (4) and the optimal order J.[8] At the 5% significance level all tests reject the independence assumption with the exception of the CMS test statistic applied the narrowest range. To better understand this result the upper part of Figure 2 shows the quantile coefficients on experience and experience squared with 95% pointwise confidence intervals.[9] For these two variables the coefficients show a U-shape and inverted U-shape pattern as a function of the quantile. This explains why the CMS test is not able to detect heterogeneity when we consider only the 60% in the middle of the distribution.

5.2 Female wage distribution in the USA

While the first application shows that our test can be quite powerful even in medium-size samples (428 observed wages), considerably larger data sets are available for our second application. In their study on US women's relative wages, Mulligan and Rubinstein (2008) estimate the conditional mean wages of married white women using a normal parametric correction for sample selection. They investigate two repeated cross-sections covering the periods 1975-1979 and 1995-1999 in the US Current Population Survey (CPS) and restrict the data to married white females aged 25-54. The outcome variable is the female's log weekly wage. Labor market participation (D) is defined as working full time and at least 50 weeks in the respective year. The 1975-79 period contains 116,843 observations, of whom 36,817 report to work full time. For the 1995-99 period, the respective numbers are 102,395 and 52,242. The regressors X include wife's education (dummies for 8 or less years of schooling, 9-11 years of schooling, high school graduate, college

[8] $b_n^{\text{opt}} = 0.14$, $J = 4$

[9] The other coefficients are available from the authors upon request.

graduate, advanced degree), potential work experience as well as squared, cubic, and quartic terms thereof, marital status, regional dummies, and interactions between all variables capturing potential experience and education. Z contains X as well as the number of children aged 0-6 and its interactions with the marital status.

TABLE 7

LABOR MARKET APPLICATION II: P-VALUES

	1975-1979		
$\tau \in$	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
KS	0.001	0.008	0.002
CMS	0.000	0.000	0.005
	1995-1999		
$\tau \in$	[0.05,0.95]	[0.1,0.9]	[0.2,0.8]
KS	0.000	0.000	0.000
CMS	0.000	0.000	0.000

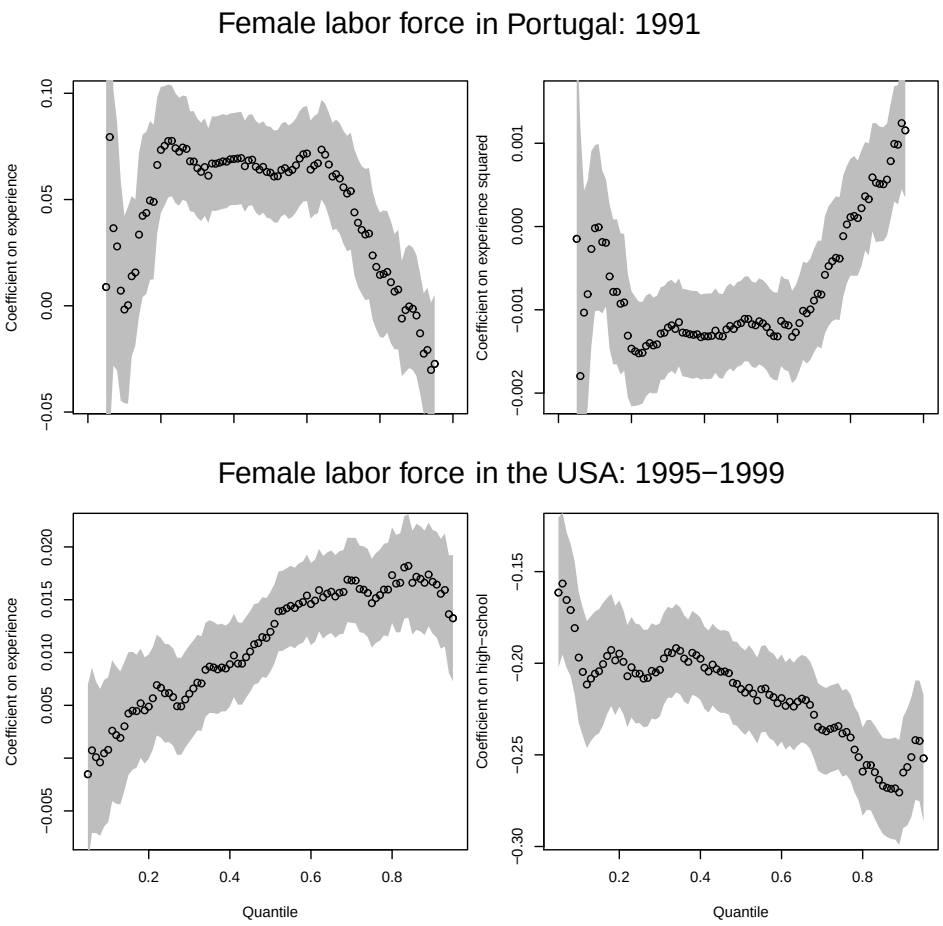
Note: 1000 bootstrap draws.

