Nonseparable Sample Selection Models

Blaise Melly, Martin Huber

Preliminary and incomplete!

First draft: December 2006, Last changes: February 1, 2011

Abstract:

This paper analyzes the nonparametric identification of nonseparable models in the presence of sample selection. Most existing sample selection models severely restrict the effect heterogeneity of the observed variables on the outcome distribution. In contrast, we allow for essential heterogeneity by assuming the outcome to be a nonparametric and nonseparable function of the explanatory variables and the disturbance term. We impose a quantile restriction on the disturbance term to derive sharp bounds on the functions of interest such as the partial effects. The identified set shrinks to a single point if separability holds or if some observations are observed with probability one. We also provide a simple estimator for the identified set in the linear quantile regression model and apply it to female wage data.

Keywords: sample selection, nonseparable model, quantile regression, partial identification
JEL classification: C12, C13, C14, C21

We have benefited from comments at the conferences 'Inference and Tests in Econometrics' at GRECAM, Marseille and 'The Evaluation of European Labour Market Programmes' at CREST, Paris and by seminar participants at Brown university, Columbia University, Duke University, and Paris I. Addresses for correspondence: Blaise Melly, Brown University, Department of Economics, Providence, RI, USA, blaise_melly@brown.edu; Martin Huber, SEW, University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland, martin.huber@unisg.ch.

1 Introduction

The sample selection problem, which was discussed by Gronau (1974) and Heckman (1974), among many others, arises whenever the outcome of interest is only observed for some subpopulation that is non-randomly selected even conditional on observed factors. Sample selection is an ubiquitous phenomenon in empirical research, e.g., when estimating the returns to schooling based on a selective subpopulation of working or the effect of a training on test scores where some individuals abstain from the test in a non-random manner. It constitutes essentially the same identification problem as attrition in the outcome related to unobserved factors.

The sample selection literature controls for this issue by instrumenting selection (and imposing a single index restriction), i.e., by exploiting a variable that shifts selection but does supposedly not enter the outcome equation. in addition, the parametric, semiparametric and nonparamteric contributions assume homogenous effects of the explanatory variables on the outcome, see Heckman (1979), Ahn & Powell (1993), Newey (1991), Das, Newey & Vella (2003) among many others. Due to the additivity of the regression function and the disturbance term, the effects can be identified for the whole population. While this eases identification separability is a very unattractive restriction given the importance of effect heterogeneity in empirical problems which has received much attention in particular in the field of treatment evaluation.

Newey (2007) discusses point identification in nonparametric and nonseparable models and is, therefore, more general in this respect. However, the price to pay is that the effects are only identified for the selected population (with observed outcomes) which may not be of substantial interest for various reasons. Firstly, results for a selective subpopulation generally appear to be less interesting than for the whole population. Secondly, even for the selected population, comparisons over time (e.g. of the returns to schooling) or across regions are not feasible in general due to its time and location dependent constitution. Finally, if one assumes a particular model (e.g., linearity) for the whole population it does not necessarily transfer to the selected population.

For these reasons, our object of interest is the whole population. Trivially, we could assume effect homogeneity in the non-selected and selected populations. This would be in the spirit of the classical sample selection literature mentioned above which invokes full independence (conditional on the participation probability) between the regressors and the disturbances. On the other side of the spectrum, allowing for arbitrary heterogeneity and nonseparability only identifies the

Manski (1989 and 1994) bounds of the effects, which are sharp, but usually very wide in typical applications.

The main contribution of this paper is the proposition of a third way between these two extreme approaches. We essentially impose the same assumptions as in conventional sample selection models, as in Das et al. (2003) and Ahn & Powell (1993) for the nonparametric and semiparametric models, respectively, with the exception of separability. Put differently, we only invoke a particular quantile restriction on the disturbance term instead of full independence (which also requires separability). This allows for heteroscedasticity and all other types of dependence between the regressors and the outcome. Under these conditions, we obtain more informative bounds than the worst case bounds of Manski. An appealing feature of our bounds is that the identified interval collapses to a single point in two special cases: if separability between the disturbances and the regressors holds and/or if a subset of observations have a selection probability of one.

The first case is obvious since we are back in the classical sample selection model. It is nevertheless important as it implies that the upper and lower bounds will be quite close when there is only limited dependence. The second case is an example of 'identification at infinity', see Chamberlain (1986). This approach has been used by Heckman (1990) and Andrews & Schafgans (1998) to identify the constant in the classical sample selection model. In an heterogenous outcome model, it even identifies the slope coefficients. Our bounds also generalize this strategy to the case where some observations are observed with a high, but below one, probability. In this case, a narrow interval for the quantile coefficients is identified even when point identification is not feasible.

The bounds take the form of the smallest and largest coefficients obtained from a family of local quantile regressions over a range of quantiles. We derive the identification results for a general nonparametric model. To be more specific about estimation and inference, we consider in details the parametric linear case. As an empirical illustration we apply our estimator to the female wage data of Mulligan & Rubinstein (2008).

The remainder of this paper is organized as follows. Section 2 defines the model and gives the sharp identified set for the structural function and its derivatives. Section 3 offers a new anatomy of the sample selection model that nests many sample selection models considered in the literature. Therefore, it offers the possibility to assess the conditions required for point identification in the

whole population and in particular subpopulations. We also consider the similarities and the differences with the instrumental variable model. Section 4 discusses estimation and inference under the linear regression model. Section 5 revisits an empirical application of sample selection models. Section 6 concludes.

