A TEST OF THE CONDITIONAL INDEPENDENCE ASSUMPTION IN SAMPLE SELECTION MODELS

MARTIN HUBER[a] AND BLAISE MELLY[b]*

[a] *Department of Economics, University of Fribourg, Switzerland*
[b] *Department of Economics, Bern University, Switzerland*

SUMMARY

Identification in most sample selection models depends on the independence of the regressors and the error terms conditional on the selection probability. All quantile and mean functions are parallel in these models; this implies that quantile estimators cannot reveal any—per assumption non-existing—heterogeneity. Quantile estimators are nevertheless useful for testing the conditional independence assumption because they are consistent under the null hypothesis. We propose tests of the Kolmogorov–Smirnov type based on the conditional quantile regression process. Monte Carlo simulations show that their size is satisfactory and their power sufficient to detect deviations under plausible data-generating processes. We apply our procedures to female wage data from the 2011 Current Population Survey and show that homogeneity is clearly rejected. Copyright © 2014 John Wiley & Sons, Ltd.

1. INTRODUCTION

Estimation of economic models is frequently complicated by the problem of sample selection: the variable of interest is only observed for a non-random subsample of the population. A prominent example in labor economics is the estimation of the determinants of female wages. Individuals are assumed to offer a positive labor supply only if their potential wage exceeds their reservation wage. Starting with Gronau (1974) and Heckman (1974, 1976, 1979), the last 40 years have seen a proliferation of work addressing this difficult problem. Although the suggested estimators have progressively weakened the distributional and parametric assumptions originally made, none of them is robust to the presence of heteroscedasticity or higher-order dependence between the error terms and the covariates (after controlling for the selection probability).

However, dependence in general—and heteroscedasticity in particular—is ubiquitous in the fields where sample selection models are applied. 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 quantile regression applications in labor economics find significant heterogeneity in the returns to education and experience. Thus the conditional independence assumption has neither theoretical nor empirical support in most economic applications. In this paper we discuss the importance of this assumption in sample selection models and the implication for quantile functions, and we suggest a test of its validity.

Our test exploits the fact that, under conditional independence, all quantile (and mean) regression functions are parallel after controlling for sample selection. Thus we can test the conditional independence assumption by comparing the slope coefficients at different quantiles. This approach is in the spirit of Koenker and Bassett (1982), who apply quantile regression for testing heteroscedasticity in a linear model without sample selection. In the absence of sample selection, the lack of independence between the error terms and the regressors does not affect the consistency of the estimators; it is only a minor nuisance for inference. In the presence of sample selection, the conditional independence

---

* Correspondence to: Blaise Melly, Department of Economics, Bern University, Schanzeneckstrasse 1, CH-3001 Bern, Switzerland. E-mail: blaise.melly@vwi.unibe.ch

assumption is an identifying assumption; its violation leads to the inconsistency of most sample selection estimators (including those only concerned with mean effects). To the best of our knowledge, ours is the first test of this identifying assumption.

To implement this testing idea we use the sample selection correction for quantile regression proposed by Buchinsky (1998a, 2001). These papers extended the series estimator of Newey (2009) to the estimation of quantiles. The estimator has been applied to many data—among others in both original articles—to analyze the heterogeneous effects of the explanatory variables on the distribution of the outcome. Another contribution of this paper is to show that this estimator is consistent only in two cases: when all quantile regression slopes are equal or when selection is random. The reason is that Buchinsky (1998a) assumes that the error terms are independent of the regressors given the selection probability. This implies that all quantile slope coefficients and the mean coefficients are identical; i.e. it excludes heterogeneous effects even though their analysis has been the main motivation for using quantile regression in recent applications.

The Buchinsky estimator nevertheless remains useful for two reasons, which were the initial motivations for quantile regression: it is robust and can be used as the basis for a test of independence. Koenker and Bassett (1978) also assume independence in their seminal paper and motivate quantile regression purely from a robustness and efficiency point of view in the presence of non-Gaussian errors. Similarly, the estimator of Buchinsky has a bounded influence function (in the direction of the dependent variable), so that it is more robust than mean regression. Under the null hypothesis of conditional independence the slope coefficients are constant as a function of the quantile, and the procedure proposed by Buchinsky (1998a) consistently estimates them. Under the alternative hypothesis, the estimates do not converge to the true values but will be a non-trivial function of the quantile. These two properties justify testing the conditional independence assumption by testing whether the coefficients vary across quantiles.

In order to detect any violation of the null hypothesis, our testing approach is based on the Kolmogorov–Smirnov (KS) and Cramér–von Mises (CM) statistics, applied to the entire conditional quantile process (after trimming the tails and conditioning on the selection probability). The presence of an unknown nuisance parameter under the null hypothesis jeopardizes the nuisance-free asymptotic distribution of these statistics; Koenker and Xiao (2002) call this situation the Durbin problem. Chernozhukov and Fernández-Val (2005) overcome this complication by estimating the critical values using resampling procedures. More precisely, they obtain the critical values by subsampling the KS and CMS statistics applied on the appropriately recentered empirical quantile process. This approach requires recomputing the whole quantile process for all subsamples, which is feasible for standard quantile regression but too demanding for instrumental variable (IV) quantile regression or the sample selection correction for quantile regression proposed by Buchinsky (1998a). For this reason, Chernozhukov and Hansen (2006) resample the influence functions instead of recomputing the whole process for all draws. They show that this procedure provides asymptotically valid critical values under some high-level conditions. They also show that these conditions are satisfied for standard and IV quantile regression; we check their conditions for our sample selection model. Monte Carlo simulations calibrated to fit characteristics of typical applications show that this procedure performs well in reasonably sized samples.

Finally, we apply our tests to examine whether violations of the conditional independence assumption are an empirically relevant phenomenon. We use data from the 2011 outgoing rotation groups of the Current Population Survey (CPS) to estimate a traditional Mincer wage regression for women. The number of children (measured in three age brackets) is used as an instrument for selection and excluded from the outcome equation. This is a typical application of sample selection models; a prominent recent example is given by Mulligan and Rubinstein (2008). We reject the conditional independence assumption at any conventional significance level. The working paper version of this article (Huber and Melly, 2011) contains two other applications with similar conclusions. Therefore, we

suspect that this problem is not limited to a few cases but is widespread in fields where sample selection models have been used.[1]

This paper contributes to the literature on sample selection and quantile regression. 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 may yield 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. Finally, Das *et al.* (2003) considered fully nonparametric sample selection models.

Quantile regression has progressively emerged as the method of choice to analyze the effects of covariates on the distribution of the outcome. In the absence of selection, Koenker and Bassett (1978) proposed a parametric (linear) estimator for conditional quantile models. Owing to its ability to capture heterogeneous effects, its theoretical properties have been studied extensively and it has been used in many empirical studies. Guntenbrunner and Jurečková (1992) derived the uniform asymptotic representation for the quantile regression process. Koenker and Xiao (2002) suggested tests based on this process. Angrist *et al.* (2006) analyzed the misspecified quantile regression process. Chernozhukov *et al.* (2013) suggested tests based on the counterfactual unconditional quantile processes. Buchinsky (1994) is an early and influential application of quantile regression to the estimation of wage regression. Buchinsky (1998b), Koenker and Hallock (2000) and Koenker (2005) provide a comprehensive discussion of quantile regression models and recent developments.

The remainder of this paper is organized as follows. In Section 2 we describe the sample selection model of Buchinsky (1998a) and discuss the role of the independence assumption in sample selection models. Section 3 outlines the test procedure and proves its asymptotic validity. In Section 4 Monte Carlo simulations document the power and size properties of the proposed test. Section 5 offers an empirical application showing the relevance of the problem at the heart of this paper. Section 6 concludes.

2. THE CONDITIONAL INDEPENDENCE ASSUMPTION IN SAMPLE SELECTION MODELS

2.1. The Model

We consider the quantile sample selection model of Buchinsky (1998a), which is similar to the semiparametric models for the mean in Cosslett (1991), Powell (1987) and Newey (2009). The outcome equation and the latent selection function are assumed to be linear in the covariates as in the seminal work of Heckman (1974, 1976, 1979). 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 unknown. While we derive formal results only for this specific model, we would like to emphasize that the importance of the independence assumption, its implication for quantile regression and its testability are all preserved in more general models. This includes models with a nonparametric selection equation as in Ahn and Powell (1993) or a nonparametric outcome equation as in Das

---

[1] 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. Interested researchers can therefore easily verify whether conditional independence holds in their applications.

et al. (2003).[2] Bearing this in mind, we maintain the following assumption, which is equation (2) in Buchinsky (1998a).

Assumption 1: potential outcome. The potential outcome of interest, Y^*, is determined by

$$Y^* = X_2'\beta_0 + U$$

where X_2 denotes a vector of regressors, β_0 is the vector of slope coefficients and U is an error term.

X_2 does not include a constant term, so that β_0 only comprises the slope coefficients and the intercept is incorporated in the error term. We use this simplified notation because the intercept is not relevant for testing conditional independence. Note that we do not observe the latent variable Y^* but only Y, which is defined in Assumption 2.

Assumption 2: observation rule and selection equation.

$$Y = Y^* \text{ if } D = 1 \text{ and is not observed otherwise}$$

$$D = 1\left(X_1'\gamma_0 + \varepsilon \geq 0\right)$$

where $1(\cdot)$ is the indicator function, X_1 is a strict superset of X_2 and ε is a second error term.

The indicator function for being selected, D, is determined by a latent single-index crossing model, which corresponds to equation (5) in Buchinsky (1998a). We do not impose any parametric restriction on the distribution of (U, ε) but we require X_1 to include at least one continuous variable which is not in X_2 and has a non-zero coefficient in the selection equation.

Assumption 3: conditional independence. The joint density of (U, ε) is absolutely continuous with respect to Lebesgue measure. It is denoted by $f_{U,\varepsilon}$ and is independent of X_1 conditional on the latent selection index:

$$f_{U,\varepsilon}(\cdot|X_1) = f_{U,\varepsilon}\left(\cdot|X_1'\gamma_0\right)$$

Assumption 3 is the conditional independence that, combined with the additivity of the outcome in Assumption 1, is at the center of this paper. It corresponds exactly to Assumptions C and E in Buchinsky (1998a).[3] This assumption is naturally implied by the unconditional independence between (U, ε) and X_1 but it is less restrictive because it allows X_1 to affect the error terms through the index governing the selection probability. For instance, heteroscedasticity with respect to the selection probability is allowed. We make and test only this conditional independence assumption because it is

---

[2] In this paper we assume that the selection equation is semiparametric, which allows estimating the index governing selection at the $\sqrt{n}$ rate. A test similar to the one developed in this paper could be built on the quantile version of Ahn and Powell (1993). Our asymptotic results likely extend to this model by applying the general uniform results in Escanciano and Zhu (2013). The extension to a nonparametric outcome equation could be based on the quantile version of Das *et al.* (2003). However, deriving the limiting distribution for this estimator appears very challenging. Mammen *et al.* (2012) provide results for this type of model but only results for conditional means; we are not aware of similar results for the conditional quantiles. These two possible extensions are left for future research.

