INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2010-10

Another Look at the Identification at Infinity of Sample Selection Models*

X. D'HAULTFOEUILLE[1]
A. MAUREL[2]

* We are grateful to Magali Beffy, Edwin Leuven and Arthur Lewbel for helpful comments.
[1] CREST-INSEE. Email : xavier.dhaultfoeuille@ensae.fr
[2] CREST-INSEE, ENSAE, PSE (Paris) et IZA (Bonn). Email : maurel@ensae.fr

Abstract

It is often believed that without instrument, endogenous sample selection models are identified only if a covariate with a large support is available (see Chamberlain, 1986, and Lewbel, 2007). We propose a new identification strategy mainly based on the condition that the selection variable becomes independent of the covariates when the outcome, not one of the covariates, tends to infinity. No large support on the covariates is required. Moreover, we prove that this condition is testable. We finally show that our strategy can be applied to the identification of generalized Roy models.

Keywords: Identification at infinity, sample selection model, Roy model
JEL classification: C21

Résumé

Il est souvent admis qu'en l'absence d'instrument, les modèles de sélection généralisée ne sont identifiés que sous une condition de large support d'une covariable (cf. Chamberlain, 1986, et Lewbel, 2007). Nous proposons une nouvelle stratégie d'identification basée principalement sur une hypothèse d'indépendance entre la variable de sélection et les covariables lorsque la variable dépendante, et non l'une des covariables, tend vers l'infini. Dans ce cas, aucune condition de large support sur les covariables n'est requise. Nous montrons également que cette condition est testable. Enfin, nous appliquons cette stratégie à l'identification des modèles de Roy généralisés.

Mots-clés : identification à l'infini, modèle de sélection généralisée, modèle de Roy.
Classification JEL : C21.

1 Introduction

Since the seminal work of Heckman (1974), the issue of endogenous selection has been an active topic of research in both applied and theoretical econometrics (see Vella, 1998, for a survey). The usual strategy to deal with this issue is to rely on instruments that determine selection but not the outcome. However, the search of a valid instrument may be difficult if not impossible in some applications. Another strategy, which has been sometimes advocated, relies on the fact that, loosely speaking, the selection problem becomes negligible "at the limit". Following this idea, Chamberlain (1986) proved that the effects of covariates on an outcome are identified under the linearity of the model and a large support assumption on at least one covariate. Lewbel (2007) generalized this result by proving that identification can be achieved without imposing any structure on the outcome equation, provided that a special regressor has a large support and under restrictions on the selection equation.[1]

The main drawback of the latter approach is that it requires the existence of a covariate with a large support. Thus, it breaks down when all covariates are discrete, a case which is fairly common in practice. In this paper, we consider another route for identifying the model at infinity. Intuitively, if selection is truly endogenous, then we can expect the effect of the outcome on the selection variable to dominate those of the covariates for large values of the outcome. Following this idea, our main identifying condition states that the selection variable is independent of the covariates at the limit, i.e., when the outcome tends to infinity. Under this condition, the model is identified without any large support condition on these covariates. Only an exogeneity assumption and a mild restriction on the residuals are required. Moreover, we show that the main condition is testable. Apart from the standard selection model, we apply our result to a generalization of the Roy model (1951) of self-selection accounting for non-pecuniary factors. We show that, in this framework, the effects of covariates on the outcomes are identified without exclusion restriction under a moderate dependence condition on the residuals.

The note is organized as follows. Section 2 presents the model and establishes the main identification result. Section 3 proves the testability of our main condition. Section 4 applies this result to generalized Roy models, and Section 5 concludes.

[1] These restrictions entail that the probability of selection tends to zero or one when the special regressor takes arbitrary large values.