2 Identification of nonseparable sample selection models

2.1 The model

This section introduces the model and the identifying assumptions. We consider a nonparametric, nonseparable sample selection model. It is more general than the parametric and semiparametric sample selection models considered by Heckman (1979), Powell (1987), Ahn & Powell (1993), and Newey (1991), among others, as well as the nonparametric, but separable model of Das et al. (2003). The outcome equation is the same as in Manski (2003) or Chesher (2003):

Assumption 1 (outcome equation)

$$Y^* = m(X, \varepsilon) \tag{1}$$

where Y is the scalar dependent variable, X is the vector of regressors and ε is the scalar disturbance term. $m(\cdot)$ denotes a general regression function that we restrict to be strictly increasing in ε. Sometimes, it will be useful to separately consider a particular regressor, X_1 with support $\mathcal{X}_1$, separately from the remaining observed variables X_{-1} with support $\mathcal{X}_{-1}$. Note that Y is only observed conditional on selection, i.e.,

$$Y = Y^* \text{ if } S = 1 \text{ and is not observed otherwise,} \tag{2}$$

where S denotes the binary selection indicator.

Equation (1) nests the nonparametric separable model in in Das et al. (2003) where

$$m(X, \varepsilon) = g(X) + \varepsilon$$

and the separable parametric model

$$m(X, \varepsilon) = X\beta + \varepsilon$$

as special cases. In Section 5 we consider in details the parametric nonseparable quantile model

$$m(X, \varepsilon) = X\beta(\varepsilon).$$

While most contributions are concerned with the the identification of (conditional) mean effects, this paper focusses on the quantile function. The first reason is that in many applications, distributional features appear to be particularly relevant. E.g., the returns to a training or schooling are likely to differ at different parts of the outcome distribution. In fact, the earnings effects for disadvantaged individuals at lower ranks of the income distribution might bear more policy relevance than the mean returns. Secondly, in our nonparametric nonseprable framework, quantiles are naturally bounded while bounding mean functions requires bounding the support of the outcome. Thirdly, quantile regression allows for (partial) identification of the structural function and not just its average.

In the presence of sample selection and in the absence of a fully parametric model, Chamberlain (1986) shows that the constant is identfied only by an 'identification at infinity´ argument. We will not make such an assumption.[1] While we would prefer identifying the quantile function of Y, identification of the quantile function up to a constant is sufficient in many applications.

To this end, define $Q_A(\tau)$ as the τth quantile of some variable A and $Q_A(\tau|B=b)$ as the τth conditional quantile of A given that $B=b$, respectively, with $\tau \in (0,1)$. $F(\cdot), F(\cdot|\cdot)$ denote the cdf and the conditional cdf, respectively. We impose the following quantile insensitivity assumption w.r.t. ε:

Assumption 2 (quantile insensitivity)

$$Q_\varepsilon(\tau|X, Z) = Q_\varepsilon(\tau) = \tau \tag{3}$$

where Z is observed with support $\mathcal{Z}$. Note that Z is excluded from equation (1) which will be crucial for identification. The first equality implies that the τth quantile of ε is independent of X and Z. It is a selection on observable assumption. The second equality, if applied to all quantiles τ, imply that ε is $U(0,1)$. This is not restrictive in the sense that the function $m(\cdot)$ is identified only up to a strictly increasing transformation. The normalization is convenient for our subsequent discussion because it implies that, in the absence of sample selection, .we could

[1] Of course, we don't exclude this possibility but our results hold generally.

identify the structural function evaluated at $\varepsilon = \tau$ by the τ conditional quantile regression:

$$\begin{aligned} Q_Y(\tau|X=x) &= m(x, Q_\varepsilon(\tau|X=x)) \\ &= m(x,\tau). \end{aligned}$$

Apart from $m(x,\tau)$, we might also want to learn about the partial effect of one particular regressor, say X_1, on the quantile function of Y. If X_1 is continuous, this is the partial derivative of the quantile function with respect to X_1. If X_1 is discrete, we are interested in the ceteris paribus difference between the quantile functions evaluated at distinct values $X_1 = x_1$ and $X_1 = x_1'$. To summarize, the parameters to be identified are (for some $\tau \in (0,1)$):

$$\begin{aligned} &m(x,\tau), \\ &\frac{\partial m(x,\tau)}{\partial x_1}, \\ &m(x_1, x_{-1}, \tau) - m\left(x_1', x_{-1}, \tau\right). \end{aligned}$$

In the absence of sample selection, the identification of these estimands would be standard, as discussed for instance in Manski (2003). However, Y is only observed conditional on $S = 1$. For this reason, we need to impose further structure on the selection mechanism. As in Ahn & Powell (1993), we don't make any parametric assumption about the selection probability, $P = \Pr(S=1|X,Z)$ and $p(x,z) = \Pr(S=1|X=x, Z=z)$. As in the conventional sample selection literature our identification strategy is based on an exclusion restriction for Z. For the identification of partial effects, Z needs to have sufficient predictive power on S as it requires observations with distinct X_1 but equal P (conditional on X_{-1} and $S = 1$). Put differently, Z must shift the selection probability such that common support of P across different values of X_1 holds (ceteris paribus) in the selected population. This is formally stated in Assumption 3.

Assumption 3 (first stage)

a) Continuous treatment effects: $m(x,\tau)$ is continuously differentiable with respect to x_1. For any $\varepsilon > 0$ and $p \in \mathcal{P}_{S=1,X=x}$, X_1 is a nondegenerate random variable conditionally on $|X_1 - x_1| < \varepsilon, X_{-1} = x_{-1}, P = p$ and $S = 1$.

b) Discrete treatment effects

$$Support(P|X_1 = x_1, X_{-1} = x_{-1}) \cap Support\left(P|X_1 = x_1', X_{-1} = x_{-1}\right) \neq \varnothing$$

Furthermore, we impose the following single index restriction on the distribution of the disturbances in the selected population:

Assumption 4 (index restriction)

$$F_{\varepsilon}\left(\tau|S=1,x,z\right)=F_{\varepsilon}\left(\tau|S=1,P\left(x,z\right)\right).$$

Assumption 4 requires the conditional quantile of the disturbance term in the selected population to only depend on the conditional selection probability. This is the quantile equivalent of an assumption invoked by Das et al. (2003), Newey (1991), and Powell (1987), among others, for the estimation of mean effects. Since this paper considers a nonseparable model, Assumption 4 does *not* imply that the distribution of Y given S and P is homoscedastic or independent of X. Unrestricted outcome heterogeneity is still allowed by equation 1, e.g., through interaction of X and ε. However, Assumption 4 imposes restrictions on the way observations are (not) selected. For instance, if there is positive selection w.r.t. the outcome at one value of X conditional on $P=p$, there must be the same amount of positive selection at any other value of X given $P=p$.