[3] In some sample selection models concerned with identifying the mean function, only a conditional moment restriction (such as $E[U|X_1, D = 1] = E[U|X_1'\gamma_0, D = 1]$) is imposed without assuming full independence. However, the latter often serves as a justification for the moment restriction. Departures from full independence that still satisfy the moment condition are not substantial. For example, 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).

enough to guarantee the consistency of the Buchinsky (1998a) estimator. The increase in generality should nonetheless not be overemphasized. Because of Assumption 2 we could equivalently write Assumption 3 as

$$f_{U,\varepsilon}(\cdot|X_1) = f_{U,\varepsilon}(\cdot|\Pr(D = 1|X_1))$$

which implies that X_1 would be unconditionally independent of (U, ε) in the absence of sample selection (i.e. if the entire population was observed with probability one).

2.2. Identification and Testability

Let $Q_V(\tau|W)$ denote the τth conditional quantile of V given W, for any random variables V and W. Under Assumptions 1–3 for any $0 < \tau < 1$:

$$\begin{aligned}
Q_Y(\tau|X_1, D = 1) &= X_2'\beta_0 + Q_U(\tau|X_1, D = 1) \\
&= X_2'\beta_0 + Q_U(\tau|X_1, X_1'\gamma_0 \geq -\varepsilon) \\
&= X_2'\beta_0 + Q_U(\tau|X_1'\gamma_0, X_1'\gamma_0 \geq -\varepsilon) \\
&= X_2'\beta_0 + Q_U(\tau|X_1'\gamma_0, D = 1) \\
&\equiv X_2'\beta_0 + h_\tau(X_1'\gamma_0)
\end{aligned}$$

The first line holds by Assumption 1, the second line by Assumption 2, the third line by Assumption 3 and the fourth line again by Assumption 2. In the last line we define $h_\tau(X_1'\gamma_0) \equiv Q_U(\tau|X_1'\gamma_0, D = 1)$.

This result has three important consequences. First, Assumptions 1–3 point identify β_0 if the regressors X_2 are not linearly dependent after conditioning on $X_1'\gamma_0$ and $D = 1$.[4] This implies that β_0 can be estimated by the τth quantile regression of Y on X_2 and a nonparametric function of $X_1'\hat{\gamma}$ in the selected sample ($\hat{\gamma}$ being a consistent estimate of γ_0). Second, since this is true for *any* quantile τ, Assumptions 1–3 imply that the slopes of all quantile regressions are identical. In addition, these slopes are also equal to the slopes of the corresponding mean regression. 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, the identifying assumptions exclude such a possibility.

Third, this result also implies that the model defined by Assumptions 1–3 can be tested by comparing the slope coefficients obtained at different quantiles. Heterogeneity of the slope coefficients across several quantiles points to the violation of (at least) one identifying assumption. 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 conditional independence assumption is a necessary condition for the consistency of the estimators suggested in Heckman (1979) for parametric sample selection models, but also in more general models as considered in Cosslett (1991), Gallant and Nychka (1987), Powell (1987), Das *et al.* (2003) and Newey (2009) that allow relaxing Assumptions 1 and/or 2. 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 to suggest a formal test.

Our test is specifically designed to detect deviations from the conditional independence assumption (Assumption 3) while maintaining the other assumptions. The next subsection gives examples for which our test is consistent for deviations from this assumption. Of course, it may also detect some

---

[4] This last assumption is formally stated below as Assumption 5. Its main substantial restriction (X_1 must be a strict superset of X_2) is already imposed by Assumption 2.

violations of the maintained assumptions such as the validity of the exclusion restriction or the correct parametric specifications but it is not designed to be very powerful in these cases, if at all. For this reason, we call our procedure a test of the conditional independence assumption. In any case, a rejection of the null hypothesis indicates that the model is misspecified and the estimates may be misleading.

2.3. Two Examples

Two examples will convey the intuition about the inconsistency of traditional estimators when independence is violated and about the testability of this assumption. These two examples correspond to the data-generating processes (DGPs) used in Section 4 to assess the finite-sample properties of our tests. We first consider a simple multiplicative heteroscedastic model:

$$Y^* = \alpha_0 + X_2'\beta_0 + \left(1 + X_2'\delta\right) \cdot V \tag{1}$$

where α_0 is the constant and V is a mean zero error term that is independent of X_1. In the notation of Assumption 1, $U = \alpha_0 + \left(1 + X_2'\delta\right) \cdot V$ and the conditional independence assumption is satisfied if and only if $\delta = 0$. It follows that

$$E[Y|X_1, D = 1] = \alpha_0 + X_2'\beta_0 + \left(1 + X_2'\delta\right) \cdot E\left[V|X_1'\gamma_0, D = 1\right]$$

We see that controlling for the selection probability (as done by sample selection mean estimators) is no longer sufficient to identify β_0. If we fix $X_1'\gamma_0$ to a specific value, the slope estimates in the selected sample will converge to

$$\beta_0 + \delta \cdot E\left[V|X_1'\gamma_0, D = 1\right]$$

There is no bias only if (i) $\delta = 0$ (i.e. Assumptions 1–3 are satisfied) or if (ii) $E\left[V|X_1'\gamma_0, D = 1\right] = 0$ (i.e. selection is random).

The same issue also arises for the quantile regression estimator of Buchinsky (1998a):

$$Q_Y(\tau|X_1, D = 1) = \alpha_0 + X_2'\beta_0 + \left(1 + X_2'\delta\right) \cdot Q_v\left(\tau|X_1'\gamma_0, D = 1\right)$$

On the bright side, this demonstrates the possibility of testing the validity of our model because the slope of the τth quantile regression in the selected population after fixing $X_1'\gamma_0$ is

$$\beta_0 + \delta \cdot Q_v\left(\tau|X_1'\gamma_0, D = 1\right)$$

Thus the slope is constant as a function of τ if and only if $\delta = 0$, i.e. if conditional independence holds. In other words, our test is asymptotically capable of detecting any deviation from independence when the true model is given by equation (1).

Figure 1 illustrates this example with 500 simulated realizations of (X_2, Y^*) from model (1) with a scalar X_2 and $\beta_0 = 1$. X_1 has been chosen such that the conditional selection probability is 0.5 for each observation. Selection is positive, which can be seen from the fact that we observe more realizations (symbolized by boxes around the crosses) above than below the median. The error terms are independent in the left panel and heteroscedastic in the right panel ($\delta = 0$ vs. $\delta = 0.5$).[5] In the first case, selection induces a shift in the conditional median of Y, but this bias is constant across observations due to the constant participation probability, resulting in a correct slope. In the heteroscedastic

---

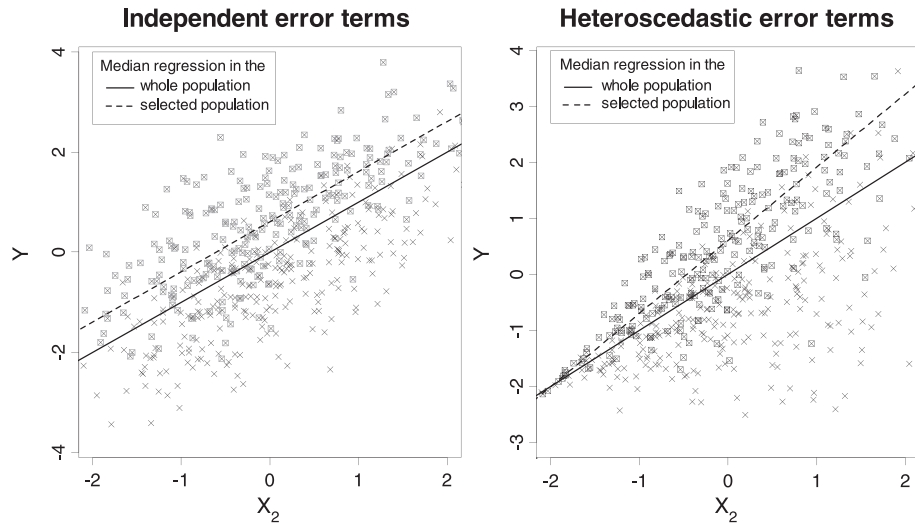[5] More details concerning the DGP can be found in Section 4.1.

Figure 1. Sample selection under independence and heteroscedasticity. Random samples of 500 observations. Only the observations surrounded by a box are observed. All observations have X_1 set such that $\Pr(D = 1|X_1) = 0.5$

case, the bias is increasing as a function of X_2 because the variance of the errors is increasing with X_2, resulting in an excessive slope. In general, the sign of the bias corresponds to the product of the sign of δ and the direction of the selection. Figure 1 makes also clear that the independence assumption can be tested by comparing the slopes of several quantile regressions.

After the location scale shift model, we consider a more general linear quantile regression model:

$$Y^* = \alpha(V) + X_2'\beta(V), V|X_1 \sim U(0,1)$$

where $\tau \mapsto \alpha(\tau) + x_2'\beta(\tau)$ is strictly increasing in τ for each value of x_2 in support of X_2. Using the notation of Assumption 1, we can define $U = \alpha(V) + X_2'(\beta(V) - \beta_0)$ and the conditional independence assumption is satisfied if and only if $\beta(\tau) = \beta_0$ for all $\tau \in (0,1)$.

In this model

$$Q_Y(\tau|X_1, D = 1) = \alpha\left(Q_V\left(\tau|X_1'\gamma_0, D = 1\right)\right) + X_2\beta\left(Q_V\left(\tau|X_1'\gamma_0, D = 1\right)\right) \qquad (2)$$

Thus, if we fix $X_1'\gamma_0$, the slope of the quantile function is given by $\beta\left(Q_V\left(\tau|X_1'\gamma_0, D = 1\right)\right)$. There is no bias only if (i) $\beta(\tau) = \beta(0.5)$ for all τ (i.e. Assumptions 1–3 are satisfied) or if (ii) $Q_V\left(\tau|X_1'\gamma, D = 1\right) = \tau$ (i.e. selection is random at the quantile of interest).

We suggest testing the model defined by Assumptions 1–3 by testing whether $\beta\left(Q_V\left(\tau|X_1'\gamma_0, D = 1\right)\right)$ is constant across all quantiles τ. Our test asymptotically detects any violation of independence under a minor additional assumption: if there exist at least two quantiles $\tilde{\theta}$ and $\bar{\theta}$ with $\beta\left(\tilde{\theta}\right) \neq \beta\left(\bar{\theta}\right)$ and $\Pr\left(D = 1|V = \tilde{\theta}\right) > 0$ and $\Pr\left(D = 1|V = \bar{\theta}\right) > 0$, then the slopes of the $\tilde{\tau} = \Pr\left(V \leq \tilde{\theta}|D = 1\right)$ and $\bar{\tau} = \Pr\left(V \leq \bar{\theta}|D = 1\right)$ conditional quantile regressions in the selected sample will differ. In other words, when the true model is a linear quantile regression model then our test has power against any deviation from independence as long as the heterogeneous quantiles are not completely unobserved.