2 Main result

Let Y^* denote the outcome of interest, X denote covariates and D denote the selection dummy. Let us consider the following model, with $\sigma(X) > 0$:

$$Y^* = \psi(X) + \sigma(X)\varepsilon \tag{2.1}$$

The econometrician observes D, $Y = DY^*$ and X. Without loss of generality, we suppose that $\psi(x_0) = 0$ and $\sigma(x_0) = 1$ for a given $x_0 \in \text{Supp}(X)$, where $\text{Supp}(T)$ denotes the support of the random variable T. Our main result is based on the following assumptions.

Assumption 1 *(Exogeneity)* $X \perp\!\!\!\perp \varepsilon$.

Assumption 2 *(Tails of the residual)* $\sup(Supp(\varepsilon)) = +\infty$. *Moreover, there exists* $\beta > 0$ *such that* $E(\exp(\beta\varepsilon)) < \infty$.

Assumption 3 *(Independence at infinity) There exists* $l > 0$ *such that for all* $x \in Supp(X)$, $\lim_{y\to\infty} P(D = 1|X = x, Y^* = y) = l$.

Assumption 1 is usual in selection models and weaker than the exogeneity assumption imposed by Chamberlain (1986), since heteroskedasticity is allowed for here. Assumption 2 puts some weak restrictions on the tails of the distribution of ε. In the example of a wage equation where Y^* denotes the logarithm of the wage W, it is satisfied if $E[W^\beta] < \infty$ for a given $\beta > 0$. Thus, it holds even if wages have very fat tails, Pareto-like for instance. Note that Assumption 2 also implies that the supremum of the support of ε is infinite. We discuss below the implications of relaxing this condition. Finally, Assumption 3 is the main condition here. It requires the probability of selection to be independent of X at the limit, i.e., for those who have very large outcomes. In other terms, the effect of Y^* on selection becomes prominent when Y^* takes arbitrary large values.[2] To illustrate Assumption 3, let us consider the following selection rule:

$$D = \mathbb{1}\{\varphi(X) + \eta \geq 0\}. \tag{2.2}$$

Endogenous selection stems from the correlation between η and ε. Suppose that the following decomposition holds:

$$\eta = h(\varepsilon) + \nu, \ \nu \perp\!\!\!\perp (\varepsilon, X).$$

[2]Exogenous selection is thus ruled out (except when $(x, y) \mapsto P(D = 1|X = x, Y^* = y)$ is constant). As shown in Section 3, it is actually possible in this case to reject Assumption 3 from the data.

Then we get:

$$D = \mathbb{1}\left\{\varphi(X) + h\left(\frac{Y^* - \psi(X)}{\sigma(X)}\right) + \nu \geq 0\right\}.$$

Thus, Assumption 3 is satisfied (with $l = 1$) provided that $h(x) \to \infty$ as $x \to \infty$. In particular, when $h(x) = ax$, this condition holds provided that $a > 0$. Hence, in the Gaussian case, Assumption 3 is satisfied as soon as $Cov(\eta, \varepsilon) > 0$. Note that neither additive separability nor monotonicity in η of the index in (2.2) is needed. If $D = \mathbb{1}\{\varphi(X, \eta) \geq 0\}$, the same reasoning applies provided that for all x, $\liminf_{u\to\infty} \varphi(x, u) > 0$.

In the examples above, $l = 1$ but Assumption 3 also holds with $0 < l < 1$. This is the case (under the preceding assumptions on h and φ) if $D = U\mathbb{1}\{\varphi(X, \eta) \geq 0\}$, where $U \in \{0, 1\}$ is a random shock independent of (X, ε, η) satisfying $P(U = 1) > 0$. For instance, this framework may be used to model participation to the labor market, with U denoting in that case an unobserved random shock related to, e.g., health conditions that could prevent individuals from entering the labor market. One could also rely on this framework to model women's labor supply in the presence of family disruptions (such as separation or divorce, which is denoted here by U) inducing them to participate to the labor market.