This assumption is untestable because the outcome is never observed for non-selected units. In contrast, the stronger full independence assumption in separable models is testable because it restricts the distribution of Y in the selected sample. We propose such a test in our companion paper, Melly & Huber (2010).

Finally, we assume that the outcome Y is a continuous random variable in the selected sample to ensure that the observed quantiles are well-defined:

Assumption 5 (continuity) $F_{Y}\left(m\left(x,\tau\right)|X=x,S=1\right)$ is continuously differentiable with respect to y with density $f_{Y}\left(m\left(x,\tau\right)|X=x,S=1\right)$ which is bounded above and away from 0.

2.2 Sharp identified sets

Most sample selection models considered in the literature are separable in the error term. This crucial assumption allows for the point identification of partial effects. Identification at infinity can then be used in addition to point identify the constant term. On the other side of the spectrum, Manski (1989 and 1994) derives the worst case bounds for the conditional mean and

quantile functions implied by a minimal set of assumptions. These bounds are typically very wide, especially when one is interested in the effects of a change in X on the conditional mean or quantile of Y. In this case, they are simply the difference between the upper (lower) bound at one value and the lower (upper) bounds at another value of X. The width of the bounds is the sum of the widths of the bounds at both values. Bounding the partial derivative of the conditional mean (or quantile) requires bounds on the partial derivative of the function, which is often not available. In the absence of such a restriction, the partial derivatives are unbounded. Thus, these bounds do not collapse to a point when we have separability.

In this paper, we suggest an intermediate path between the worst case bounds of Manski and the classical sample selection models. The main difference with Manski is that we impose the nonparametric index restriction defined in Assumption 4. This restriction is implied by the independence assumption made by most sample selection models, which are separable in the error term. We derive the sharp bounds on the structural function and on the derivatives of this function. The nice feature of these bounds is that they collapse to a point when the outcome is separable in the disturbance term or when there is identification at infinity. Thus, there is no cost in terms of identification by relaxing these two assumptions.[2]

Theorem 1 gives the main result of this paper. It states that given our assumptions, the parameters of interest (i.e., the structural functions and the partial effects) at the τth quantile of the parameter of interest in the whole population (given X) lies within the intersection of a particular interval of quantiles defined upon the conditional outcome distribution (given X and P) of the selected population, which is observed. The admissible ranks of the interval, denoted by Θ_p, are obtained in two steps. First, we derive the admissible interval of quantiles ($\Theta_{\tau,x}$) for the conditional outcome distribution of the whole population at some rank θ by invoking Assumption 2 and the results in Manski (1994) on sharp bounds. Together they imply that θ must lie within $\left[\frac{\tau-(1-P(x,z))}{P(x,z)}, \frac{\tau}{P(x,z)}\right]$ and that $\Theta_{\tau,x}$ is obtained by taking the intersection over Z. Obviously, if $P = 1$ for some value of Z the interval collapses to a point. Second, we evaluate the conditional outcome distribution of the selected population at the values in $\Theta_{\tau,x}$, which are conditional on X, and finally obtain the admissible ranks Θ_p as the intersection across different values in X.

Theorem 1 (Identified set) *Assumptions 1, 2, 4 and 5 hold at some $\tau \in (0,1)$. The sharp*

[2] Of course, there are costs w.r.t. the precision of the estimator.

identified set for $m(x,\tau)$ *is given by*

$$m(x,\tau) \in \underset{p\in\mathcal{P}_{S=1,X=x}}{\cap} \{Q_Y(\theta|S=1,x,p) : \theta \in \Theta_p\}.$$

Assumption 3-a and existence of the derivatives imply the following sharp bounds for the τ^{th} *quantile treatment effect of the continuous variable* X_1

$$\frac{\partial m(x,\tau)}{\partial x_1} \in \underset{p\in\mathcal{P}_{S=1,X=x}}{\cap} \left\{ \frac{\partial Q_Y(\theta|S=1,x,p)}{\partial x_1} : \theta \in \Theta_p \right\}.$$

Assumption 3-b implies the following sharp bounds for the τ^{th} *quantile treatment effect of shifting* X_1 *from* x_1 *to* x'

$$m(x_1,x_{-1},\tau) - m\left(x_1',x_{-1},\tau\right) \in \underset{p\in\left(\mathcal{P}_{S=1,X=x}\cap\mathcal{P}_{X=x',S=1}\right)}{\cap} \left\{Q_Y(\theta|S=1,x,p) - Q_Y\left(\theta|S=1,x',p\right) : \theta\in\Theta_p\right\}.$$

Θ_p *is the identified set for* $F_\varepsilon(\tau|S=1,p)$ *and is given by*

$$\begin{aligned}
\Theta_p &= \underset{x\in\mathcal{X}_{S=1}}{\cap} \{F_Y(q|S=1,x,p) : q\in\Theta_x\}, \\
\Theta_x &= \underset{Z\in\mathcal{Z}_{S=1,X=x}}{\cap} \left\{ Q_Y(\theta|S=1,x,z) : \theta \in \left[\frac{\tau-(1-P(x,z))}{P(x,z)}, \frac{\tau}{P(x,z)} \right] \right\}.
\end{aligned}$$

Proof.