Table 7 reports the p-value of our tests for the same range of quantiles than for the first application.[10] 1000 bootstrap draws of the scores were sampled. The null hypothesis is rejected at the 1% significance level for both years, both test statistics and all ranges. The low p-values leave little doubt about the violation of the independence assumption. One may argue that small deviations from the null hypothesis may lead to a rejection in so large samples. However, the quantile coefficients plotted in the lower half of Figure 2 show that there are important and systematic deviation from the null hypothesis for economically important regressors. This means that there is at least the potential for an economically significant bias. In addition, since the selection probability is changing over time, the researchers have to be extremely careful when they compare the wage functions between the 70s and the 90s. The apparent differences may be the results of a changing bias.

[10] b_n^{opt} is 0.15 for 1975-1979 and 1995-1999. J is 5 and 6 for 1975-1979 and 1995-1999, respectively.

FIGURE 2

QUANTILE REGRESSION COEFFICIENTS ON SELECTED VARIABLES



Note: The coefficients have been estimated using the estimator suggested by Buchinsky (1998a). The samples, the other variables and the details of the implementation are described in the text. 95% pointwise confidence intervals are also plotted. Caution: Given the result of our test, these results are not consistent for the true parameters.

6 Conclusion

Assuming additivity and independence of the error term in the outcome equation is rather restrictive. It implies that all units with the same observable variables react to changes in the latter in the same way. However, the unobservable random terms may have important economic interpretations. The recent econometric literature has considerably relaxed restrictions on the in-

teraction of observables and unobservables. Advances have been reported in models based selection on observables, instrumental variables, and panel data, among many others, see for instance Matzkin (2007) for a discussion.

Somewhat surprisingly, the sample selection model has been excluded from this trend. Almost all sample selection estimators still assume that the error terms are independent.[11] This is also the case in the quantile regression model of Buchinsky (1998a). However, in the quantile regression framework the independence assumption implies that the quantile slope coefficients are equal to the mean slope coefficients and all quantile curves are parallel. In other words, the heterogeneity that we want to analyze is excluded by assumption. Applications of the sample selection correction for quantile regression that have found significant differences between the coefficients estimated at distinct quantiles have merely proven the violation of the underlying assumptions and the inconsistency of the estimator.[12]

Given the importance of the independence assumption for the identification of sample selection models, this assumption should be tested whenever this is possible. In this paper we propose the (to the best of our knowledge) first formal test for this assumption. Our method is based on the quantile estimator of Buchinsky (1998a), which is consistent under the null hypothesis, and compares the coefficients obtained at different quantiles. It is relevant for the consistency of both mean and quantile regression. Monte Carlo simulations provide evidence on the satisfactory power and size properties of our test procedures. We also present two applications to representative labor market data. The results foster the suspicion that the independence assumption may be violated in many empirical problems.

The question that naturally follows is: What can be done in the case of the rejection of this critical assumption? In our companion paper, Melly and Huber (2011), we derive the sharp bounds on the quantile regression parameters when the independence assumption (separability) is no longer imposed. It appears that point identification can be attained only

[11] In addition to the papers discussed below, Newey (2007) is a notable exception. However, he is interested in outcomes defined in the selected population and not in the whole population.

[12] Restrictions similar to the independence assumption also appear in some instrumental variable models, see for example Amemiya (1982), Powell (1983), Chen and Portnoy (1996), Lee (2007), Blundell and Powell (2007), and Carneiro and Lee (2009). These restrictions are particularly useful to justify a control function or a fitted value approach in order to tackle endogeneity problems. Also in these models, this assumption implies that the coefficients do not vary across quantiles. Therefore, these estimators are not useful for analyzing heterogeneity (which was *not* the intention of their authors).

by an identification at infinity argument or by a parametric assumption. In the absence of observations that are observed with probability one, only the second strategy can help recovering point identification. Donald (1995) and Chen and Khan (2003) make one step away from independence and allow for multiplicative heteroscedasticity. Donald (1995) identifies the model by a normality assumption while Chen and Khan (2003) use de facto an exclusion restriction for the conditional variance (see Appendix B). Arellano and Bonhomme (2010) obtain point identification by a clever parametrization of the copula between both error terms (selection and outcome equations) while keeping the marginal distributions of the error terms nonparametric. This weaker parametric restriction does not restrict the relationship of the outcome and the observables.