Theorem 2.1 *Under Assumptions 1-3, $\psi(.)$ and $\sigma(.)$ are identified.*

Proof: Subsequently, S_T denotes the survival function of the random variable T. Besides, we use the notation $f(y) \sim g(y)$ if there exists $r(.)$ such that $f(y) = g(y)(1 + r(y))$ with $\lim_{y\to\infty} r(y) = 0$. The result is based on the following lemma.

Lemma 2.1 *Let T be a real random variable such that $\sup(Supp(T)) = +\infty$ and $E(|T|) < \infty$. Suppose also that $S_T(y) \sim S_T(lf(y))$, where $\lim_{y\to\infty} f'(y) = 1$ and $l > 0$. Then $l = 1$.*

Proof of Lemma 2.1: Suppose that $l > 1$. Then there exists $\eta > 0$ such that $l > 1 + \eta$. Moreover, because $\sup(\text{Supp}(T)) = +\infty$, $S_T(lf(y)) > 0$ for all y. Thus, $S_T(y) \sim S_T(lf(y))$ implies that there exists y_0 such that for all $y \geq y_0$,

$$S_T(y) < (1 + \eta)S_T(lf(y)).$$

Note that $E(|T|) < \infty$ implies that $\int_0^\infty S_T(u)du < \infty$. Consequently, for all $y \geq y_0$,

$$\int_y^\infty S_T(u)du < (1 + \eta)\int_y^\infty S_T(lf(u))du. \tag{2.3}$$

By assumption, the derivative of the function $m(y) = lf(y)$ tends to $l > 1$ when $y \to \infty$. Thus, there exists y_1 such that for all $y \geq y_1$, $m'(y) > 1 + \eta$. Integrating between y_1 and

$y \geq y_1$ shows that $m(y) > (1+\eta)(y-y_1) + m(y_1)$. Thus, there exists $y_2 \geq y_1$ such that $m(y) > y$ for all $y \geq y_2$. Hence, for all $y \geq y_2$, m is one-to-one and

$$\begin{aligned}
\int_y^{\infty} S_T(lf(u))du &= \int_{m(y)}^{\infty} \frac{S_T(v)}{m'(m^{-1}(v))} dv \\
&< \frac{1}{1+\eta} \int_{m(y)}^{\infty} S_T(v)dv \\
&< \frac{1}{1+\eta} \int_y^{\infty} S_T(v)dv. \qquad (2.4)
\end{aligned}$$

Inequalities (2.3) and (2.4) imply that $\int_y^{\infty} S_T(u)du < \int_y^{\infty} S_T(u)du$ for all $y \geq \max(y_0, y_2)$, a contradiction. Similarly, one can show that $l < 1$ is impossible. Thus $l = 1$. $\square$

Now let us prove Theorem 2.1. Let $q(y, x) = P(D = 1, Y^* \geq y | X = x)$. We have

$$q(y, x) = \int_y^{\infty} P(D = 1 | X = x, Y^* = u) dP^{Y^*|X=x}(u)$$

By Assumption 3, as $u \to \infty$, we have $P(D = 1|X = x, Y^* = u) \to l > 0$. Thus, using standard results on integrals, we get as $y \to \infty$,

$$q(y, x) \sim lP(Y^* \geq y | X = x).$$

By Assumption 1, $P(Y^* \geq y|X = x) = S_\varepsilon((y - \psi(x))/\sigma(x))$, where $S_\varepsilon(.)$ denotes the survival function of ε . Thus,

$$q(y, x) \sim lS_\varepsilon\left(\frac{y - \psi(x)}{\sigma(x)}\right). \qquad (2.5)$$

Similarly,

$$q(y, x_0) \sim lS_\varepsilon(y) \qquad (2.6)$$

In other words,

$$q(y, x) \sim q\left(\frac{y - \psi(x)}{\sigma(x)}, x_0\right) \qquad (2.7)$$

Note that the function q is identified. Thus, $\sigma(x)$ and $\psi(x)$ are identified if, as $y \to \infty$,

$$q(y, x) \sim q(sy + u, x_0)\,(s > 0) \implies (s, u) = \left(\frac{1}{\sigma(x)}, -\frac{\psi(x)}{\sigma(x)}\right) \qquad (2.8)$$