We first calculate the conditional bounds given $P=p$. To this end, assume that we know the scalar $\theta(p)$ that satisfies

$$F_\varepsilon(\tau|S=1,x,z) = F_\varepsilon(\tau|S=1,P(x,z)) = \theta(p).$$

This means that

$$Q_\varepsilon(\theta(p)|S=1,x,z) = Q_\varepsilon(\theta(p)|S=1,p) = \tau.$$

By m being strictly increasing in ε, Assumption 2, and the equivariance property of quantiles

$$\begin{aligned}
Q_Y(\theta(p)|S=1,x,z) &= m(x, Q_\varepsilon(\theta(p)|S=1,x,z)) \\
&= m(x, Q_\varepsilon(\theta(p)|S=1,p)) = m(x,\tau).
\end{aligned}$$

When considering a continuous regressor, the partial quantile effect (given the existence of the required derivatives) is

$$\begin{aligned}
\frac{\partial Q_Y(\theta(p)|S=1,x,z)}{\partial x_1} &= \frac{\partial m(x, Q_\varepsilon(\theta(p)|S=1,p))}{\partial x_1} + \frac{\partial m(x,t)}{\partial t}\frac{\partial Q_\varepsilon(\theta(p)|S=1,p)}{\partial x_1} \\
&= \frac{\partial m(x,\tau)}{\partial x_1}.
\end{aligned}$$

When considering a discrete regressor, the partial quantile effect is

$$Q_Y\left(\theta\left(p\right)|S=1,x,z\right)-Q_Y\left(\theta\left(p\right)|S=1,x',z\right)=m\left(x,\tau\right)-m\left(x',\tau\right)$$

Thus, point identification of the derivatives or of the discrete change is obtained conditional on knowing $\theta\left(p\right)$.

In our case, $\theta\left(p\right)$ is not known, but falls into an set of admissible ranks Θ_p. To show that this set is sharp, note that Manski (1994) proves the sharpness of identified set of quantiles Θ_x for $F_Y^{-1}\left(\theta|X=x\right)=m\left(x,\theta\right)$. Therefore, sharp bounds on $F_Y\left(m\left(x,\tau\right)|S=1,X=x,P=p\right)$ are given by $\{F_Y\left(q|S=1,X=x,P=p\right):q\in\Theta_x\}$. By m being strictly increasing in ε, Assumption 2, and the equivariance property of quantiles

$$\begin{aligned}F_Y\left(m\left(x,\tau\right)|S=1,x,p\right) &= F_\varepsilon\left(\tau|S=1,x,p\right)\\ &= F_\varepsilon\left(\tau|S=1,p\right)=\theta\left(p\right)\end{aligned}$$

Thus, the bounds on $F_Y\left(m\left(x,\tau\right)|S=1,X=x,P=p\right)$ are the bounds on $\theta\left(p\right)$ which do not depend on X. Therefore, we can take the intersection of the bounds given X which yields $\theta\left(p\right)\in\Theta_p$. We have, thus, shown that the bounds conditional on $P=p$ are sharp. Finally, as $m(x,\tau)$ is independent of P (or, equivalently, of Z, as X is fixed at x), the intersection of parameter bounds under Θ_p across different p conditional on $X=x,S=1$ gives the results. ■

We will discuss the implications of Theorem 1 in greater details in section 4. A first remark is that empty sets may be observed. The upper bound in the intersection of Θ_p over P can be lower than the lower bound, which disqualifies Z as valid instrument that may be excluded from the outcome equation. To see this, note that bounds crossing implies that $m(x,\tau)$ is in fact a function of P and, thus of Z because X is fixed at x. In this case, Y and Z are directly related such that the latter cannot be a valid instrument as assumed in the model.

Even though this provides us with a testable implication, the violation of the instrument validity does not necessarily provoke bounds crossing such that testing will not be uniformly powerful. Furthermore, since bounds can cross in finite samples even when they do not in the population, an estimate of the precision of the bounds is necessary in order to implement a formal test. Testing based on bounds crossing is considered by Blundell, Gosling, Ichimura & Meghir (2007). Kitagawa (2009) proposes a different approach to test for the instrument validity (or exclusion restriction) in sample selection models. He develops an inferential procedure for whether the integral of the envelope over the conditional densities of the selected Y given Z (for

X fixed) is larger than one. This is useful because an integrated envelope larger than one is equivalent to an empty identification set, which has already been noticed by Manski (2003).

2.3 Alternative models that produce the same bounds

Section 2.1 defined the least restrictive model that justifies the bounds presented in theorem 1. In this subsection we discuss slightly more restrictive models that have been considered in the literature and imply the same bounds.

While the selection equation was left unspecified in Section 2.1, our assumptions are implied by a familiar single crossing model. In the rather general form considered by Newey (2007), the selection indicator is generated by

$$S = 1\left(\upsilon \leq \Pi\left(X, Z\right)\right), \tag{4}$$

where υ is an unobserved factor and Π is a general function. It is assumed that $F_{\upsilon}(t)$ is strictly monotonic in t such that $P = F_{\upsilon}\left(\Pi\left(X, Z\right)\right)$. Furthermore,

$$\left(\varepsilon, \upsilon\right) \perp X, Z \tag{5}$$

where '$\perp$' denotes independence. This could be slightly relaxed to independence conditional on P. Note that independence can also be expressed based on the concept of copulae if the marginal distributions of ε and υ are normalized to $U\left(0, 1\right)$, as in Arellano & Bonhomme (n.d.):

$$C\left(\varepsilon, \upsilon | X, Z\right) = C\left(\varepsilon, \upsilon | P\right). \tag{6}$$

That is, the copula of ε and υ is independent of X and Z after conditioning on P. All parametric families in Arellano & Bonhomme (n.d.) satisfy this assumption. Therefore, our bounds can be considered as the worst case bounds if the copula was nonparametrically specified.

Vytlacil (2002) shows that the latent single index structure for S defined in equation (4) is equivalent to the following monotonicity assumption:

$$\begin{aligned} \text{For all } \left(x, z\right) \text{ and } \left(x', z'\right) &\in \mathcal{X} \times \mathcal{Z}, \\ \text{either } S_i\left(x, z\right) &\leq S_i\left(x', z'\right) \text{ for all } i \text{ or } S_i\left(x, z\right) \geq S_i\left(x', z'\right) \text{ for all } i, \end{aligned} \tag{7}$$

where $S_i\left(x, z\right)$ is the potential selection indicator that is observed when X_i and Z_i are exogenously set to x and z. This assumption is similar to the monotonicity assumption of Imbens & Angrist

(1994), however, with respect to the selection indicator (and not the treatment indicator). Lee (2009) bounds treatment effects in the presence of sample selection and invokes a monotonicity assumption similar to (7). The main difference is that he has no instrument such that Z is empty.

