Inference on an Extended Roy Model, with an Application to Schooling Decisions in France *

Xavier D'Haultfœuille † Arnaud Maurel‡

First version: December 2009
This version: October 2012

Abstract

This paper considers the identification and estimation of an extension of Roy's model (1951) of sectoral choice, which includes a non-pecuniary component in the selection equation and allows for uncertainty on potential earnings. We focus on the identification of the non-pecuniary component, which is key to disentangle the relative importance of monetary incentives versus preferences in the context of sorting across sectors. By making the most of the structure of the selection equation, we show that this component is point identified from the knowledge of the covariate effects on earnings, as soon as one covariate is continuous. Notably, and in contrast to most results on the identification of Roy models, this implies that identification can be achieved without any exclusion restriction nor large support condition on the covariates. As a byproduct, bounds are obtained on the distribution of the *ex ante* monetary returns. We propose a three-stage semiparametric estimation procedure for this model, which yields root-n consistent and asymptotically normal estimators. Finally, we apply our results to the educational context, by providing new evidence from French data that non-pecuniary factors are a key determinant of higher education attendance decisions.

JEL Classification: C14, C25 and J24

Keywords: Roy model, nonparametric identification, schooling choices, *ex ante* returns to schooling.

*We are grateful to the editor, Han Hong, the associate editor and two anonymous referees for helpful comments. We also thank Victor Aguirregabiria, Christian Belzil, Gerard van den Berg, Federico Bugni, Stephen Cosslett, Philippe Février, Marc Gurgand, Marc Henry, Bo Honoré, Shakeeb Khan, Simon Lee, Thierry Magnac, Enno Mammen, Salvador Navarro, David Neumark, Jean-Marc Robin, Christopher Taber and participants at numerous seminars and conferences for useful discussions and comments.

†CREST. E-mail address: xavier.dhaultfoeuille@ensae.fr.

‡Duke University and IZA. E-mail address: apm16@duke.edu.

1 Introduction

Self-selection is probably one of the major issue economists have to deal with when trying to measure causal effects such as, among others, wage returns to education, migration and occupation wage premia. The seminal Roy's model (1951) of occupational choice can be seen as an extreme setting of self-selection, where agents choose between two sectors by maximizing their wage. The idea underlying this model has been very influential in the analysis of choices of participation to the labor market (Heckman, 1974), union versus nonunion status (Lee, 1978, Robinson & Tomes, 1984), public versus private sector (Dustmann & van Soest, 1998), college attendance (Willis & Rosen, 1979), migration (Borjas, 1987), training program participation (Ashenfelter & Card, 1985, Ham & LaLonde, 1996) as well as occupation (Dolton et al., 1989).

The standard Roy model, is, however, restrictive in at least two dimensions. First, non-pecuniary aspects matter much in general. For instance, in the context of educational choice, it is most often assumed that individuals consider not only the investment value of schooling, which is related to wage returns, but also the non-pecuniary consumption value of schooling, which relates to preferences and schooling ability. Recent empirical evidence suggests that these non-pecuniary factors are indeed a key determinant of schooling decisions (see, e.g., Carneiro et al., 2003, Arcidiacono, 2004, and Beffy et al., 2012). Non-pecuniary aspects such as working conditions may also matter when choosing an occupation. Similarly, migration decisions are likely to be driven both by monetary returns and the psychic costs associated with the decision to migrate (see, e.g., Bayer et al., 2011). Second, as emphasized by a recent stream of the literature on schooling choices (see Cunha & Heckman, 2007, for a survey), agents most often do not anticipate perfectly their potential earnings in each sector at the moment of their decision. Because of this *ex ante* uncertainty, their decision depends on expectations of these potential earnings rather than on their true values.