To prove (2.8), suppose that $s > 0$ and u satisfy $q(y, x) \sim q(sy + u, x_0)$. Then it follows from (2.5) and (2.6) that

$$S_\varepsilon(t(y + v)) \sim S_\varepsilon(y), \qquad (2.9)$$

where $t = s\sigma(x)$ and $v = (1/\sigma(x))(\psi(x)+u/s)$. Besides, by Assumption 2, $\sup(\text{Supp}(\varepsilon)) = +\infty$ and $E(|\varepsilon|) < \infty$. Thus, by Lemma 2.1, $t = 1$, i.e. $s = 1/\sigma(x)$. Thus, $\sigma(x)$ is identified. Besides, by (2.9),

$$S_{e^{\beta\varepsilon}}(wy) \sim S_{e^{\beta\varepsilon}}(y),$$

where β is defined in Assumption 2 and $w = \exp(\beta v)$. Because $E(\exp(\beta\varepsilon)) < \infty$, we can apply Lemma 2.1 once more. This yields $w = 1$, which is equivalent to $u = -\psi(x)/\sigma(x)$. Thus, $\psi(x)$ is identified. $\square$

The intuition of the proof is that by Assumption 3, the conditional survival function of Y is equivalent at infinity (up to a constant) to the one of a location-scale model. Then the normalization $(\psi(x_0), \sigma(x_0)) = (0, 1)$ and the restrictions on ε ensure that the parameters of this location-scale model can be identified at infinity. Note that, unlike Lewbel (2007), we rely on additive separability in the outcome equation. On the other hand, no structure is imposed on the selection process, apart from Assumption 3.

Because of its argument at infinity, the proof of the theorem breaks down when the supremum of the support of the residual is finite. However, in this situation, identification can still be achieved if the infimum of the support is also finite, or by assuming homoskedasticity. In both cases, indeed, the functions of interest can be recovered by using support variation and the normalizations $(\psi(x_0), \sigma(x_0)) = (0, 1)$. Within this framework, Assumption 3 is not required anymore.

Note also that Theorem 2.1 does not provide any information on the intercept of (2.1), that is, on $E(\varepsilon)$. Actually, one can show that this intercept is not identified in general in our context. Basically, this stems from the fact that contrary to the framework of Andrews and Schafgans (1998), for instance, there is in general here no individual for whom $P(D = 1|X)$ is arbitrarily close to one.[3] Besides, apart from Assumption 3, our model puts no restriction on the probability $P(D = 1|X, Y^*)$. As a result, it is possible to define a distribution for ε and a conditional probability of selection different from the true ones but observationally equivalent, leading to different values for $E(\varepsilon)$.[4]

[3] It can be shown that if $P(D = 1|X = x) \to 1$ when $x \to x_1$ (x_1 being finite or infinite), $E(\varepsilon)$ is identified by $\lim_{x\to x_1}(E(Y|D = 1, X = x) - \psi(x))/\sigma(x)$.

[4] The formal proof of this non-identification result is available from the authors upon request.

3 Testability

The main identifying condition in the setting above is Assumption 3, so one may wonder whether this assumption is refutable or not. The answer turns out to be affirmative. Indeed, together with Assumptions 1 and 2, this condition implies (2.7), which can be stated as

$$\forall x \in \text{Supp}(X),\ \exists (s(x), u(x)) \in \mathbb{R}^{*+} \times \mathbb{R} :\ q(y, x) \sim q(s(x)y + u(x), x_0), \tag{3.1}$$

where $q(y, x) = P(D = 1, Y^* \geq y | X = x)$. Because the function q is identified, Condition (3.1) can be tested in the data. Then one can reject Assumption 3 when there is no $(s(x), u(x))$ satisfying (3.1). Theorem 3.1 below shows that the reverse also holds: under a slight reinforcement of Assumption 2 and another mild condition, Condition (3.1) and Assumption 3 are equivalent. Hence, there is no risk of misspecification here: one can reject Assumption 3 whenever it fails to hold.