Another strategy consists in changing the estimand by considering a different population. Newey (2007) analyzes a nonseparable model but shows identification only in the selected population instead of the entire population. In the absence of an exclusion restriction, Lee (2009), Lechner and Melly (2010), and Huber and Mellace (2010) provide sharp bounds for several subpopulations. This is of interest in some applications but clearly not in all. For instance, it does not allow the researcher to compare female and male wages or wages across different years.

A Appendix A: Score function

In this appendix, we derive the score function used for the test proposed in Section 3. We first consider the estimator of the selection probability. Define the short-hand notation $P_i(a) = \Pr(D = 1|Z_i'a)$ (where a denotes the vector of first stage regressors the true value of which is α). Klein and Spady (1993) show in their equation (49) that, under some regularity conditions,

$$\sqrt{n}(\hat{\alpha} - \alpha) = \Delta_P^{-1} \sum_{i=1}^{n} K_i + o_p(1), \tag{A-1}$$

where

$$\Delta_P = \mathrm{E}\left[\left.\frac{\partial P_i(a)}{\partial a}\right|_{a=\alpha} \left.\frac{\partial P_i(a)'}{\partial a}\right|_{a=\alpha} \frac{1}{P(\alpha)(1-P(\alpha))} \right] \tag{A-2}$$

$$\text{and } K_i = \left.\frac{\partial P_i(a)}{\partial a}\right|_{a=\alpha} \cdot \frac{D_i - P_i(\alpha)}{P_i(\alpha)(1-P_i(\alpha))}. \tag{A-3}$$

We heavily draw from the appendix in Buchinsky (1998a) to derive the score function for the second step quantile estimator. $\hat{\beta}(\tau)$ solves the moment condition for the τth quantile regression:

$$\Psi(Z, Y, D, a, b) = D[\tau - I\left\{Y < X'b - h_\tau(Z'a)\right\}]X.$$

Following the arguments in Buchinsky (1998a), we can combine his equations (A3) to (A8) with the simplications in (A14) to (A16) to obtain the following representation:

$$\sqrt{n}(\hat{\beta}(\tau) - \beta) = \Delta_{b,\tau}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\ell_i(\tau) - \Delta_{a,\tau}\sqrt{n}(\hat{\alpha} - \alpha)) + o_p(1). \tag{A-4}$$

$\Delta_{b,\tau}$ is the derivative of the expected value of $\Psi(Z, Y, D, a, b)$ with respect to b and $\Delta_{a,\tau}$ is the derivative of the same expected value with respect to a. Precisely,

$$\begin{aligned}
M_i &= D_i\left(X_i - E\left[X_i|Z_i'a\right]\right), \\
\Delta_{b,\tau} &= E[f_{\varepsilon(\tau)}(0|Z_i'a))M_iM_i'], \\
\Delta_{a,\tau} &= E[f_{\varepsilon(\tau)}(0|Z_i'a)M_i\left(\frac{\partial h(Z'a)}{\partial a}\right)'], \\
\ell_i(\tau) &= (\tau - I\{Y_i < X_i'\beta(\tau) + h_\tau(Z_i'\alpha)\})M_i.
\end{aligned}$$

We now insert (A-1) into (A-4):

$$\begin{aligned}
\sqrt{n}(\hat{\beta}(\tau) - \beta) &= \Delta_{b,\tau}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\ell_i(\tau) - \Delta_{a,\tau}\Delta_P^{-1}K_i) + o_p(1) \\
&\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_i(\tau) + o_p(1).
\end{aligned}$$

Our test statistics exploit the differences between $\hat{\beta}(\tau)$ and $\hat{\beta}(0.5)$. Therefore, the test's score function $s_i(\tau)$ is obtained by subtracting one score from the other. Thus,

$$\sqrt{n}(\hat{\beta}(\tau) - \hat{\beta}(0.5)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i(\tau) + o_p(1),$$

where

$$s_i(\tau) = A_i(\tau) - A_i(0.5).$$

B Appendix B: Chen and Khan (2003)

Chen and Khan (2003) discuss the estimation of sample selection models subject to conditional heteroscedasticity in both the selection and outcome equations. They consider a model similar to that of Donald (1995) but relax the normality assumption on the errors. They propose a three-step estimator and show that it is $\sqrt{n}$ consistent.