Figure 2 shows the biases that can arise in such a DGP when the parameters are calibrated to empirical data as discussed in Section 4.2. The true quantile function is the male wage function in the 2011 CPS, the selection equation is the female selection equation during that year, and the copula between V and ε is Gaussian. We plot the values to which the estimator of Buchinsky (1998a) converges when the correlation between V and ε is set to -0.99, -0.5, 0, 0.5 and 0.99. The true values correspond to the values for zero correlation. The quantile regression coefficients—in particular, the way they change with the quantile of interest—are quite different for different correlations. In an application, the correlation between V and ε is, however, not identified without distributional assumptions, possibly leading to an incorrect analysis.

Sample selection procedures are often applied to compare outcomes (e.g. the gender wage gap) between different time periods. Our results imply not only that the estimates can be biased in any period but also that the bias will change over time if the correlation between V and ε changes. This seems to be particularly an issue for the results in Mulligan and Rubinstein (2008) because one of their main findings is precisely that selection into employment changed from negative to positive over time.
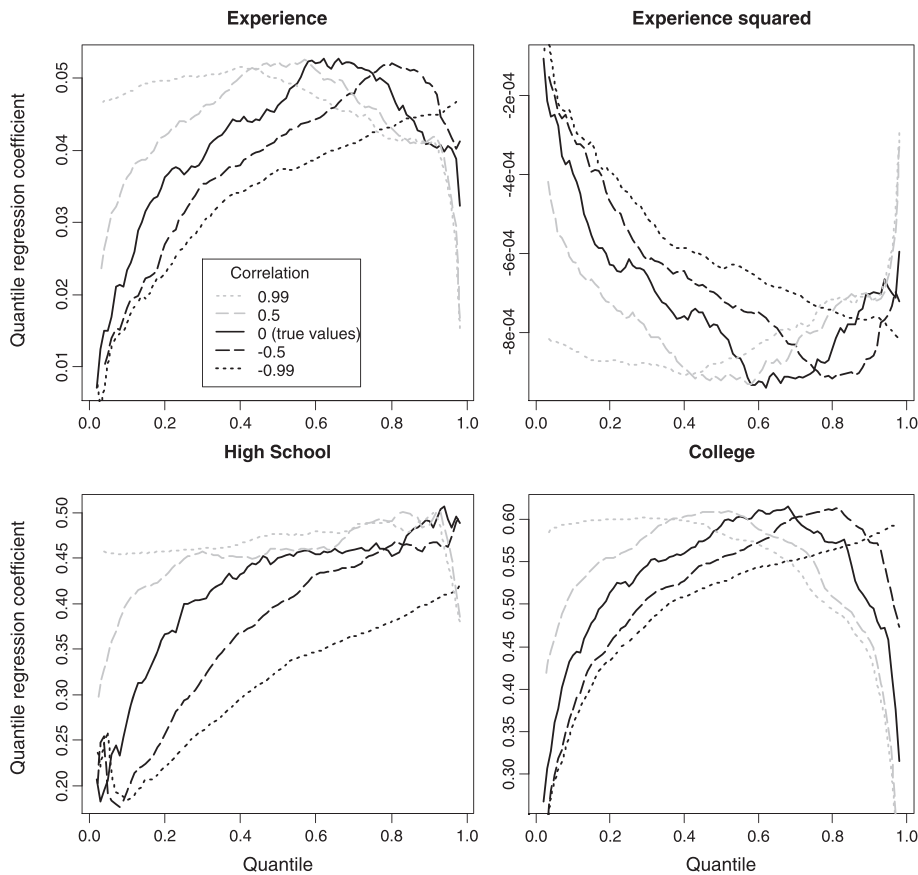

Figure 2. Asymptotic estimates in realistic DGPs. Details of the DGP can be found in Section 4.2. The true coefficients correspond to the male coefficients in log wage regressions in CPS 2011. Plotted lines give the values to which the estimates of Buchinsky (1998a) converge for five different correlations between the error terms in the selection and outcome equation, ε and U

3. TEST PROCEDURE

The discussion in Section 2 shows that we can test the validity of the independence assumption by testing whether the coefficients on X_2 in the quantile regression of Y on X_2 and $h_\tau(X_1'\gamma_0)$ are constant as a function of the quantile. Our testing approach is based on the Kolmogorov–Smirnov and Cramér–von Mises statistics, which are defined on the empirical quantile regression process. We follow Buchinsky (1998a) and approximate the unknown function $h_\tau(X_1'\gamma_0)$ by a power series in $X_1'\gamma_0$. We estimate the unknown parameter γ_0 using the Klein and Spady (1993) estimator for binary choice models. In Sections 3.1 and 3.2 we describe step by step the estimation and testing algorithms. In Section 3.3 we give the formal inference results.

3.1. Estimation

The first step consists of estimating the selection equation, i.e. the parameter γ_0. The formal inference results do not depend on the specific estimator of γ_0 but only require $\sqrt{n}$ consistency and asymptotic linearity of the influence function. In our simulations and application we estimate the selection equation by the semiparametric binary choice estimator suggested in Klein and Spady (1993):

$$\hat{\gamma} \equiv \arg\max_\gamma \sum_{i=1}^n \left\{ D_i \cdot \log\left[\hat{E}\left(D|x_{1i}'\gamma\right)\right] + (1 - D_i) \cdot \log\left[1 - \hat{E}\left(D|x_{1i}'\gamma\right)\right] \right\} \tag{3}$$

where

$$\hat{E}\left(D|X_{1i}'\gamma\right) = \frac{\sum_{j=1}^n D_j \cdot \kappa_{KS}\left(\left(x_{1i}'\gamma - x_{1j}'\gamma\right)/w_{KS}\right)}{\sum_{j=1}^n \kappa_{KS}\left(\left(x_{1i}'\gamma - x_{1j}'\gamma\right)/w_{KS}\right)} \tag{4}$$

in which w_{KS} is a bandwidth that depends on the sample size and $\kappa_{KS}(\cdot)$ denotes a kernel function. This estimator satisfies our regularity conditions under the assumptions detailed in Klein and Spady (1993) and it attains the semiparametric efficiency bound for this model. Heteroscedasticity is allowed to depend on the regressors only through the linear index.

In the second step, the function $h_\tau(X_1'\gamma_0)$ is approximated by a power series expansion. As suggested by Buchinsky (1998a), we use a power series expansion of the inverse Mill's ratio of the normalized estimated index but other bases of power functions also satisfy our assumptions. The order of the approximation, J, must increase with the sample size to ensure the consistency of the estimator. $\hat{\beta}(\tau)$ solve the following optimization problem:

$$\left(\hat{\beta}(\tau), \hat{\pi}(\tau)\right) = \arg\min_{(\beta,\pi)} \sum_{i:d_i=1} \hat{t}_i \cdot \rho_\tau\left(y_i - x_{2i}'\beta - \Pi_J\left(x_{1i}'\hat{\gamma}\right)\pi\right) \tag{5}$$

where $\rho_\tau(A) = A \cdot (\tau - 1(A \le 0))$ is the check function suggested by Koenker and Bassett (1978). Furthermore, $\hat{t}_i$ is a trimming function:

$$\hat{t}_i = 1\left(x_{1i}'\hat{\gamma} \in \mathcal{X}\right)$$

where $\mathcal{X}$ is a one-dimensional compact set. Trimming is convenient for proving several asymptotic results (some of which we borrow from Lee, 2007) such as the boundedness of the influence function. We do not expect it to be a critical tuning parameter in most applications. $\Pi_J\left(x_{1i}'\hat{\gamma}\right)$ is a polynomial row vector in the inverse Mill's ratio, $\Pi_J\left(x_{1i}'\hat{\gamma}\right) = \left(1, \lambda\left(x_{1i}'\hat{\gamma}\right), \lambda\left(x_{1i}'\hat{\gamma}\right)^2, \ldots, \lambda\left(x_{1i}'\hat{\gamma}\right)^J\right)$, where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the ratio of the normal probability density function to the normal cumulative distribution function.

3.2. Tests

Our testing problem is given by

$$H_0: \beta(\tau) = \beta(0.5), \forall \tau \epsilon \mathcal{T} \tag{6}$$

$$\text{versus } H_1: \beta(\tau) \neq \beta(0.5), \quad \text{for some } \tau \epsilon \mathcal{T} \tag{7}$$

where $\mathcal{T}$ is a closed subset of $[e, 1-e], 0 < e < 1$, and $\beta(\tau)$ denotes the true τ quantile regression coefficient defined as

$$\beta(\tau) = \arg\min_{\beta} E \left[\rho_\tau \left(Y_i - X'_{2i}\beta - h_\tau \left(X'_{1i} \gamma_0 \right) \right) \right] \tag{8}$$

Buchinsky (1998a) gives the joint asymptotic distribution of $\hat{\beta}(\tau)$ for a finite number of quantiles. Based on his results, we can use a finite number of quantile regressions and apply a Wald test as proposed by Koenker and Bassett (1982) in the absence of sample selection. Even asymptotically, this test does not have power against all deviations from the null because the number of quantiles considered is fixed. In addition, it is tempting to choose the quantiles used by the test after having seen the results, which may distort the size of the test. Our tests use the whole quantile process and are therefore consistent against all global departures from H_0.

We measure the deviations from the null hypothesis by the Kolmogorov–Smirnov (KS) and Cramér–von Mises (CM) statistics applied to the empirical process $\hat{\beta}(\tau) - \hat{\beta}(0.5)$:

$$T_n^{\text{KS}} = \sup_{\tau \epsilon \mathcal{T}} \sqrt{n} \left|\left| \hat{\beta}(\tau) - \hat{\beta}(0.5) \right|\right|_{\hat{\Lambda}_\tau} \quad \text{and} \quad T_n^{\text{CM}} = n \int_{\mathcal{T}} \left|\left| \hat{\beta}(\tau) - \hat{\beta}(0.5) \right|\right|^2_{\hat{\Lambda}_\tau} \mathrm{d}\tau \tag{9}$$

where $||a||_{\hat{\Lambda}_\tau}$ denotes $\sqrt{a' \hat{\Lambda}_\tau a}$ 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.

Inference requires a knowledge of the asymptotic distributions of $T_n^{\text{KS}}, T_n^{\text{CM}}$. In Section 3.3 we derive the asymptotic distribution of the quantile regression process $\sqrt{n} \left(\hat{\beta}(\tau) - \beta(\tau) \right)$, which would allow us to derive the distribution of the test statistic if the null hypothesis was not estimated. The presence of a nuisance parameter (the median coefficients vector) jeopardizes this possibility.