Theorem 3.1 *Suppose that Assumption 1 holds,* $\sup(Supp(\varepsilon)) = +\infty$*, there exists* $\alpha > 1$*,* $\beta > 0$ *such that* $E[\exp(\beta|\varepsilon|^\alpha)] < \infty$ *and there exists* $l(x) > 0$ *such that*

$$\lim_{y \to \infty} P(D = 1 | X = x, Y^* = y) = l(x). \tag{3.2}$$

Then Assumption 3 is equivalent to Condition (3.1).

Proof: We shall first prove a result similar to the one of Lemma 2.1.

Lemma 3.1 *Let* T *be a real random variable such that* $\sup(Supp(T)) = +\infty$ *and* $E(|T|) < \infty$*. Suppose also that when* $y \to \infty$*,* $S_T(y) \sim l S_T(f_\delta(y))$*, where* $l > 0$ *and* $f_\delta(.)$ *is strictly increasing for* y *large enough and satisfies (i)* $f'_\delta(y) \to 0$ *if* $\delta < 0$*, (ii)* $f'_0(y) \to C > 0$ *and (iii)* $f'_\delta(y) \to \infty$ *if* $\delta > 0$*. Then* $\delta = 0$*. Moreover, if* $f_0(y) = y$*, then* $l = 1$*.*

Proof of Lemma 3.1: Suppose that $\delta > 0$. By assumption, there exists $l' > 0$ and y_0 such that for all $y \geq y_0$,

$$S_T(y) < l' S_T(f_\delta(y)). \tag{3.3}$$

Besides, there exists y_1 such that $f_\delta(.)$ is one-to-one on $[y_1, \infty)$, with $f'_\delta(y) > l'$ and $f_\delta(y) > y$ for all $y \geq y_1$. Thus, for all $y \geq y_1$,

$$\begin{aligned} \int_y^\infty S_T(f_\delta(u))du &= \int_{f_\delta(y)}^\infty \frac{S_T(v)}{f'_\delta(f_\delta^{-1}(v))}dv \\ &< \frac{1}{l'}\int_{f_\delta(y)}^\infty S_T(v)dv \\ &< \frac{1}{l'}\int_y^\infty S_T(v)dv. \end{aligned} \tag{3.4}$$

Inequalities (3.3) and (3.4) imply that $\int_y^\infty S_T(u)du < \int_y^\infty S_T(u)du$ for all $y \geq \max(y_0, y_1)$, a contradiction. The proof that $\delta < 0$ is impossible follows similarly. Thus $\delta = 0$. Finally, if $f_0(y) = y$, then $S_T(y) \sim lS_T(y)$, which implies directly that $l = 1$. $\square$

Now let us prove Theorem 3.1. By the proof of Theorem 2.1, Assumption 3 implies Condition (3.1). Thus, it suffices to prove that Condition (3.1) implies Assumption 3. For all $x \in \text{Supp}(X)$, by a similar reasoning as in the previous proof,

$$q(y, x) \sim l(x)S_\varepsilon\left(\frac{y - \psi(x)}{\sigma_(x)}\right).$$

The same holds for $q(y, x_0)$. Thus, by Condition (3.1), there exists $\mu > 0$ and $\nu \in \mathbb{R}$ such that

$$S_\varepsilon(y) \sim lS_\varepsilon(\mu y + \nu), \tag{3.5}$$

where $l = l(x)/l(x_0)$. This implies that

$$S_{\exp(\beta\varepsilon)}(y) \sim lS_{\exp(\beta\varepsilon)}(\exp(\beta\nu)y^\mu).$$