In terms of our notation, their model is defined as follows:

$$\begin{aligned} D_i &= I\{\mu(Z_i) - \sigma_1(Z_i) \cdot \upsilon_i \geq 0\}, \\ Y_i^* &= X_i'\beta + \sigma_2(X_i) \cdot \epsilon_i, \\ Y_i &= Y_i^* \text{ if } D_i = 1, \end{aligned}$$

where β are the parameters of interest, X_i, Z_i and D_i are observed, $\mu(Z_i)$, $\sigma_1(Z_i)$ and $\sigma_2(X_i)$ are unknown functions, and υ_i and ϵ_i are unobserved disturbances, which are independent of the regressors but not necessarily of each other.

They show (equation 2.13) that

$$F_{Y_i^*}^{-1}(\tau | Z_i, D_i = 1) = X_i'\beta + \sigma_2(X_i)\lambda_\tau(P_i),$$

where λ_τ is an unknown function (different at each quantile) and $P_i = \Pr(D_i = 1 | Z_i)$. This implies that the inter-quartile range is

$$\begin{aligned} \Delta Q(Z_i) &= F_{0.75}^{-1}(\tau | Z_i, D_i = 1) - F_{0.25}^{-1}(\tau | Z_i, D_i = 1) \\ &= \sigma_2(X_i)(\lambda_{0.75}(P_i) - \lambda_{0.25}(P_i)) \equiv \sigma_2(X_i)\Delta\lambda(P_i). \end{aligned}$$

The more conventional selection correction equation is given by

$$E[Y_i^* | Z_i, D_i = 1] = X_i'\beta + \sigma_2(X_i)\lambda(P_i),$$

where $\lambda(P_i) = E[\epsilon_i | D_i = 1, Z_i]$. We now define the transformed variables

$$\tilde{Y}_i = \frac{Y_i}{\Delta Q(Z_i)}, \tilde{X}_i = \frac{X_i}{\Delta Q(Z_i)}, \tilde{\lambda}(P_i) = \frac{\lambda(P_i)}{\Delta\lambda(P_i)}$$

and obtain the new selection correction equation

$$E\left[\tilde{Y}_i^* | Z_i, D_i = 1\right] = \tilde{X}_i'\beta + \tilde{\lambda}(P_i).$$

This looks like the partial linear form of the conditional expectation function in the homoscedastic sample selection model. Chen and Khan (2003) propose to use the same kernel procedure as Ahn and Powell (1993) to estimate β.

In their regularity assumption I, Chen and Khan (2003) directly assume identification of the parameters of interest. Here, we show that this assumption excludes the simplest case of linear multiplicative heteroscedasticity:

$$\sigma_2(X_i) = X_i'\gamma.$$

In this case

$$\tilde{X}_i = \frac{X_i}{\Delta Q(Z_i)} = \frac{X_i}{X_i'\gamma\Delta\lambda(P_i)},$$

such that $\tilde{X}_i\gamma = \frac{1}{\Delta\lambda(P_i)}$. Since we have to condition on P_i in order to control for selection bias, this implies that the transformed regressors $\tilde{X}_i$ are multicollinear given the propensity score. In other words, the parameters β are not identified. Identification of β when $\sigma_2(X_i)$ is linear requires a new type of exclusion restrictions: a variable that affects the conditional variance but has no effect on the conditional mean of Y. If $\sigma_2(X_i)$ is nonlinear, the model is identified without exclusion restriction with $\sigma_2(X_i)$ minus its linear projection on X_i serving as excluded regressor.

Ahn and Powell (1993) make a similar assumption in a homoscedastic sample selection model ($\sigma_2(X_i) = 0$). They note that Z must include a variable excluded from X if $\mu(Z)$ is linear in Z. If $\mu(Z)$ is nonlinear, then the model is identified without exclusion restriction with $\mu(X)$ minus its linear projection on X serving as the excluded regressor. This is very similar to the identification of the model in Chen and Khan (2003).