As suggested by Chernozhukov and Fernández-Val (2005), we could calculate the critical values by resampling the centered inference process $\hat{\beta}(\tau) - \hat{\beta}(0.5)$. But the repeated computation of the coefficients (in our case especially of $\hat{\gamma}$) for each bootstrap sample can be quite costly, in particular when the sample size is large. For this reason, we follow Chernozhukov and Hansen (2006) and resample the score functions instead, which are linear approximations of the empirical processes. We show in Appendix A that

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

A precise definition of the score function $s_i(\tau)$, as well as its estimator $\hat{s}_i(\tau)$, is given in the next subsection. We draw B samples of size b and denote such samples as $\Upsilon_j, 1 \leq j \leq B$. The KS and CM statistics for the jth iteration are

$$T_{n,b,j}^{\text{KS}} \equiv \sup_{\tau \epsilon \mathcal{T}} \sqrt{b} \left| \left| 1/b \sum_{i \epsilon \Upsilon_j} \hat{s}_i(\tau) \right| \right|_{\hat{\Lambda}_\tau} \quad \text{and} \quad T_{n,b,j}^{\text{CM}} \equiv b \int_{\mathcal{T}} \left| \left| 1/b \sum_{i \epsilon \Upsilon_j} \hat{s}_i(\tau) \right| \right|^2_{\hat{\Lambda}_\tau} \mathrm{d}\tau$$

The critical values of a size α test are the $1 - \alpha$th quantiles of $T_{n,b,j}^{\mathrm{KS}}$ and $T_{n,b,j}^{\mathrm{CM}}$, i.e. $c_{n,b}^{\mathrm{KS}}(\alpha) = \inf\left\{c : B^{-1} \sum_{j=1}^{B} 1\left(T_{n,b,j}^{\mathrm{KS}} \leq c\right) \geq 1 - \alpha\right\}$ and similarly for the CM statistic. The level α tests reject H_0 if $T_n^{\mathrm{KS}} > c_{n,b}^{\mathrm{KS}}(\alpha)$ and $T_n^{\mathrm{CV}} > c_{n,b}^{\mathrm{CM}}(\alpha)$.

3.3. Formal Inference Results

We prove that our tests have asymptotically the correct size and are consistent against all global alternative hypotheses by verifying the conditions of Theorem 5 in Chernozhukov and Hansen (2006). To do so, we borrow heavily from the results in Lee (2007), who considers an instrumental variable estimator that is almost identical to the Buchinsky sample selection estimator from a computational point of view. To establish these results, we make the following additional assumptions.

Assumption 4: sampling. The data $\{(y_i, x_{1i}, d_i) : i = 1, \ldots, n\}$ are i.i.d. realizations of (Y, X_1, D) and take values in a compact set.

Assumption 5: non-singularity.
(a) Define

$$\Delta_{b,\tau} = E\left[f_U\left(Q_U\left(\tau | x'_{1i} \gamma_0, d_i = 1\right) | x'_{1i} \gamma_0, d_i = 1\right) \cdot m_i m'_i\right],$$
$$m_i = t_i \cdot d_i \cdot \left(x_{2i} - E\left[X_2 | x'_{1i} \gamma_0, d_i = 1\right]\right)$$

$\Delta_{b,\tau}$ is non-singular uniformly in $\tau \in \mathcal{T}$, i.e.

$$\sup_{\tau \in \mathcal{T}} \det(\Delta_{b,\tau}) > c$$

for some constant $c > 0$.
(b) Define

$$\Phi_J(\tau) = E\left[f_U\left(Q_U\left(\tau | x'_{1i} \gamma_0, d_i = 1\right) | x'_{1i} \gamma_0, d_i = 1\right) \cdot P_{Ji} P'_{Ji}\right]$$
$$P_{Ji} = t_i \cdot d_i \cdot \left(x'_{2i}, \Pi_J\left(x'_{1i} \gamma_0\right)\right)'$$

Uniformly in $\tau \in \mathcal{T}$ the smallest eigenvalue of $\Phi_J(\tau)$ is bounded away from zero for all J, and the largest eigenvalue of $\Phi_J(\tau)$ is bounded for all J, i.e.

$$\inf_{\tau \in \mathcal{T}} \lambda_{\min}(\Phi_J(\tau)) > \underline{c} \quad \text{and} \quad \sup_{\tau \in \mathcal{T}} \lambda_{\max}(\Phi_J(\tau)) < \overline{c}$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and the largest eigenvalues of a symmetric matrix A and $\underline{c}$ and $\overline{c}$ are two strictly positive constants.

Assumption 6: continuity.
(a) The density of U in the selected sample, $f_U(u | X_1, D = 1)$, is bounded above and away from zero and is twice continuously differentiable with respect to u uniformly in u and X_1.
(b) $h_\tau(X'_1 \gamma_0)$ is r-times continuously differentiable with respect to $X'_1 \gamma_0$ uniformly in $\tau \in \mathcal{T}$.
(c) As a function of $X'_1 \gamma_0$, $E\left[t_i \cdot f_U\left(Q_U\left(\tau | X'_1 \gamma_0, d_i = 1\right) | X'_1 \gamma_0, d_i = 1\right) \cdot x_{2i} | X'_1 \gamma_0, d_i = 1\right]$ is continuously differentiable uniformly in $\tau \in \mathcal{T}$.

Assumption 7: number of approximating functions. $J = Cn^{\upsilon}$ for some constant C satisfying $0 < C < \infty$ and some υ satisfying $\frac{1}{2r} < \upsilon < \frac{1}{9}$.

Assumption 8: first stage. There exists an invertible matrix Δ_P and a bounded $\dim(X_1) \times 1$ column vector $k_i \equiv K(x_{1i}, d_i, \gamma_0)$ such that $E[k_i] = 0$, $E\left[k_i k_i'\right]$ exists and is non-singular and

$$\sqrt{n}\left(\hat{\gamma} - \gamma_0\right) = \Delta_P^{-1} \frac{\sum_{i=1}^{n} k_i}{\sqrt{n}} + o_p(1)$$

There is an estimator $\hat{\Delta}_P$ of Δ_P such that $\hat{\Delta}_P \to \Delta_P$.

These assumptions are the sample selection equivalent of the assumptions in Lee (2007). Assumption 5(a) is satisfied if X_1 includes at least one continuous variable which is not in X_2 and has a non-zero coefficient in the selection equation (and the components of X_2 are not linearly dependent). Assumption 5(b) ensures the non-singularity of the second moment matrix of the estimator. Assumption 6(a) is standard in quantile models because quantiles are well defined and have a standard asymptotic distribution only for continuous random variables. The restrictions imposed on $f_U(u|X_1, D = 1)$ are implied by the same restrictions on the density of U in the whole population together with $\Pr(D = 1|X_1, U = u) > 0$. In other words, they are implied by the standard regularity conditions on U when the probability to observe the outcome at the quantile of interest is strictly positive. Assumptions 6(b) and 6(c) require that the approximated functions are sufficiently smooth. Assumption 7 restricts the growth rate of J. Together with Assumption 6(b), it requires that $h_\tau(X_1'\gamma_0)$ is at least five times differentiable ($r \geq 5$). Assumption 8 imposes regularity conditions on the estimation of γ_0. The estimator suggested by Klein and Spady (1993) satisfies this condition with

$$\Delta_P = E\left[\frac{\partial P_i(\gamma)}{\partial \gamma}\bigg|_{\gamma=\gamma_0} \frac{\partial P_i(\gamma)'}{\partial \gamma}\bigg|_{\gamma=\gamma_0} \frac{1}{P_i(\gamma_0) \cdot (1 - P_i(\gamma_0))}\right]$$

$$\text{and } k_i = \frac{\partial P_i(\gamma)}{\partial \gamma}\bigg|_{\gamma=\gamma_0} \frac{D_i - P_i(\gamma_0)}{P_i(\gamma_0) \cdot (1 - P_i(\gamma_0))}$$

where $P_i(\gamma) = \Pr\left(D = 1|X_1'\gamma = x_{1i}'\gamma_0\right)$.

Under Assumptions 1–8 the influence function of $\hat{\beta}(\tau)$ is given by (see Appendix A for details)

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) = \Delta_{b,\tau}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\ell_i(\tau) + \Delta_\gamma \Delta_P^{-1} k_i\right) + o_p(1) \tag{11}$$

where

$$\ell_i(\tau) = \left(\tau - 1\left(y_i < x_{2i}'\beta_0 + h_\tau(x_{1i}'\gamma_0)\right)\right) \cdot m_i,$$
$$\Delta_\gamma = E\left[t_i \cdot f_\varepsilon\left(Q_\varepsilon(\tau|x_{1i}'\gamma_0) \cdot \frac{dh_\tau(x_{1i}'\gamma_0)}{d\gamma_0} m_i\right]\right.$$

$s_i(\tau)$ is simply the difference between the influence function of $\hat{\beta}(\tau)$ and the influence function of $\hat{\beta}(0.5)$. We estimate the influence functions by their sample analogs. In particular, we estimate the densities using the kernel method as in Powell (1986), Buchinsky (1998a) and Lee (2007):

$$\hat{\ell}_i(\tau) = \left(\tau - 1\left(y_i < x'_{2i}\hat{\beta}(\tau) + \Pi_J\left(x'_{1i}\hat{\gamma}\right)\hat{\pi}(\tau)\right)\right) \cdot \hat{m}_i,$$

$$\hat{\Delta}_{b,\tau} = (nw_n)^{-1} \sum_{i=1}^{n} \hat{t}_i \cdot \kappa\left(\frac{\hat{\varepsilon}_i(\tau)}{w_n}\right) \cdot \hat{m}_i\hat{m}'_i,$$

$$\hat{\Delta}_{\gamma} = (nw_n)^{-1} \sum_{i=1}^{n} \hat{t}_i \cdot \kappa\left(\frac{\hat{\varepsilon}_i(\tau)}{w_n}\right) \cdot \frac{d\Pi_J\left(x'_{1i}\hat{\gamma}\right)}{d\gamma}\hat{m}_i,$$

$$\hat{\varepsilon}(\tau) = y_i - x'_{2i}\hat{\beta}(\tau) - \Pi_J\left(x'_{1i}\hat{\gamma}\right)\hat{\pi}(\tau)$$

where $\hat{m}_i$ is the vector of residuals from the regression of X_2 on $\Pi_J\left(X'_1\hat{\gamma}\right)$ evaluated at x_{1i} and multiplied by $\hat{t}_i d_i$. We impose the following restrictions on the kernel function $\kappa(\cdot)$ and the bandwidth w_n.

Assumption 9: kernel and bandwidth. The kernel function κ has support $[-1, 1]$, is bounded and symmetrical about 0 and satisfies $\int_{-1}^{1} \kappa(u)\, \mathrm{d}u = 1$, $\int_{-1}^{1} u\kappa(u)\mathrm{d}u = 0$, and $\int_{-1}^{1} u^2\kappa(u)\mathrm{d}u < \infty$. $w_n = c_h n^{-\mu}$ for some positive finite c_h and some μ satisfying $\upsilon/2 < \mu < (1 - 4\upsilon)/4$.

Under these conditions our tests are asymptotically correctly sized and consistent against any global alternative hypothesis.

Proposition 1. Suppose Assumptions 1–9 hold. Then, as $B \to \infty, b \to \infty, n \to \infty$: (i) under the null hypothesis $P\left(T_n^{\mathrm{KS}} > c_{n,b}^{\mathrm{KS}}(\alpha)\right) \to \alpha$ and $P\left(T_n^{\mathrm{CM}} > c_{n,b}^{\mathrm{CM}}(\alpha)\right) \to \alpha$; (ii) under the alternative hypothesis $P\left(T_n^{\mathrm{KS}} > c_{n,b}^{\mathrm{KS}}(\alpha)\right) \to 1$ and $P\left(T_n^{\mathrm{CM}} > c_{n,b}^{\mathrm{CM}}(\alpha)\right) \to 1$.