By assumption, $E[\exp(\beta\varepsilon)] < \infty$. Thus, by applying Lemma 3.1 to $f_\delta(y) = \exp(\beta\nu)y^{\exp(\delta)}$ (with $\delta = \ln\mu$), we get $\mu = 1$. Hence, by (3.5),

$$S_{\exp(\beta\varepsilon^\alpha)}(\exp(\beta y^\alpha)) \sim lS_{\exp(\beta\varepsilon^\alpha)}(\exp(\beta(y + \nu)^\alpha)).$$

After some manipulations, we obtain

$$S_{\exp(\beta\varepsilon^\alpha)}(y) \sim lS_{\exp(\beta\varepsilon^\alpha)}(f_\nu(y)),$$

where

$$f_\nu(y) = y^{\left(1+\nu\left(\frac{\beta}{\ln y}\right)^{1/\alpha}\right)^\alpha}.$$

Some computations show that f_ν is strictly increasing for y large enough and (i) $f_\nu'(y) \to 0$ if $\nu < 0$, (ii) $f_0(y) = y$ and (iii) $f_\nu'(y) \to \infty$ if $\nu > 0$. Thus, by Lemma 3.1, $\nu = 0$ and $l = 1$. In other terms, $l(x) = l(x_0)$ for all $x \in \text{Supp}(X)$, which proves that Assumption 3 holds. $\square$

To illustrate Theorem 3.1, suppose for instance that in the true model, selection is exogenous, i.e. $P(D = 1|X = x, Y^* = y) = P(D = 1|X = x)$ for all y, and that $x \mapsto P(D = 1|X = x)$ is nonconstant. In this setting, Condition (3.2) is satisfied with $l(x) = P(D = 1|X = x)$. Thus, by Theorem 3.1, Condition (3.1) fails to hold, since Assumption 3 is not satisfied. Hence, the "independence at infinity" assumption can be rejected by the data in this case.

4 Application to generalized Roy Models

Let us consider a class of generalized Roy models where each individual chooses the sector $D \in \{0, 1\}$ that provides him with the higher utility. Suppose that the utility U_i associated with each sector $i \in \{0, 1\}$ is the sum of the potential log-earnings $Y_i = \psi_i(X) + \varepsilon_i$ and a random nonpecuniary component $G_i(X) + \eta_i$. Thus, $D = \mathbb{1}\{Y_1 \geq Y_0 + G(X) + \eta\}$ with $G(X) = G_0(X) - G_1(X)$ and $\eta = \eta_0 - \eta_1$, and the econometrician only observes $Y = DY_1 + (1 - D)Y_0$, as well as D and X. For the sake of simplicity, we do not account for uncertainty on potential outcomes. Nevertheless, it would be straightforward to adapt our identification strategy to the case where sectoral decisions depend on expectations of Y_0 and Y_1 rather than on their true values (see D'Haultfœuille and Maurel, 2009). Without loss of generality, we assume that there exists $x_0 \in \text{Supp}(X)$ such that $\psi_0(x_0) = \psi_1(x_0) = 0$.

The generalized Roy models we consider in this section can be used in a broad range of economic settings. The standard Roy model, in which the chosen sector is the one yielding the higher earnings, corresponds to $\eta = 0$ and $G(X) = 0$. This framework also encompasses Heckman (1974)'s model of labor market participation. In this latter case, Y_1 corresponds to the logarithm of the potential wage, $G_1(X) = \eta_1 = 0$, $Y_0 = 0$ and $G_0(X)$ (resp. η_0) is the observable (resp. unobservable) part of the logarithm of the reservation wage. More generally, generalized Roy models are well suited for most of the situations in which self-selection between two alternatives is driven both by the relative pecuniary and non-pecuniary returns. They can be used for instance to model the decision to attend higher education after graduating from high school, thus extending Willis and Rosen (1979) by accounting for non-pecuniary factors entering the schooling decision (see, e.g., Carneiro et al., 2003). Other examples of applications include occupational choice (see, e.g., Dagsvik and Strøm, 2006 for the choice between private and public sector) as well as migration decisions (see, e.g., Borjas, 1987 and Bayer et al., 2008) accounting for non-pecuniary factors.[5] Theorem 2.1 can be applied to provide identification of (ψ_0, ψ_1) without an exclusion restriction nor any large support condition on the covariates, as the following result shows.