References

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

Amemiya, T. (1982): "Two Stage Least Absolute Deviations Estimators," *Econometrica*, 50, 689–711.

Angrist, J. (1997): "Conditional Independence in Sample Selection Models," *Economics Letters*, 54, 103–112.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563.

Arellano, M., and S. Bonhomme (2010): "Quantile selection models," *unpublished manuscript*.

Blundell, R. W., and J. L. Powell (2007): "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*, 141(1), 65–83.

Buchinsky, M. (1994): "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica*, 62, 405–458.

——— (1998a): "The dynamics of changes in the female wage distribution in the USA: A quantile regression approach," *Journal of Applied Econometrics*, 13, 1–30.

——— (1998b): "Recent advances in quantile regression models: A practical guidline for empirical research," *The Journal of Human Resources*, 33, 88–126.

——— (2001): "Quantile regression with sample selection: Estimating women's return to education in the U.S.," *Empirical Economics*, 26, 87–113.

Carneiro, P., and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149(2), 191–208.

Chaudhuri, P. (1991): "Global nonparametric estimation of conditional quantile functions and their derivatives," *Journal of Multivariate Analysis*, 39, 246–269.

Chen, L., and S. Portnoy (1996): "Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equations models," *Comm. Statist. Theory Methods*, 25, 1005–1032.

Chen, S., and S. Khan (2003): "Semiparametric estimation of a heteroskedastic sample selection model," *Econometric Theory*, 19, 1040–1064.

Chernozhukov, V., and I. Fernandez-Val (2005): "Subsampling inference on quantile regression processes," *Sankhya: The Indian Journal of Statistics*, 67, 253–276.

Chernozhukov, V., and C. Hansen (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.

Cosslett, S. (1991): "Distribution-Free Estimator of a Regression Model with Sample Selectivity," in *Nonparametric and semiparametric methods in econometrics and statistics*, ed. by W. Barnett, J. Powell, and G. Tauchen, pp. 175–198. Cambridge University Press, Camdridge, UK.

Craven, P., and G. Wahba (1979): "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, 31, 377–403.

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

Donald, S. G. (1995): "Two-step estimation of heteroskedastic sample selection models," *Journal of Econometrics*, 65, 347–380.

Gallant, A., and D. Nychka (1987): "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

Gronau, R. (1974): "Wage comparisons-a selectivity bias," *Journal of Political Economy*, 82, 1119–1143.

Guntenbrunner, C., and J. Jurečková (1992): "Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics," *Annals of Statistics*, 20, 305–330.

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

——— (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

Huber, M., and G. Mellace (2010): "Sharp bounds on causal effects under sample selection," *mimeo*.

Klein, R. W., and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

Koenker, R. (2005): *Quantile Regression*. Cambridge University Press.

Koenker, R., and G. Bassett (1978): "Regression quantiles," *Econometrica*, 46, 33–50.

——— (1982): "Robust tests for heteroskedasticity based on regression quantiles," *Econometrica*, 50, 43–62.

Koenker, R., and K. F. Hallock (2001): "Quantile regression," *Journal of Economic Perspectives*, 15, 143–156.

Koenker, R., and Z. Xiao (2002): "Inference on the Quantile Regression Process," *Econometrica*, 70, 1583–1612.

Lechner, M., and B. Melly (2010): "Partial Identification of Wage Effects of Training Programs," *Brown University working paper*.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Lee, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141, 1131–1158.

Martins, M. (2001): "Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal," *Journal of Applied Econometrics*, 16, 23–39.

Matzkin, R. (2007): "Nonparametric Identification," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, pp. 5307–5368. Elsevier, Amsterdam.

Melly, B., and M. Huber (2011): "Nonseparable sample selection models," *unpublished manuscript*.

Mincer, J. (1973): *Schooling, Experience, and Earnings*. NBER, New York.

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

Newey, W. K. (2007): "Nonparametric continuous/discrete choice models," *International Economic Review*, 48, 1429–1439.

——— (2009): "Two-step series estimation of sample selection models," *Econometrics Journal*.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Powell, J. (1983): "The asymptotic normality of two-stage least absolute deviations estimators," *Econometrica*, 51, 1569–1575.

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

——— (1987): "Semiparametric Estimation of Bivariate Latent Variable Models," *unpublished manuscript*, University of Wisconsin-Madison.

Ronchetti, E., and F. Trojani (2001): "Robust inference with GMM estimators," *Journal of Econometrics*, 101, 37–69.