Chernozhukov and Hansen (2006) develop new inference tools in the context of their instrumental variable quantile regression estimator but they state the conditions in a form that is general enough to apply to other estimators. We prove Proposition 1 by verifying the conditions for inference of their Theorem 5. Similar tests based on resampling the estimates (instead of the scores) are also valid by Theorem 1 in Chernozhukov and Fernández-Val (2005) because their conditions are weaker than the conditions that we check in Appendix A. In addition to the statement of our Proposition 1, Chernozhukov and Fernández-Val (2005) also show that this test has non-trivial power against local alternatives.

3.4. Practical Details

3.4.1. Smoothing Parameters
We use the Gaussian density for all kernel functions. We follow a pragmatic approach and use the generalized cross-validation criterion (GCV) proposed in Craven and Wahba (1979) to select the smoothing parameters w_{KS} and J. The same method was used in Buchinsky (1998a) and Newey *et al.* (1990), among others. We do not claim any optimality of this method for our estimation and inference

problem but it works well in our simulations.[6] Finally, we use the Hall and Sheather (1988) rule for w_n, which is recommended by Koenker (2005).

3.4.2. Discretization

The theoretical results still hold if, instead of $\mathcal{T}$, we use a grid $\mathcal{T}_n$ with the largest cell size converging to 0 as $n \rightarrow \infty$. In the simulations and application, we use for simplicity a grid with mesh size 0.01 but the results are not sensitive to the choice of mesh size as long as it is small enough. The differences between the results with a mesh of size 0.005, 0.01 and 0.02 are completely negligible.

3.4.3. Choice of Λ_τ

In order to improve the power of the test we set $\Lambda_\tau = \text{var}[s_i(\tau)]^{-1}$. Note that by construction $\text{var}[s_i(0.5)] = 0$, which forces us to cut out the part of $\mathcal{T}_n$ close to the median. In our simulations and application we disregard $[0.46, 0.54]$. We estimate $\text{var}[s_i(\tau)]$ by $\text{var}[\hat{s}_i(\tau)]$ (with $\text{var}[\hat{s}_i(\tau)]$ denoting the sample variance of $\hat{s}_i(\tau)$).

3.4.4. Choice of b and $\mathcal{T}$

We experiment in our simulations and application with several rules for the choice of block size, b. Although our empirical results are not sensitive to the choice of block size, the bootstrap has better size and power properties in our simulations, while the tests are conservative when the subsample size is smaller. Concerning $\mathcal{T}$, in principle we would like to use a large range of quantiles in order to increase the asymptotic power of the tests. However, the quantile regression estimator performs poorly in the tails of the distribution if we do not make strong assumptions about the density of Y. In the simulations we experiment with several ranges ($[0.05, 0.95]$, $[0.1, 0.9]$ and $[0.2, 0.8]$).

4. SIMULATIONS: POWER AND SIZE PROPERTIES OF THE TESTS

In this section, we present Monte Carlo evidence about the size and power properties of the tests that we have proposed in Section 3. We first consider a simple DGP that was used by Koenker and Xiao (2002) and Chernozhukov and Fernández-Val (2005), with the difference that we observe only a non-random sample from the population. Second, we will use a more realistic multivariate example calibrated to empirical data used in Section 5.

4.1. A Simple Heteroscedastic DGP

$$
\begin{aligned}
D &= 1(X_2 + Z + \varepsilon > 0), \\
Y^* &= X_2 + (1 + X_2\delta) \cdot V, \\
X_2 &\sim N(0,1), \quad Z \sim N(0,1)
\end{aligned}
\tag{12}
$$

where

$$
(V, \varepsilon) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right)
$$

---

[6] The optimal choice of smoothing parameters in semiparametric models remains an important topic for future research. Although they are probably not optimal, our choice of w_n is in the range of values allowed by Delecroix et al. (2006) for $\sqrt{n}$ consistency and, as noted by Newey (2009) for a similar model, our choice of J should satisfy the rate condition of Assumption 7. Alternatively, Härdle *et al.* (1993) and Escanciano and Zhu (2014) suggested estimating simultaneously the bandwidth and the parameters for the binary choice and the sample selection model, respectively.

Table I. Empirical rejection frequencies for 5% bootstrap tests: normal distribution, 1000 replications, 250 bootstrap draws

$\tau \epsilon$	$\delta = 0$			$\delta = 0.2$			$\delta = 0.5$		
	[0.05, 0.95]	[0.10, 0.90]	[0.20, 0.80]	[0.05, 0.95]	[0.10, 0.90]	[0.20, 0.80]	[0.05,0.95]	[0.10,0.90]	[0.20,0.80]
Kolmogorov–Smirnov test									
$n = 400$	0.016	0.012	0.013	0.133	0.054	0.024	0.774	0.494	0.169
$n = 800$	0.027	0.020	0.013	0.556	0.333	0.147	0.998	0.988	0.881
$n = 1600$	0.036	0.025	0.012	0.936	0.849	0.585	1.000	1.000	1.000
$n = 3200$	0.041	0.036	0.030	1.000	0.996	0.927	1.000	1.000	1.000
$n = 6400$	0.049	0.042	0.035	1.000	1.000	1.000	1.000	1.000	1.000
Cramér–von Mises test									
$n = 400$	0.008	0.007	0.007	0.149	0.105	0.048	0.912	0.790	0.465
$n = 800$	0.019	0.019	0.016	0.571	0.447	0.236	1.000	0.999	0.973
$n = 1600$	0.032	0.029	0.018	0.957	0.900	0.721	1.000	1.000	1.000
$n = 3200$	0.033	0.027	0.028	1.000	0.999	0.969	1.000	1.000	1.000
$n = 6400$	0.049	0.046	0.043	1.000	1.000	1.000	1.000	1.000	1.000

We consider three different values for the parameter δ, which gauges heteroscedasticity. Under the null hypothesis, $\delta = 0$, such that the regressor X_2 is independent of U, which then corresponds to v, and has a pure location shift effect. 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 with $\delta = 0.2$ and 0.5, such that errors U are heteroscedastic.

In the simulations, we consider five sample sizes from $n = 400$ to $n = 6400$. 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. Results are presented for three different regions $\mathcal{T}$ over which the quantile coefficients are estimated: $[0.05, 0.95]$, $[0.1, 0.9]$ and $[0.2, 0.8]$.[7]

The empirical rejection frequencies reported in Table I suggest that the bootstrap score tests have good size and power properties. In the presence of independent errors ($\delta = 0$), both the KS and CM 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. Under heteroscedasticity, the rejection probabilities correctly converge to 100% as n becomes larger. As expected, this happens at a faster pace for $\delta = 0.5$ than for $\delta = 0.2$. The power properties of the KS and CM tests are rather similar, albeit the latter become relatively more powerful in larger samples and/or for a higher δ. The empirical power increases as the range of quantiles considered increases and this holds true for both test statistics and both values of δ. Summing up, the KS and CM 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.

The results in Koenker and Xiao (2002) and Chernozhukov and Fernández-Val (2005) provide useful benchmarks. In order to compare our results with theirs we must take into account that we observe Y^* only for half of the sample. Therefore, it seems fair to compare our results for 1600 observations

---

[7] Whether a particular region $\mathcal{T}$ entails good testing power depends on the conditional outcome distribution in two ways. First, the location of deviations from the null plays a role. For example, if violations of independence are concentrated in the center of the distribution, using a smaller region is likely to be more powerful than the inclusion of extreme quantiles. Second, also the shape of the conditional distribution of Y affects power. Extreme quantiles are less informative when the tails of the distribution are fat.

with theirs for 800 observations. In this case, we see for instance that with $\delta = 0.2$ our empirical power is 0.936 and 0.957, the power in Koenker and Xiao (2002) is 0.982 (in the best case) and the power in Chernozhukov and Fernández-Val (2005) is 1. In general, the price to pay for allowing for non-random selection is moderate in this DGP.

4.2. A Calibrated Multivariate DGP

In this subsection we analyze the properties of our tests in a more realistic setup. We calibrate our DGP to fit several characteristics of the CPS data used in the application in Section 5. We draw the Monte Carlo samples of X_1 from the female sample distribution in 2011 (containing 45,296 observations). We consider a model with four regressors (experience, experience squared, a high school dummy and a college dummy) and three exclusion restrictions (number of children in three age brackets). These covariates are only a subset of those used in our application because we also consider samples of moderate sizes, which would lead to frequent multicollinearity problems with the 19 regressors of the application.

The conditional distribution of the potential wage is generated by a linear quantile model:

$$Y^* = \alpha(V) + X_2'\beta(V), V|X_1 \sim U(0, 1)$$

We consider two sets of functions for $\alpha(\cdot)$ and $\beta(\cdot)$. The first one satisfies the null hypothesis with a constant $\beta(\tau)$ set to the value of the median regression coefficient estimates in the male sample.[8] In the second case, we set the coefficients to the unrestricted quantile regression estimates in the male sample.[9] This ensures that the violation of the null hypothesis is realistic and allows us to see the power of our tests in a typical application. The selection equation is

$$D = 1\left(X_1'\gamma + \varepsilon > 0\right)$$

where $\varepsilon \sim N(0, 1)$ and γ is set to the probit estimates of the female participation equation in 2011. Finally, the copula between ε and V is Gaussian with a correlation coefficient of 0.233, which is the estimate obtained by Mulligan and Rubinstein (2008) in the period 1990–1995.

When there are several regressors, the identifying Assumptions 1–3 imply that the whole vector $\beta(\tau)$ is constant as a function of τ. Thus the tests suggested in Section 3 simultaneously use all coefficients and weight them by the inverse of the covariance matrix (of the difference between $\hat{\beta}(\tau)$ and $\hat{\beta}(0.5)$). In addition, we can also separately test whether each element of the vector is constant as a function of the quantile. This will inform the researcher about the sources of the rejection when the global test rejects the null hypothesis.

The results in Table II give the empirical rejection probabilities of our tests when the coefficients are homogeneous (empirical size) and heterogeneous (empirical power). Our tests are somewhat conservative, with none of the empirical sizes being above the theoretical size of 5%. However, under the null hypothesis the rejection probability does converge to the intended one even if it does not yet attain it with 6400 observations. The power of the tests is very low when $n = 400$ but attains 98.6% when $n = 6400$ for the KS statistic, which tends to be more powerful than the CM statistic. The dataset

---

[8] $\alpha(\cdot)$ and $\beta(\cdot)$ are calibrated using the male sample to avoid the sample selection problem, as the employment probability is high for men. If we took the female sample then we could find violations of the independence assumption due to the selection bias. Alternatively, we could use the female sample and correct for the selection bias but this would require making distributional assumptions that seem quite arbitrary.

[9] We approximate the $U(0, 1)$ distribution of V with a discrete uniform distribution with 1000 points. This means that 1000 quantile regression are used to approximate the conditional distribution.