Corollary 4.1 *Suppose that $(\varepsilon_0, \varepsilon_1, \eta) \perp\!\!\!\perp X$, the suprema of the supports of ε_0 and ε_1 are*

[5]Note that generalized Roy models are also used as a structural underlying framework for the estimation of treatment effects, with D corresponding in that case to the treatment status and $G + \eta$ to the cost of receiving treatment (see Heckman and Vytlacil, 2005). Here however, we cannot recover average treatment effects in general since we do not identify $E(\varepsilon_0)$ and $E(\varepsilon_1)$. Yet, the distribution of treatment effects can be point or set identify under additional restrictions (see D'Haultfœuille and Maurel, 2009).

infinite and there exists $\beta_0, \beta_1 > 0$ *such that* $E[\exp(\beta_i \varepsilon_i)] < \infty$ *for* $i \in \{0,1\}$ *and*

$$\lim_{u \to \infty} P(\varepsilon_i + (1-2i)\eta \leq a + u | \varepsilon_{1-i} = u) = l_{1-i} > 0 \tag{4.1}$$

for all $a \in \mathbb{R}$ *and* $i \in \{0,1\}$. *Then* $\psi_0(.)$ *and* $\psi_1(.)$ *are identified.*

Proof: Since $(\varepsilon_0, \varepsilon_1, \eta) \perp\!\!\!\perp X$, Condition (4.1) implies that

$$\lim_{u \to \infty} P(Y_1 \geq Y_0 + G(X) + \eta | X = x, Y_1 = u) = l_1.$$

In other words,

$$\lim_{u \to \infty} P(D = 1 | X = x, Y_1 = u) = l_1.$$

Thus, we can apply Theorem 2.1 to (D, DY_1, X) and ψ_1 is identified. The same result holds for ψ_0. □

To the best of our knowledge, this is the first identification result on the effects of covariates in generalized Roy models without exclusion restriction. Identification without exclusion restriction of the competing risk model, which is strongly related to the standard Roy model, has already been considered in the literature by Heckman and Honore (1989),[6] Abbring and van den Berg (2003), Lee (2006) and Lee and Lewbel (2009).[7] However, all of the strategies proposed in these papers break down when turning to generalized Roy models. Indeed, they rely extensively on the fact that the observed duration is the minimum of potential durations, whereas the observed outcome does not satisfy such a simple property in generalized Roy models.

Identification of (ψ_0, ψ_1) is obtained in Corollary 4.1 under rather mild restrictions on the unobservables. In particular, Condition (4.1) can be understood as a moderate dependence assumption between the unobservables. It is automatically satisfied for instance if ε_0, ε_1 and η are independent. It also holds if $(\varepsilon_0, \varepsilon_1, \eta)$ is Gaussian, provided that

$$|Cov(\varepsilon_i, \varepsilon_{1-i} + (2i-1)\eta)| < V(\varepsilon_i), \quad i \in \{0,1\}.$$

Noteworthy, the latter condition does not put drastic restrictions on the dependence between the unobservables. For instance, it will be satisfied in the standard log-normal Roy model if $V(\varepsilon_0) = V(\varepsilon_1)$, as long as $(\varepsilon_0, \varepsilon_1)$ is not degenerated. It is also satisfied for instance in Heckman (1974)'s empirical application to labor market participation of married women, although the estimated correlation between ε and η is quite large (0.83).