Lemma 2 *Expressions (4) and (5) and the monotonicity of $F_\upsilon(t)$ imply Assumption 4.*

Proof. *By the monotonicity of $F_\upsilon(t)$, $\{S=1\}=\{\upsilon \leq Q_\upsilon\}(P)$. Thus, $F_\varepsilon(\tau|S=1,X,Z,P) = F_\varepsilon(\tau|\{\upsilon \leq Q_\upsilon\}(P),X,Z,P) = F_\varepsilon(\tau|\{\upsilon \leq Q_\upsilon\}(P),P)$. As P serves as control function for υ, any bounded function of ϵ (including the cdf) given $S=1$ and P is independent of X,Z. This is Assumption 4.* ■

A natural assumption in quantile models is the rank invariance assumption (also called comonotonicity). This assumption appears already in the early motivations for considering quantile treatment effects, for instance in Lehmann (1974) and Doksum (1974). It has been considered more recently by Koenker & Xiao (2006) and Chernozhukov & Hansen (2005), among others. We define the potential outcomes as $Y(x) = m(x,\varepsilon_x)$ and state the rank invariance assumption as

$$\text{Conditional on } Z \text{ and } \upsilon,\ \{\varepsilon_x\} \text{ are identically distributed.} \tag{8}$$

Lemma 3 *Expressions (4) and (8) imply Assumption 3.*

Proof. To be written. ■

3 A new anatomy of the sample selection model

This section, the title of which alludes to Manski (1989), offers a new anatomy of the sample selection model. It first discusses the conditions under which the structural function is point identified. Then, it shows that the effects are identified for some sub-populations without further assumptions other than the ones of Section 2. Finally, the similarities and differences between the sample selection model and the instrumental variable model is discussed.

3.1 Conditions for point identification in the whole population

Since our model nests many traditional sample selection models, it allows us to discuss the conditions required for point identification. When the parameter of interest is the structural

function, two strategies point identify $m(x,\tau)$: identification at infinity, see Chamberlain (1986), and a parametric specification of the copula. When the interest is limited to the discrete change or the derivative of $m(x,\tau)$, the separability between x and ε offers a third way to identify these objects.

Given our assumptions, point identification of $m(x,\tau)$ over all $x \in \mathcal{X}$ based on identification at infinity does not require $P(x,z) = 1$ for some z across all values x. What has to be satisfied is that there exists a value of X, $x\prime$, and a value of Z, $z\prime$, such that $P(x\prime, z\prime) = 1$. This immediately yields $m(x\prime,\tau)$. Furthermore, it must hold for some $x\prime\prime$, $z\prime\prime$, and $z\prime\prime\prime$ that $P(x\prime, z\prime\prime) = P(x\prime\prime, z\prime\prime\prime)$. This allows identifying $m(x\prime\prime,\tau)$, as the distribution of ε given P is equal across different x in the selected population by Assumption 2. The same argument applies to any further values of the regressor such that any $m(x,\tau)$ is identified. For this reason, $F_{\varepsilon}(\tau|S=1,p)$ has to be point identified for only one observed value of $P(x,Z)$ and Z has to be sufficiently rich to satisfy common support in Z across different x. This is different to the bounds of Manski (1994) that collapse to a point only if $P(x,z) = 1$ for some z at each x.

Given these results, an obvious solution yielding point identification consists in parameterizing $F_{\varepsilon}(\tau|S=1,p)$. Under the single index crossing model, the identification of the copula of ε and v identifies $F_{\varepsilon}(\tau|S=1,p)$. Arellano & Bonhomme (n.d.) assume that this copula belongs to a parametric family. Note that this does not restrict the outcome heterogeneity. Previous model specifications, e.g., Heckman (1974) and Donald (1995), simultaneously parameterized the distribution of the errors in the outcome equation and the copula. This is unnecessary, severely restricts the model, and can be rejected by the data.

Both identification at infinity and a parametric specification of the copula yield point identification of the partial effect. In addition, the latter is also point identified when $m(X,\varepsilon)$ is separable and ε is globally independent of X and Z, which is the leading case considered in most semi- or non-parametric models. Assuming the following separable model

$$Y = m(x) + \varepsilon,\ (X,Z) \perp \varepsilon$$

implies that

$$\frac{\partial F_Y^{-1}(\theta|S=1,x,p)}{\partial x_1} = \frac{\partial m(x)}{\partial x_1},$$

which does not depend on θ. Therefore, the identified set shrinks to a point. Here, full independence is required while Assumption 1 in Section 2.1 only imposes a local quantile independence.

The separability assumption has stringent consequences, especially if one considers quantile effects. The latter are restricted to be location shift effects. Heteroscedasticity and higher order dependence are excluded even if they are ubiquitous in the empirical literature.

Note that slightly weaker assumptions are sufficient to point identify the parameters of interest. First, if one is interested in mean effects alone, only mean independence (instead of full independence) has to hold. Second, full independence can be eased somewhat by conditioning on the selection probability:

$$Y(x) = g(x) + \varepsilon,\ X \perp \varepsilon | P \tag{9}$$

This is a moderate relaxation that only allows for heteroscedasticity related to the conditional selection probability. Note that in the absence of sample selection (implying that P is a constant and equal to 1), full independence is assumed. Thus, even this weaker form is a strong restriction. Given its importance in the literature, Melly & Huber (2010) suggest a test for the conditional independence assumption in expression (9) for the linear quantile regression model.

Independence implies that the partial effects are constant across τ such that the lower bound coincides with the upper bound, yielding a single admissible value.[3] Thus, there is no cost for allowing for a violation of the independence in terms of identification. If the errors are indeed independent, the partial effects are point identified in the same manner as before. If they are not independent, we obtain a consistent identification region that covers the true parameter.