Table II. Empirical rejection frequencies for 5% bootstrap tests: empirical data, 1000 replications, 250 bootstrap draws

	$\tau \in [0.05, 0.95]$					$\tau \in [0.10, 0.90]$					$\tau \in [0.20, 0.80]$				
	all	exp	exp²	hs	col	all	exp	exp²	hs	col	all	exp	exp²	hs	col
Kolmogorov–Smirnov test: size															
$n = 400$	0.033	0.006	0.007	0.017	0.005	0.026	0.004	0.003	0.021	0.000	0.028	0.002	0.004	0.015	0.002
$n = 800$	0.037	0.006	0.007	0.032	0.005	0.032	0.004	0.006	0.025	0.004	0.028	0.004	0.005	0.022	0.005
$n = 1600$	0.021	0.020	0.023	0.024	0.016	0.009	0.016	0.013	0.018	0.010	0.009	0.009	0.009	0.015	0.011
$n = 3200$	0.030	0.023	0.030	0.029	0.034	0.017	0.020	0.022	0.017	0.031	0.010	0.019	0.019	0.015	0.028
$n = 6400$	0.035	0.041	0.043	0.022	0.041	0.025	0.031	0.030	0.024	0.038	0.022	0.027	0.026	0.017	0.035
Kolmogorov–Smirnov test: power															
$n = 400$	0.026	0.006	0.004	0.019	0.009	0.025	0.004	0.002	0.020	0.003	0.023	0.004	0.002	0.019	0.002
$n = 800$	0.042	0.015	0.013	0.020	0.089	0.028	0.011	0.010	0.015	0.058	0.022	0.009	0.009	0.014	0.042
$n = 1600$	0.229	0.112	0.069	0.043	0.346	0.080	0.047	0.038	0.018	0.136	0.020	0.020	0.022	0.009	0.043
$n = 3200$	0.694	0.261	0.156	0.109	0.755	0.340	0.150	0.105	0.053	0.421	0.090	0.093	0.072	0.022	0.146
$n = 6400$	0.986	0.518	0.343	0.178	0.972	0.808	0.329	0.228	0.091	0.815	0.329	0.200	0.155	0.041	0.352
Cramér–von Mises test: size															
$n = 400$	0.002	0.005	0.003	0.004	0.005	0.001	0.003	0.002	0.002	0.003	0.000	0.001	0.001	0.003	0.004
$n = 800$	0.004	0.012	0.009	0.007	0.012	0.003	0.011	0.009	0.006	0.010	0.002	0.009	0.007	0.005	0.009
$n = 1600$	0.011	0.020	0.023	0.018	0.022	0.003	0.022	0.020	0.011	0.020	0.003	0.014	0.016	0.010	0.015
$n = 3200$	0.026	0.039	0.040	0.029	0.042	0.020	0.038	0.040	0.020	0.041	0.013	0.032	0.035	0.015	0.041
$n = 6400$	0.034	0.043	0.035	0.038	0.043	0.032	0.034	0.034	0.032	0.039	0.022	0.029	0.029	0.024	0.038
Cramér-von-Mises test: power															
$n = 400$	0.002	0.008	0.005	0.003	0.013	0.001	0.005	0.003	0.002	0.010	0.001	0.003	0.001	0.003	0.004
$n = 800$	0.014	0.017	0.014	0.009	0.080	0.009	0.015	0.012	0.008	0.061	0.007	0.013	0.010	0.007	0.048
$n = 1600$	0.134	0.095	0.071	0.033	0.260	0.068	0.076	0.057	0.019	0.157	0.014	0.041	0.035	0.009	0.071
$n = 3200$	0.491	0.212	0.157	0.069	0.570	0.298	0.166	0.126	0.042	0.380	0.101	0.104	0.098	0.022	0.176
$n = 6400$	0.904	0.455	0.338	0.106	0.901	0.700	0.361	0.284	0.069	0.717	0.361	0.246	0.206	0.051	0.364

used in our application contains 45,296 observations such that we can expect our tests to be extremely powerful. The widest range of quantiles that we considered, [0.05, 0.95], delivers the best performance.

The tests of the constancy of a single coefficient are also slightly conservative. When n is large the empirical power of all tests are above 5%, which shows that all variables contribute to the rejection of the null hypothesis. The return to a college degree is the most heterogeneous coefficient, with rejection rates close to those obtained when jointly testing all variables.

5. FEMALE WAGE DISTRIBUTION IN THE USA

The estimation of the female wage function is the most frequent application of sample selection estimators (see, for example, Buchinsky (1998a) and Mulligan and Rubinstein (2008). We use the 2011 Merged Outgoing Rotation Groups of the CPS, which offer a good measure of hourly wages. Similarly to many examples in the wage regression literature, we restrict our sample to white non-Hispanic adults who are between 25 and 54 years old and exclude individuals who are self-employed, have allocated earnings, or work in the military, agricultural and public sectors.

We classify females as either working or not, according to whether they worked at least 35 hours during the week preceding the surveys. The dependent variable, Y, is the log hourly wage. Our vector of regressors, X_2, contains five educational dummies (at least high school, some college, associate degree, bachelor degree and advanced degree), potential experience and potential experience squared as well as their interaction with educational attainment in years, three regional dummy variables and an indicator for being married. The vector X_1 consists of the same elements plus the numbers of

Table III. Female wage distribution in the USA

	Selection equation	Outcome equation: Buchinsky (1998a)		
	Klein and Spady	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
Experience	0.013 (0.012)	0.021 (0.003)	0.031 (0.003)	0.034 (0.003)
Experience squared	−0.001 (3e−4)	−3e−4 (7e−5)	−5e−4 (7e−5)	−5e−4 (8e−5)
High school	1.848 (0.255)	0.147 (0.032)	0.260 (0.033)	0.273 (0.035)
Some college	0.216 (0.071)	0.117 (0.017)	0.132 (0.015)	0.146 (0.018)
Associate	0.630 (0.100)	0.088 (0.020)	0.128 (0.020)	0.203 (0.022)
Bachelor	0.234 (0.078)	0.229 (0.020)	0.255 (0.021)	0.230 (0.024)
Advanced	0.437 (0.097)	0.232 (0.023)	0.236 (0.023)	0.241 (0.024)
Exp. × education	−0.005 (0.002)	2e−3 (5e−4)	3e−3 (5e−4)	4e−3 (7e−4)
Exp.^{2} × education	5e−5 (6e−5)	−4e−5 (1e−5)	−6e−5 (1e−5)	−8e−5 (2e−5)
Married	−0.392 (0.065)	0.035 (0.013)	0.033 (0.012)	0.020 (0.014)
# children [0, 2]	−1.000			
# children [3, 5]	−0.530 (0.130)			
# children [6, 13]	−0.249 (0.067)			
Married × # children [0, 2]	−0.262 (0.152)			
Married × # children [3, 5]	−0.339 (0.113)			
Married × # children [6, 13]	−0.336 (0.069)			

Note: Column 1 contains the coefficients of the selection equation obtained by the Klein and Spady (1993) estimator. Columns 2–4 show the quantile regression coefficients obtained by the Buchinsky (1998a) estimator. Standard errors are reported in parentheses. Three regional dummy variables are also part of the regressors but are not shown, for brevity. These coefficients are consistent only if the independence assumption is satisfied or if selection is random. Our tests in Table IV reject the independence assumption.

children in three age brackets (0–2, 3–5 and 6–13 years) and their interaction with the marital indicator.[10] Descriptive statistics can be found in Appendix B. We use the CPS earning weights for all our calculations. We implement our test procedure as described in Section 3. The estimates of the selection equation by the Klein and Spady estimator, reported in Table III, show that the coefficients on all six variables excluded from the outcome equation are strongly significant. The signs of the other coefficients are also as expected, with a higher employment probability for non-married and high-educated women. The bandwidth, determined by GCV, was set to 0.25 times the standard deviation of the fitted index $X_1'\hat{\gamma}$.

Table III also contains the 0.25, 0.5 and 0.75 quantile coefficients estimated using the Buchinsky (1998a) estimator. Figure 3 shows all quantile coefficients between the 2nd and the 98th percentile of four of the most important covariates. The order of the power series in the inverse Mill's ratio, also determined by GCV, is 5, which is large compared to the literature. Note that the coefficients are consistent only if the independence assumption is satisfied or if selection is random.[11] They are provided here because they serve—among other quantile coefficients—as a basis for our tests. The linear term in experience is more or less monotonically increasing as we move up the distribution. At the same time, the quadratic term is decreasing as a function of the quantile. In other words, the whole shape of the experience–wage function is changing with the quantile. This is a clear hint against the independence assumption. In addition, the returns to education are also increasing in the quantile.

---

[10] Owing to data coding, in about 5% of cases we are only able to determine lower and upper bounds on the number of children in the three age brackets. We impute the average between the lower and upper bound in these cases. In contrast, the total number of children is always known.

The excluded regressors are strictly speaking discrete. However, in our dataset we observe 77 different combinations of values for these excluded variables. Therefore, the theoretical results should give a good approximation.

Huber and Mellace (2014) use novel hypothesis tests to check the validity (in particular, the exclusion restriction) of the number of children as instrument for female labor supply. They found no statistical evidence for its violation in four different empirical applications, including data from the CPS. As a word of caution, however, note that even asymptotically these tests cannot detect all possible violations of instrument validity, as they rely on a partial identification approach.

[11] We focus here on our test of the conditional independence assumption. We also tested for random selection and clearly rejected random selection at the mean and at most quantiles (results available from the authors).
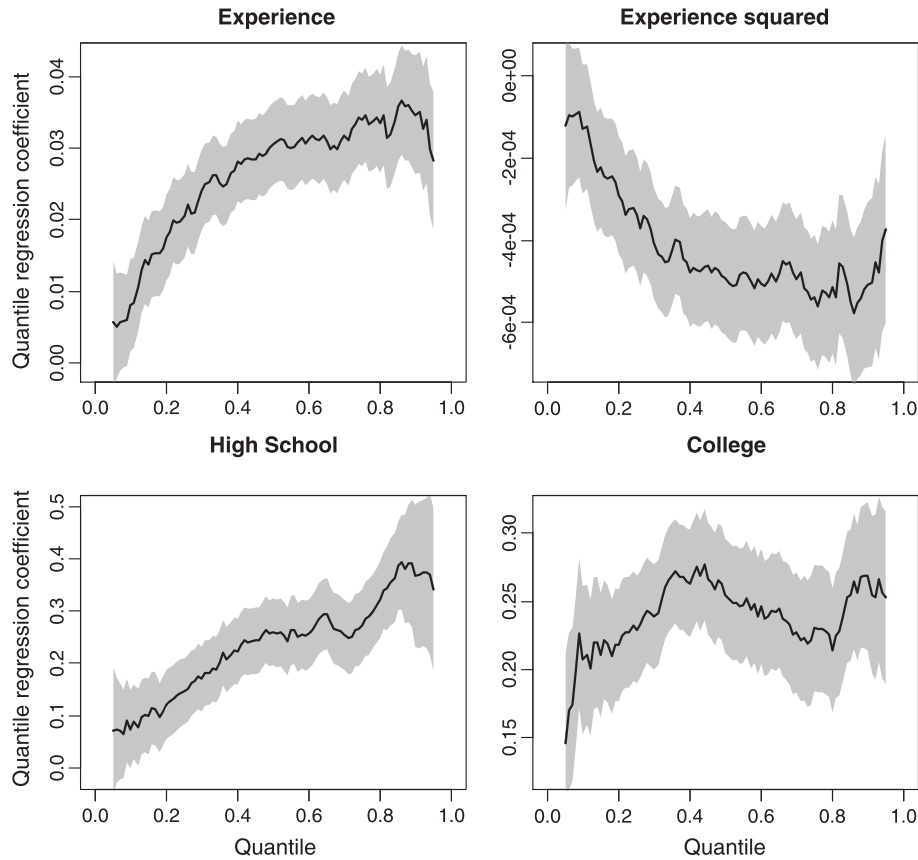
Figure 3. Quantile regression coefficients corrected for sample selection. The coefficients are estimated using the estimator of Buchinsky (1998a). Pointwise 95% confidence intervals are also plotted. In addition to the four variables whose coefficients are plotted, the vector of regressors also contains: three regional dummy variables, indicators for some college, associate degree, advanced degree and being married, and interactions between experience and experience squared with education in years. These coefficients are consistent only if the independence assumption is satisfied or if selection is random. Our tests in Table IV reject the independence assumption

The returns are becoming flatter as a function of the quantile only for the highest levels of education, perhaps due to a more homogeneous population.