[6] Heckman and Honore (1989) use exclusion restrictions but only to identify the distribution of the underlying durations. Their proof shows that the effects of covariates are identified without such restrictions.

[7] Interestingly, these last two papers do not rely on identification at the limit.

Condition (4.1) is appealing because of its fairly simple interpretation in terms of dependence between the unobservables. Nevertheless, ψ_0 and ψ_1 may be identified even if it fails, as soon as the "independence at infinity" conditions hold in this context, namely as soon as for all $x \in \text{Supp}(X)$, $\lim_{u\to\infty} P(D = 0|X = x, Y_0 = u) = l_0 > 0$ and $\lim_{u\to\infty} P(D = 1|X = x, Y_1 = u) = l_1 > 0$. Furthermore, Section 3 shows that one can in any case test for these conditions by checking whether Condition (3.1) holds or not.

5 Concluding remarks

This note shows that identification of endogenous sample selection models can be achieved without instrument by letting the outcome, not a covariate, tend to infinity. The main condition, apart from the exogeneity of the covariates, is the "independence at infinity" of the selection variable and the covariates. In particular, unlike Chamberlain (1986) and Lewbel (2007), our identification strategy does not rely on the existence of a covariate with a large support. Besides, another attractive feature of the proposed identification strategy lies in its testability. Noteworthy, our identification proof is constructive, and an estimator of $\psi(.)$ and $\sigma(.)$ could be based on (2.8) for instance. One possible route for estimation would be to use trimmed means, in a similar spirit as Andrews and Schafgans (1998). In this case, we conjecture that the rate of convergence would depend on the thickness of the tail of the distribution of the outcome, as in Andrews and Schafgans (1998) and Khan and Tamer (2009). The derivation of the estimators and their properties seems quite intricate, however, and we leave this issue for future research.

References

Abbring, J. and van den Berg, G. (2003), 'The identifiability of the mixed proportional hzards competing risks model', *J.R. Statist. Soc. B* **65**, 701–710.

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

Bayer, P. J., Khan, S. and Timmins, C. (2008), Nonparametric identification and estimation in a generalized roy model. NBER Working Paper.

Borjas, G. (1987), 'Self-selection and the earnings of immigrants', *American Economic Review* **77**, 531–553.

Carneiro, P., Hansen, K. and Heckman, J. (2003), 'Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice', *International Economic Review* **44**, 361–422.

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

Dagsvik, J. and Strøm, S. (2006), 'Sectoral labour supply, choice restrictions and functional form', *Journal of Applied Econometrics* **21**, 803–826.

D'Haultfœuille, X. and Maurel, A. (2009), Inference on a generalized Roy model, with an application to schooling decisions in france. IZA Discussion Paper No 4606.

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

Heckman, J. J. and Honore, B. (1989), 'The identifiability of competing risks models', *Biometrika* **76**, 325–330.

Heckman, J. and Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation', *Econometrica* **73**, 669–738.

Khan, S. and Tamer, E. (2009), Irregular identification, support conditions and inverse weight estimation. Working Paper, Northwestern University.

Lee, S. (2006), 'Identification of a competing risks model with unknown transformations of latent failure times', *Biometrika* **93**, 996–1002.

Lee, S. and Lewbel, A. (2009), Nonparametric identification of accelerated failure time competing risks models. Working Paper, Boston College.

Lewbel, A. (2007), 'Endogenous selection or treatment model estimation', *Journal of Econometrics* **141**, 777–806.

Roy, A. D. (1951), 'Some thoughts on the distribution of earnings', *Oxford Economic Papers(New Series)* **3**, 135–146.

Vella, F. (1998), 'Estimating models with sample selection bias: a survey', *Journal of Human Ressources* **33**, 127–169.

Willis, R. and Rosen, S. (1979), 'Education and self-selection', *Journal of Political Economy* **87**, S7–S36.