3.2 Point identification in some sub-populations

Since point identification is often not obtained under reasonable assumptions, an alternative strategy that has been prominently applied in the instrumental variable literature consists in moving the goalposts and identifying the parameters only for a subgroup instead of the full population. Even this approach requires the assumption of the single crossing selection equation (4). Newey (2007) shows that the latter and a strong support assumption for Z identifies the distribution of Y in the selected population. If the support condition is not satisfied, we can move the goalposts further and contend ourselves with identifying the effects for the selected population that satisfies the common support restriction.

The traditional semiparametric estimators handle the sample selection problem by conditioning on the propensity score, see for instance Powell (1987), Newey (1991), Cosslett (1991), and

[3] Therefore, all $m(X, \tau)$ are parallel across different $\tau \in (0, 1)$.

Buchinsky (1998). They consistently estimate the parameters in the whole population if the outcome function is separable as discussed in Section 3.1. However, if one allows for outcome heterogeneity by means of a nonseparable model, Newey (2007) results indicates that these methods estimate the mean and quantile effects consistently (only) in the selected population. This is in the spirit of the distinction between the local and global interpretation of IV estimators. The latter estimate the effects in the whole population if effect homogeneity is assumed and the effects for the so-called compliers (see Section 3.3) if heterogeneity is allowed for.

Note that the selected population is *not* the largest population for which we can identify the distribution of the outcome. Define $z^*(x)$ as the value of $z \in \mathcal{Z}$ that maximizes the observed participation probability at some $x \in \mathcal{X}$:

$$z^*(x) = \arg\max_{z \in \mathcal{Z}} \Pr(S = 1 | X = x, Z = z).$$

The potential outcome distribution for some hypothetical $X = x$, denoted as $Y(x)$, is identified for the population with $S(X, z^*(X)) = 1$.

It is obvious that the group with $S(X, z^*(X)) = 1$ is a super-population of the selected population. It encounters the latter plus those individuals who would have switched from non-selection to selection, had their instrument value been set to z^*. Note that if the common support assumption is not satisfied, we can derive a similar result for the population with $\min_{x \in \mathcal{X}} S(x, z^*(x)) = 1$, i.e., conditional on the subpopulation satisfying common support.

3.3 Links to the IV literature

This section discusses the analogies (and distinctions) of the sample selection framework considered in this paper and the instrumental variable (IV) treatment effects model. First of all, IV models for binary treatments may be written in terms of sample selection models, recall the equivalence result in Vytlacil (2002). The main difference between both frameworks is, however, that outcomes are uncensored in the IV model while they are not observed when $S = 0$ in the sample selection model. Therefore, the rank invariance of Chernozhukov & Hansen (2005) has no power at all in the latter case while it is powerful in the IV model because the rank of each individual is observed in one of the outcome distributions (under treatment and non-treatment).[4]

[4]Note that the rank invariance assumption discussed in Section 2.3 is different from the rank invariance assumption in Chernozhukov & Hansen (2005). Our rank invariance is with respect to X while their rank invariance is

The similarities are stronger when considering the LATE model of Imbens & Angrist (1994). In Section 3.2, we allow for outcome heterogeneity, restrict the first step heterogeneity,[5] and identify the function for a sub-population.[6] To see the analogy with the sample selection model, define $z_*(x)$ in the following way:

$$z_*(x) = \arg\min_{z \in \mathcal{Z}} \Pr(S = 1 | X = x, Z = z).$$

Using the potential selection notation we can now define three types of individuals (defiers are excluded by the monotonicity condition):

$$\begin{aligned} \text{never-selected:} \quad & S(x, z^*(x)) = 0, \\ \text{always-selected:} \quad & S_i(x, z_*(x)) = 1, \\ \text{compliers:} \quad & S_i(x, z_*(x)) < S_i(x, z^*(x)). \end{aligned}$$

The point is that we identify the distribution of Y for the compliers *and* for the always-selected, not just of the complying population, as in the IV model. (Recall, however, that compliance in the IV model is w.r.t. to the treatment state whereas it refers to selection in the sample selection model.) In an IV model, identification depends on variation in the treatment as a response to variation in an instrument in order to observe outcomes in both treatment states. Therefore, we cannot identify treatment effects on the always-takers because their potential outcomes under non-treatment are unobserved. In the sample selection model, only one potential outcome needs to be observed. Therefore, we easily identify the outcome distribution of the always-selected. This is the bright side of the sample selection model.

On the dark side, the goals of a treatment effect model and of a sample selection model are not necessarily identical. While we may be satisfied with estimating treatment effects in the complying population, we often aim at identifying the outcome distribution in the whole population when applying a sample selection correction. The identification of functions of populations that depend on the instrument and more generally on the propensity score is not satisfying. We illustrate this point by means of two examples. First, sample selection models have been used to correct for the selection bias that may arise when estimating the gender wage gap. It is typically assumed

with respect to D (S in the sample selection model).

[5] Remember that the single crossing condition is equivalent to their monotonicity condition.

[6] Machado (2009) makes a similar analogy using the monotonicity assumption for a binary Z. She identifies the effects for the always-selected individuals.

that there is no selection problem for men. The female wage equation is corrected and compared to the observed male wage distribution. The problem is that these two distributions will not be comparable because of the inherently different populations consisting of all males, but only the selected females. A second example is the comparison of wage functions across different time periods. Since the employment probability changed over time, the wage functions are estimated w.r.t. different populations such that the comparison is not meaningful.

There are obviously cases where the identification for some sub-population is interesting. For instance, when one is interested in the effect of a treatment and the outcome is only observed for a selected subpopulation. This is the case considered by Lee (2009), among others, who imposes the monotonicity assumption outlined in expression (7), but assumes no instrument for selection. In the absence of Z, point identification is not attained (except if the selection probability is the same for the treated and non-treated). Still, the effect for the population selected when treated and non-treated ($S(X=1) = S(X=0) = 1$) can be bounded. The advantage of considering this population is that the bounds for the average treatment effect don't depend on the bounds of the support of Y. Lechner & Melly (2010) bounds average and quantile treatment effects for the population selected when treated ($S_i(X=1) = 1$).