Table IV reports the p-values of our tests of the conditional independence assumption. The null hypothesis tested in the first row is the constancy of the whole quantile regression vector, i.e. the validity of the model defined by Assumptions 1–9. The remaining rows give the p-values for each coefficient separately. The results are reported for two choices of block size ($b = n$ and $b = 20 + n^{1/2.01}$) and two ranges of quantiles ($\mathcal{T} = [0.02, 0.98]$ and $\mathcal{T} = [0.2, 0.8]$).[12] The number of bootstrap replications B was set to 1000.

The hypothesis of the validity of the sample selection model defined in Section 2 is clearly rejected in any setup as the p-values are always zero. A one-by-one investigation of the covariates shows that this rejection is mainly due to heterogeneous returns to experience and to lower-level education degrees (in particular, high school and associate degrees). These results and those for two other datasets reported in the working paper version of this article (Huber and Melly, 2011) cast serious doubts on the

---

[12] Not reported are the results for two other block sizes and two other ranges of quantiles, which are extremely similar.

Table IV. Empirical application: test results

Statistic	Kolmogorov–Smirnov				Cramér–von Mises			
$\tau \epsilon$	[0.02, 0.98]		[0.20, 0.80]		[0.02, 0.98]		[0.20, 0.80]	
b (block size)	45,296	227	45,296	227	45,296	227	45,296	227
All variables	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Experience	0.000	0.000	0.002	0.011	0.000	0.000	0.013	0.019
Experience^2	0.006	0.018	0.130	0.159	0.015	0.024	0.237	0.244
High school	0.003	0.048	0.010	0.046	0.000	0.008	0.016	0.029
Some college	0.002	0.014	0.422	0.381	0.028	0.028	0.781	0.722
Associate	0.007	0.027	0.004	0.014	0.000	0.000	0.000	0.000
Bachelor	0.113	0.179	0.387	0.351	0.195	0.218	0.275	0.300
Advanced	0.188	0.219	0.303	0.270	0.386	0.349	0.628	0.553
Exp.× education	0.338	0.317	0.242	0.190	0.131	0.169	0.076	0.113
Exp.^{2} × education	0.438	0.343	0.291	0.223	0.233	0.221	0.172	0.171
Married	0.140	0.161	0.354	0.311	0.151	0.149	0.323	0.304

Note: 45,296 observations. Critical values were obtained using 1000 bootstrap repetitions.

validity of traditional sample selection models in female wage regressions. The conclusion suggests a few tracks to explore without assuming homogeneous coefficients.

6. CONCLUSION

Assuming additivity and conditional independence of the error term in the outcome equation is rather restrictive. It implies that all individuals with the same selection probability react identically to changes in the observable characteristics. In economic models, the unobservable random terms are not simply measurement errors but they often have important economic interpretations. The recent econometric literature has relaxed many restrictions on the interaction between observed and unobserved variables. Advances have been reported in selection on observables, instrumental variables and panel data models, among many others (for a discussion see, for example, Matzkin, 2007).

Somewhat surprisingly, the sample selection model has been excluded from this trend. The consistency of almost all sample selection estimators hinges on the additivity and conditional independence of the error term. This is also the case in the quantile regression model of Buchinsky (1998a). However, in the quantile regression framework the conditional independence assumption implies that the quantile slope coefficients are equal to the mean slope coefficients. 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 across the coefficients estimated at distinct quantiles have merely proven the violation of the underlying assumptions and the inconsistency of the estimator.

A general lesson to draw from this example is the danger of importing one-to-one mean recipes to quantile models. Another example is given by the fitted value approach in endogenous models, which is justified only by a similar independence assumption. The fitted value IV estimators suggested by Amemiya (1982), Powell (1983) and Chen and Portnoy (1996) are, therefore, not useful for analyzing heterogeneity (which was *not* the intention of their authors). Similarly, the control function estimators in Lee (2007), Blundell and Powell (2007) and Carneiro and Lee (2009) are consistent only if the coefficients do not vary across quantiles.

Given the importance of conditional independence for the identification of (mean and quantile) sample selection models, we propose the first formal test for this assumption. Our test is based on the conditional quantile regression process estimated by the procedure of Buchinsky (1998a), which is consistent under the null hypothesis, and compares the coefficients obtained at different quantiles. Monte Carlo simulations provide evidence on the satisfactory power and size properties of our tests. We clearly reject the conditional independence assumption in an application to recent US wage data.

What can be done if the independence assumption is rejected? Unfortunately, the parameters of interest are then no longer point identified. In our companion paper (Melly and Huber, 2012), we derive the sharp bounds on the quantile regression parameters when this assumption is no longer imposed. It appears that there are only two ways to recover point identification.[13] The first one is to impose a parametric assumption on (at least) the copula function between the error terms in the selection and outcome equations. Donald (1995) makes one step in this direction and proposes a two-step estimator that allows for conditional heteroscedasticity but requires the error terms to be bivariate normally distributed. Arellano and Bonhomme (2011) obtain point identification by parametrizing only the copula while keeping the marginal distributions of the error terms nonparametric. The second way to recover point identification is to use an identification-at-infinity argument. The latter requires not only that some subpopulation is observed with probability (close to) one but also that the slope parameters are identified from this subpopulation.

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 (2013) provide sharp bounds for several subpopulations. This is of interest in some applications but clearly not in all. For instance, it does not allow comparing wages of different population groups like male and female or white and non-white individuals. These wage comparisons were precisely the reason for which Gronau (1974) developed corrections for the sample selection bias.

REFERENCES

Ahn H, Powell J. 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, Chernozhukov V, Fernández-Val I. 2006. Vouchers for private schooling in Colombia. *Econometrica* **74**: 539–563.
Arellano M, Bonhomme S. 2011. Quantile selection models. Walras-Bowley Lecture, *North American Summer Meeting of the Econometric Society, 10 June 2011,* St Louis.
Blundell RW, Powell JL. 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.
Buchinsky M. 1998a. The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of Applied Econometrics* **13**: 1–30.
Buchinsky M. 1998b. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of Human Resources* **33**: 88–126.

---

[13] Naturally, it is also possible to add structure to the model or use auxiliary data. For instance, Chen and Khan (2003) impose de facto a new type of exclusion restriction—the availability of a regressor that affects the variance but not the location of the dependent variable—to identify a multiplicative heteroscedastic model.

Buchinsky M. 2001. Quantile regression with sample selection: estimating women's return to education in the U.S. *Empirical Economics* **26**: 87–113.

Carneiro P, Lee S. 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.

Chen L, Portnoy S. 1996. Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equations models. *Communications in Statistics: Theory and Methods* **25**: 1005–1032.

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

Chernozhukov V, Fernández-Val I. 2005. Subsampling inference on quantile regression processes. Part 2. *Sankhya: Indian Journal of Statistics* **67**: 253–276.

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

Chernozhukov V, Fernández-Val I, Melly B. 2013. Inference on counterfactual distributions. *Econometrica* **81**: 2205–2268.

Cosslett S. 1991. Distribution-free estimator of a regression model with sample selectivity. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Barnett W, Powell J, Tauchen G (eds). Cambridge University Press: Cambridge, UK; 175–198.

Craven P, Wahba G. 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, Newey WK, Vella F. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* **70**: 33–58.

Delecroix M, Hristache M, Patilea V. 2006. On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference* **136**: 730–769.

Donald SG. 1995. Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics* **65**: 347–380.

Escanciano JC, Zhu L. 2013. Set inferences and sensitivity analysis in semiparametric partially identified models. Working paper. Indiana University.

Escanciano JC, Zhu L. 2014. A simple data-driven estimator for the semiparametric sample selection model. *Econometric Reviews*. DOI:10.1080/07474938.2014.956577.

Gallant A, Nychka D. 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, Jurečková J. 1992. Regression quantile and regression rank score process in the linear model and derived statistics. *Annals of Statistics* **20**: 305–330.

Hall P, Sheather SJ. 1988. On the distribution of a studentized quantile. *Journal of the Royal Statistical Society, Series B* **50**: 381–391.

Härdle W, Hall P, Ichimura H. 1993. Optimal smoothing in single-index models. *Annals of Statistics* **21**: 157–178.

Heckman JJ. 1974. Shadow prices, market wages and labor supply. *Econometrica* **42**: 679–694.

Heckman JJ. 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.

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* **47**: 153–161.

Huber M, Mellace G. 2013. Sharp bounds on causal effects under sample selection. *Oxford Bulletin of Economics and Statistics*. DOI: 10.1111/obes.12056.

Huber M, Mellace G. 2014. Testing exclusion restrictions and additive separability in sample selection models. *Empirical Economics* **47**: 75–92.

Huber M, Melly B. 2011. Quantile regression in the presence of sample selection. Economics working paper 1109. University of St Gallen.

Klein RW, Spady RH. 1993. An efficient semiparametric estimator for binary response models. *Econometrica* **61**: 387–421.

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

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

Koenker R, Bassett G. 1982. Robust tests for heteroskedasticity based on regression quantiles. *Econometrica* **50**: 43–62.

Koenker R, Hallock KF. 2000. Quantile regression: an introduction. Discussion paper for *Symposium on Econometric Tools, Journal of Economic Perspectives* **15**: 143–156.

Koenker R, Xiao Z. 2002. Inference on the quantile regression process. *Econometrica* **70**: 1583–1612.

Lechner M, Melly B. 2010. Partial identification of wage effects of training programs. Working paper. Brown University.

Lee DS. 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.

Mammen E, Rothe C, Schienle M. 2012. Nonparametric regression with nonparametrically generated covariates. *Annals of Statistics* **40**: 1132–1170.

Matzkin R. 2007. Nonparametric identification. In *Handbook of Econometrics*, Heckman J, Leamer E (eds). Elsevier: Amsterdam; 5307–5368.

Melly B, Huber M. 2012. Nonseparable sample selection models. Working paper.