4 The parametric linear case

This section discusses estimation of and inference about the bounds derived in Section 3 under the parametric linear regression model. The first motivation for doing so is that, while the discussion of nonparametric identification is important to allow for a maximum of generality, the dimensionality of the problems often forces applied researchers to use a parametric model. Second, a large share of the literature still relies on the parametric model, see Mulligan & Rubinstein (2008) for a recent example. It therefore seems worthwhile to see what may be gained by our approach in these studies.

4.1 Identification

In addition to the identifying assumptions of Section 2.1, we impose linearity of the conditional quantile in the whole population.

Assumption 5 (linearity)

$$m(x,\tau) = x\beta(\tau). \tag{10}$$

Similar to Newey (1991), the common support assumption can be weakened to

$$E\left[S\left(X - E\left[X|P, S=1\right]\right)\left(X - E\left[X|P, S=1\right]\right)'\right] \text{ is nonsingular.}$$

To identify the structural function and the partial effects in the linear framework, we apply Theorem 1 to $\beta(\tau) = \frac{\partial m(x,\tau)}{\partial x}$ and obtain

$$\beta(\tau) \in \underset{p\in\mathcal{P}_{X=x,S=1}}{\cap} \left\{\beta(S=1,\theta,p) : \theta \in \Theta_p\right\},$$

where $\beta(S=1,\theta,p)$ is the θ^{th} quantile regression coefficient vector in the selected population with $P=p$. Θ_p denotes the identified set for $F_\varepsilon(\tau|S=1,p)$ and is defined in Theorem 1.

Note that the maximum and the minimum in the identified interval for $\beta(\tau)$ are not necessarily attained at the boundaries of $\Theta(p)$ because $\beta(\theta,p)$ is not necessarily monotonic in θ. Thus, all quantile regression processes in $\Theta(p)$ have to be computed to determine the upper and lower bound for $\beta(\tau)$. After identifying the admissible set $\Theta(p)$ for $\beta(\tau)$ at each value of P, we take the intersection across P to further narrow the bounds on $\beta(\tau)$, as the latter does not depend on P.

Adding to the discussion in Section 3.1, we briefly investigate under which conditions point identification is achieved for the special case of linear quantile regression processes. This implies that the lower bound on $\beta(\tau)$ coincides with the upper bound such that the identified interval collapses to a single point. This is the case if the regressors are independent of the error term conditional on P, see expression (9). In the linear model independence implies that the vector of slope coefficients is constant across quantiles such that the minimum is equal to the maximum. In contrast to the estimation of slope coefficients, the constant is not point identified even if independence is satisfied, but $p < 1$ for all observations.

If independence does not hold, point identification is still feasible when the data contains observations with selection probability $P = 1$. For the constant in linear models, point identification based on identification at infinity has been discussed in Heckman (1990) and Andrews & Schafgans (1998). In order to identify the whole vector $\beta(\tau)$, it is required that all regressors are linearly independent in the sub-population with $P = 1$. If some regressors are linearly dependent, point identification of the coefficients on these regressors is not obtained. The strategy suggested

in this section extends the approach by proposing set identification of the quantile coefficients when there is no population with $P = 1$. In contrast, the mean parameters cannot be bounded without further assumptions, e.g. a bounded support for the dependent variable.

Further assumptions can be introduced to tighten the bounds or simplify their definition. For instance, if it is assumed that the regressors affect only the first two moments of Y (heteroscedasticity), the upper and lower bound on $\beta(\tau)$ are attained at the boundaries of the interval for $\theta(p)$. Another restriction is the positive selection assumption that has been imposed for instance by Blundell et al. (2007). This assumption implies that the lower bound on Θ_p becomes τ.

4.2 Estimation

We suggest a nonparametric four-step estimator for the sharp identified set. In the first step, we estimate the identified set for $F_\varepsilon(\tau|P, S = 1)$, Θ_p. This requires the nonparametric estimation of $\Pr(S = 1|X, Z)$, $F_Y^{-1}(\theta|X, Z, S = 1)$ and $F_Y(q|X, P, S = 1)$. Ahn & Powell (1993) propose to estimate the conditional selection probability by local constant kernel estimation. We, however, prefer to use local logit in the application. This has similar advantages as local linear estimation under continuous outcomes has compared to local constant estimators, namely a better convergence rate at the boundaries.[7] We estimate the conditional quantile and distribution functions by local linear quantile regression. In the second step, for each observation with $\hat{p}_i > \max(\tau, 1-\tau)$, the linear quantile regression process of Y on X in a local neighborhood of $\hat{p}_i$ is estimated. This corresponds to the nonparametric quantile regression estimator proposed by Chaudhuri (1991) using, however, an infinite bandwidth with respect to X as linearity of Y in X is assumed. Only observations with similar $\hat{p}$ can be used in the same regression as the rank of the latent quantile regression in the selected population depends on p. The function $\hat{\beta}(\theta, \hat{p}_i)$ for $\theta \in (0, 1)$ is the result of this second step. In the third step, we calculate the identified set conditional on $\hat{p}_i$ by

$$\hat{\Psi}(p) \equiv \left\{ \hat{\beta}(\theta, p) : \theta \in \hat{\Theta}_p \right\}.$$

In the fourth and final step, the identified set for $\beta(\tau)$ is estimated as the intersection of the conditional bounds

$$\beta(\tau) \in \hat{\Psi} \equiv \bigcap_{\hat{p}_i > \max(\tau, 1-\tau)} \hat{\Psi}(\hat{p}_i).$$

[7] For Monte Carlo results on the finite sample properties of local logit and alternative semi- and nonparametric estimators, see Frölich (2006).

A simple way to report the identified set is to yield the largest and smallest elements for each parameter.

4.3 Inference

Note: work in progress!

We follow a strategy similar to the one suggested by Chernozhukov, Rigobon & Stoker (2009). We start with the simple problem of estimating a $1-a$ confidence interval for the quantile regression parameter $\hat{\beta}_k(\theta, p)$ when θ and p are pre-determined. This confidence interval can be obtained in a straightforward way either analytically or by resampling and is denoted by $C_{1-\alpha}(\theta, p)$. In the next step, p is still fixed, but not θ. By proposition 2 of Chernozhukov et al. (2009) we obtain the following confidence intervals for $\hat{\Psi}_k(p)$

$$CR_{1-a}(p) \equiv \cup_{\theta \in \left[\frac{\tau-(1-p)}{p}, \frac{\tau}{p}\right]} C_{1-a}(\theta, p).$$

For any p, $CR_{1-a}(p)$ is a valid, not necessarily conservative, confidence interval for $\beta_k(\tau)$.

5 Labor Market Application

To be done!

6 Conclusion

It may seem disappointing that in a sample selection model for quantiles, invoking a quite comprehensive set of conditions (exclusion restriction, monotonicity of selection in observables, rank invariance in the conditional outcome distribution) does generally not identify the parameters of interest even in the linear regression framework. This, however, should not be too surprising when bearing in mind that the outcome is never observed for a subpopulation if there is no identification at infinity. Thus, we either have to severely restrict the generality of the model to obtain point identification or content ourselves with partial identification.

The approach suggested in this paper aims at combining the best of both worlds. If the data support even stronger conditions than those mentioned before (such as separability of observables and unobservables), which are maintained in classical sample selection models, the parameters

are point identified. If these assumptions are rejected, the suggested method yields the tightest bounds on the parameters of interest under somewhat weaker conditions.

References

Ahn, H. & Powell, J. (1993), 'Semiparametric estimation of censored selection models with a nonparametric selection mechanism', *Journal of Econometrics* **58**, 3–29.

Andrews, D. & Schafgans, M. (1998), 'Semiparametric estimation of the intercept of a sample selection model', *Review of Economic Studies* **65**, 497–517.

Arellano, M. & Bonhomme, S. (n.d.), Quantile selection models. Presented at Yale in June 2010.

Blundell, R., Gosling, A., Ichimura, H. & Meghir, C. (2007), 'Changes in the distribution of male and female wages accounting for employment composition using bounds', *Econometrica* **75**(2), 323Ű363.

Buchinsky, M. (1998), 'The dynamics of changes in the female wage distribution in the usa: A quantile regression approach', *Journal of Applied Econometrics* **13**, 1–30.

Chamberlain, G. (1986), 'Asymptotic efficiency in semiparametric models with censoring', *Journal of Econometrics* **32**, 189–218.

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

Chernozhukov, V. & Hansen, C. (2005), 'An iv model of quantile treatment effects', *Econometrica* **73**, 245–261.

Chernozhukov, V., Rigobon, R. & Stoker, T. M. (2009), 'Set identification with tobin regressors', *unpublished manuscript* .

Chesher, A. (2003), 'Identification in nonseparable models', *Econometrica* **71**, 1405–1441.

Cosslett, S. (1991), Distribution-free estimator of a regression model with sample selectivity, *in* W. Barnett, J. Powell & G. Tauchen, eds, 'Nonparametric and semiparametric methods in econometrics and statistics', Cambridge University Press, Camdridge, UK, pp. 175–198.

Das, M., Newey, W. & Vella, F. (2003), 'Nonparametric estimation of sample selection models', *Review of Economic Studies* **70**, 33–58.

Doksum, K. (1974), 'Empirical probability plots and statistical inference for nonlinear models in the two-sample case', *The Annals of Statistics* **2**, 267–277.

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

Frölich, M. (2006), 'Non-parametric regression for binary dependent variables', *Econometrics Journal* **9**, 511Ű540.

Gronau, R. (1974), 'Wage comparisons-a selectivity bias', *Journal of Political Economy* **82**(6), 1119–1143.

Heckman, J. (1974), 'Shadow prices, market wages and labor supply', *Econometrica* **42**, 679–694.

Heckman, J. (1990), 'Varieties of selection bias', *American Economic Review, Papers and Proceedings* **80**, 313–318.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**(1), 153–161.

Imbens, G. & Angrist, J. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**, 467–475.

Kitagawa, T. (2009), Testing for instrument independence in the selection model. Mimeo, UCL.

Koenker, R. & Xiao, Z. (2006), 'Quantile autoregression', *Journal of the American Statistical Association* **101**, 980–990.

Lechner, M. & Melly, B. (2010), Partial identification of wage effects of training programs. Mimeo, Brown University.

Lee, D. S. (2009), 'Training, wages, and sample selection: estimating sharp bounds on treatment effects', *Review of Economic Strudies* **76**, 1071–1102.

Lehmann, E. (1974), *Nonparametrics: Statistical Methods based on Ranks*, Holden-Day.

Machado, C. (2009), Selection, heterogeneity and the gender wage gap. Mimeo, University of Columbia.

Manski, C. F. (1989), 'Anatomy of the selection problem', *The Journal of Human Resources* **24**(3), 343–360.

Manski, C. F. (1994), The selection problem, *in* C. Sims., ed., 'Advances in Econometrics: Sixth World Congress', Cambridge University Press, pp. 143–170.

Manski, C. F. (2003), *Partial Identification of Probability Distributions*, New York: Springer Verlag.

Melly, B. & Huber, M. (2010), Quantile regression in the presence of sample selection. Working Paper, University of St. Gallen.

Mulligan, C. B. & Rubinstein, Y. (2008), 'Selection, investment, and women's relative wages since 1975', *Quarterly Journal of Economics* **123**, 1061–1110.

Newey, W. K. (1991), 'Two-step series estimation of sample selection models', *unpublished manuscript, M.I.T.* .

Newey, W. K. (2007), 'Nonparametric continuous/discrete choice models', *International Economic Review* **48**(4), 1429–1439.

Powell, J. (1987), 'Semiparametric estimation of bivariate latent variable models'. unpublished manuscript, University of Wisconsin-Madison.

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**, 331–341.