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

Mulligan CB, Rubinstein Y. 2008. Selection, investment, and women's relative wages over time. *Quarterly Journal of Economics* **123**: 1061–1110.

Newey WK. 2007. Nonparametric continuous/discrete choice models. *International Economic Review* **48**: 1429–1439.

Newey WK. 2009. Two-step series estimation of sample selection models. *Econometrics Journal* **12**: S217–S229.

Newey WK, Powell JL. 2003. Instrumental variable estimation of nonparametric models. *Econometrica* **71**: 1565–1578.

Newey WK, Powell JL, Walker J. 1990. Semiparametric estimation of selection models: some empirical results. *American Economic Review* **80**: 324–328.

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

Powell JL. 1986. Censored regression quantiles. *Journal of Econometrics* **32**: 143–155.

Powell JL. 1987. Semiparametric estimation of bivariate latent variable models. Working paper. University of Wisconsin–Madison.

van der Vaart A, Wellner J. 1996. *Weak Convergence and Empirical Processes*. Springer: New York.

APPENDIX A: PROOF OF PROPOSITION 1

Chernozhukov and Hansen (2006) present a class of instrumental variable estimators for the linear quantile regression model and develop its sampling theory. Their second contribution, which is the relevant one for our article, is to develop practical inference procedures for testing distributional hypotheses. They formulate the conditions for inference so that other estimators of the linear quantile regression model are permitted. They show in Appendix C that their IV estimator as well as the ordinary quantile regression estimator satisfy the high-level conditions for inference. We show in this Appendix that the sample selection corrected quantile regression estimator (5) also satisfies these conditions. Proposition 1 then follows from Theorem 5 in Chernozhukov and Hansen (2006).

For convenience, we restate here the four conditions for inference I.1–I.4 in Chernozhukov and Hansen (2006) but we use β instead of θ as a symbol for the vector of parameters to keep our notation consistent.

Assumption I.1. $R(\cdot)(\beta(\cdot) - r(\cdot)) = g(\cdot)$, where the functions $g(\tau), R(\cdot), r(\tau)$ are continuous and either (a) $g(\tau) = 0$ for all τ or (b) $g(\tau) \neq 0$ for some τ.

Assumption I.2. $\sqrt{n}\left(\hat{\beta}(\cdot) - \beta(\cdot)\right) \implies b(\cdot)$ and $\sqrt{n}(\hat{r}(\cdot) - r(\cdot)) \implies d(\cdot)$ jointly in $l^{\infty}(\mathcal{T})$, where $b(\cdot)$ and $d(\cdot)$ are jointly zero mean Gaussian functions that may have different laws under the null and the alternative.

Assumption I.3. The estimates admit linear representations:

$$\sqrt{n}(\hat{\beta}(\cdot) - \beta(\cdot)) = -J(\cdot)^{-1} 1/\sqrt{n} \sum_{i=1}^{n} l_i(\cdot, \beta(\cdot))\Psi_i(\cdot) + o_p(1)$$

and

$$\sqrt{n}(\hat{r}(\cdot) - r(\cdot)) = -H(\cdot)^{-1} 1/\sqrt{n} \sum_{i=1}^{n} d_i(\cdot, r(\cdot)) \Upsilon_i(\cdot) + o_p(1)$$

in $l^{\infty}(\mathcal{T})$, where $J(\cdot)$ and $H(\cdot)$ are constant invertible matrices, and $l_i(\cdot, \beta(\cdot))\Psi_i(\cdot)$ and $d_i(\cdot, r(\cdot))\Upsilon_i(\cdot)$ are i.i.d. mean zero for each τ.

Assumption I.4. (a) We have estimates $l_i\left(\cdot, \hat{\beta}(\cdot)\right)\hat{\Psi}_i(\cdot)$ and $d_i(\cdot, \hat{r}(\cdot))\hat{\Upsilon}_i(\cdot)$ that take realizations in a Donsker class of functions with a constant envelope and are uniformly consistent in τ in the $L_2(P)$ norm. (b) $Wp \to 1$, $E l_i(\tau, \beta(\tau)) f_i(\tau)|_{f=\hat{\Psi}} = 0$ and $E d_i(\tau, r(\tau)) f_i(\tau)|_{f=\hat{\Upsilon}} = 0$ for each i. (c) $E \|l_i(\tau, \beta) - l_i(\tau, \beta')\| < C \|\beta' - \beta\|$, $E \|d_i(\tau, r) - d_i(\tau, r')\| < C \|r' - r\|$ uniformly $\tau \in \mathcal{T}$ and in (β, β', r, r') over compact sets.

Our test fits into this framework by setting $R(\cdot) = R = [1, 1, \ldots, 1]$, a vector of 1 of length $d_X \equiv \dim(x_{2i})$, and $r(\cdot) = \beta(0.5)$. Our continuity assumptions imply that $\beta(\tau)$ is a continuous function of τ. Under the null hypothesis, $g(\tau) = 0$ for all τ and under the global alternative hypothesis $g(\tau) \neq 0$ for some τ. Thus Assumption I.1 is satisfied.

We must derive the limiting distribution of the sample selection corrected quantile regression process to show the validity of the remaining three conditions I.2–I.4. To do so, we build on the results of Lee (2007), who provided the pointwise limiting distribution for an estimator that closely resembles our estimator.[14] In fact, our estimator defined in equation (5) corresponds almost exactly to the IV estimator defined in equation (7) of Lee (2007). The first difference is that in our case the estimation is performed only in the selected population, whereas Lee (2007) uses the whole population. Our problem fits, however, into his framework by redefining the trimming function to be zero for observations with $D = 0$. The second difference is that the first step fitted values are estimated by the Klein and Spady (1993) estimator in our case, whereas Lee (2007) uses traditional quantile regression. This is not an issue because Lee (2007) states the assumptions in a way that allows for other $\sqrt{n}$ consistent and asymptotically linear estimators.

Our assumptions are uniform in $\tau \in \mathcal{T}$ and conditional on $D = 1$ versions of the assumptions of Lee (2007). We do not need trimming with respect to X_2 because we make a compactness assumption on the distribution of X_2. In addition, by Assumption 3, $f_U\left(u|x_{1i}'\gamma_0, x_{1i}\right) = f_U\left(u|x_{1i}'\gamma_0\right)$, such that we can simplify the definition of m_i. We apply Theorem 3.1 in Lee (2007) and obtain the following influence function (equations (39) and (40) in Lee, 2007):

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta_0\right) = A \Phi_J(\tau)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\ell_{2i}(\tau) + \Delta_{2\gamma}\Delta_P^{-1}k_i\right) + o_p(1) \tag{13}$$

where A denotes the $[d_X \times (d_X + J)]$ matrix such that $A = \left(I_{d_X}, 0_{d_X+J}\right)$, where I_{d_X} is the d_X-dimensional identity matrix, 0_{d_X+J} is the $[d_X \times J]$ matrix of zeros, d_X is the dimension of x_{2i} and

$$\ell_{2i}(\tau) = t_i\left(\tau - 1\left(y_i < x_{2i}'\beta_0 + h_\tau\left(x_{1i}'\gamma_0\right)\right)\right) p_{Ji},$$

$$\Delta_{2\gamma} = E\left[t_i f_\varepsilon\left(Q_\varepsilon\left(\tau|x_{1i}'\gamma_0\right) \frac{dh_\tau\left(X_1'\gamma_0\right)}{d\gamma_0} p_{Ji}\right]$$

---

[14] Lee (2007) developed his estimator in an instrumental variable model but noticed the similarities with the sample selection model. When applied to the sample selection model, his limiting distribution is the same as that obtained by Buchinsky (1998a).

The matrix A selects the relevant components of the $d_X + J$ score functions. The influence function can equivalently be written as

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) = \Delta_{b,\tau}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\ell_i(\tau) + \Delta_\gamma \Delta_P^{-1} k_i\right) + o_p(1)$$

in equation (11). This satisfies Assumption I.3 of Chernozhukov and Hansen (2006.[15]

We estimate the elements in the score function (11) by $\hat{\Delta}_{b,\tau}, \hat{\ell}_i(\tau), \hat{\Delta}_\gamma, \hat{\Delta}_P$, and $K(x_{1i}, d_i, \hat{\gamma})$. Under Assumptions 1-9, arguments identical to those used in Lemma A.9 in Lee (2007) imply that $\hat{\Delta}_{b,\tau}$ is uniformly consistent for the non-singular matrix $\Delta_{b,\tau}$. Similarly, arguments identical to those used in Lemma A.11 in Lee (2007) imply that $\hat{\Delta}_\gamma$ is consistent for Δ_γ. Since the trimming function is 0 when $x_{1i}'\hat{\gamma}$ is outside of a compact set and x_{1i}' takes value in a compact set, exploiting the boundedness and monotonicity of the indicator function, it follows that the class of functions

$$T\left(\tau - 1\left(Y < X_2'\beta + h_\tau\left(X_1'\gamma\right)\right)\right) M, \beta \in R^{d_X} \text{ and } \tau \in \mathcal{T}$$

is Donsker. In addition, $\Delta_\gamma, \Delta_P^{-1}$, and k_i are bounded by assumption. Since sums and products of Donsker functions are Donsker, this verifies Assumption I.4(a) of Chernozhukov and Hansen (2006). Assumption I.4(b) of Chernozhukov and Hansen (2006) holds by our Assumptions 1–3, which imply that $x_{2i}'\beta_0 + h_\tau\left(x_{1i}'\gamma_0\right)$ is the τ conditional quantile of y_i given x_{1i} and $d_i = 1$. Assumption I.4(c) of Chernozhukov and Hansen (2006) holds by the bounded density assumption.

By the central limit theorem for empirical processes indexed by Donsker classes of functions (van der Vaart and Wellner, 1996), $\Delta_{b,\cdot} \sqrt{n}\left(\hat{\beta}(\cdot) - \beta_0\right)$ converges in distribution to a zero mean Gaussian process $z(\cdot)$ with covariance function $\Sigma(\tau, \tau')$:

$$\Sigma(\tau, \tau') = \left[\min(\tau, \tau') - \tau\tau'\right] \cdot E\left[m_i m_i'\right] + \Delta_\gamma \Delta_P^{-1} var(k_i)$$

This satisfies Assumption I.2 of Chernozhukov and Hansen (2006). The result of Proposition 1 follows from Theorem 5 in Chernozhukov and Hansen (2006), parts (a) and (b), where the continuity of the distribution of the test statistics follows from part (c).

APPENDIX B: DESCRIPTIVE STATISTICS

Table B.I. Means of the relevant variables

	All observations	Employed full time	Not employed full time
High school	0.904	0.951	0.865
Some college	0.626	0.694	0.570
Associate	0.447	0.521	0.387
Bachelor	0.318	0.374	0.273
Advanced	0.089	0.113	0.070
Experience	21.76	21.41	22.044
Married	0.718	0.668	0.759
# children $[0, 2]$	0.162	0.103	0.211
# children $[3, 5]$	0.204	0.138	0.259
# children $[6, 13]$	0.496	0.392	0.582
Number of obs.	45,296	20,789	24,507

---

[15] Buchinsky (1998a) obtains the same score function in a more heuristic way.