This paper focuses on the identification of the non-pecuniary factors in an extended Roy model including these two aspects.[1] The model we consider in the paper includes a non-pecuniary component which is allowed to vary across individuals according to observed covariates. Namely, denoting by Y_d the potential earnings in sector $d \in \{0, 1\}$ (and by D the corresponding random variable), $\mathcal{I}$ the information set of the agent at the time of the choice and $G(X)$ the non-pecuniary component, we consider throughout

[1]The seminal work by Heckman & Honoré (1990) examines the identification of the standard Roy model (see Buera, 2006, for an extension to non-separable functional forms for the potential outcomes).

the paper a selection equation of the form:

$$D \quad = \quad \mathbb{1}\{E(Y_1|\mathcal{I}) > E(Y_0|\mathcal{I}) + G(X)\}$$

While much emphasis has been put in the literature on the identification of the distribution of the potential earnings (Y_0, Y_1) in the presence of endogenous selection, still relatively little attention has been geared towards the selection process itself. However, as put forward by Cunha & Heckman (2007), providing evidence on the structural determinants of sectoral choice, which correspond to what agents act on, is of clear interest. In particular, identifying the non-pecuniary factors is key to disentangle the relative importance of monetary incentives versus preferences in the context of sorting across sectors.

By making the most of the structure of the selection process, we show that this non-pecuniary component is point identified from the knowledge of the covariate effects on earnings, as soon as one covariate is continuous. When all covariates are discrete, our strategy can be naturally adapted to yield informative bounds. We then propose two alternative strategies for identifying the covariate effects on sector-specific earnings. The first one is based on exclusion restrictions. It requires either a "standard" instrument, i.e. a variable affecting the selection probability but not the potential earnings, or sector-specific variables a la Heckman & Sedlacek (1985, 1990). The second strategy builds on an argument at infinity for the potential outcomes, relying on a result from a companion paper (D'Haultfoeuille & Maurel, 2012). This latter approach does not require any exclusion restriction, nor any large support on the covariates. Taken together, these results imply that the non-pecuniary component can be identified without any exclusion restriction nor large support condition on the covariates. This conclusion contrasts sharply with the identification results for generalized Roy models, which allow for unobserved determinants of the non-pecuniary component. As stressed by French & Taber (2011), identification of this more general class of models hinges on the availability of exclusion restrictions and large support regressors. Overall, this is a key advantage of our model relative to the generalized Roy specification (see Heckman & Vytlacil, 2007 and French & Taber, 2011). Importantly, we also provide some evidence suggesting that, in the case where the data is actually generated from a generalized Roy model, the misspecification bias on the non-pecuniary component is likely to be negligible relative to the finite sample estimation error.

As a byproduct of this analysis, we obtain informative bounds on the distribution of the *ex ante* returns, which correspond to the monetary returns expected by the agent at the time of the choice and are also equal, in our setting, to marginal treatment effects

evaluated at certain margins. We also provide support conditions under which these bounds shrink to a point. In particular, standard average treatment effect parameters are point identified if the probability of selection ranges from zero to one, a result in line with that of Heckman & Vytlacil (2005) in the case of local instrumental variable strategies. Noteworthy, unlike Carneiro et al. (2003) and Cunha & Heckman (2007) who impose less structure on the selection equation, the *ex ante* returns are identified without any exclusion restriction. To the extent that convincing exclusion restrictions may in practice be hard to come by, we view this as a clear benefit from using our framework.

In a recent article investigating the identification of an extended Roy model with a focus on non-pecuniary factors, Bayer et al. (2011) also propose a strategy which does not require any exclusion restriction nor large support condition. However, they specify an extended Roy model which does not account for *ex ante* uncertainty on the outcomes and restrict the alternative-specific non-pecuniary factors to be constant across individuals. Their model also differs from ours in that they consider a setting with potentially more than two sectors, so that our framework does not nest their model. They show that the non-pecuniary factors as well as the unconditional wage distributions are identified provided that the distribution of monetary returns has a finite lower bound. Although appealing in that neither exclusion restrictions nor strong support conditions are required, the finite lower bound condition may be restrictive and the strategy hard to apply in practice, notably when using log wages which do not have a natural lower bound, as for instance in Willis & Rosen (1979) and in our application.[2]

Apart from identification, we propose a three-stage semiparametric estimation procedure under an index restriction on the effects of covariates. The first two stages allow us to estimate the covariate effects on potential earnings and correspond to Newey's method (2009) for estimating semiparametric selection models. The originality of the proposed estimation procedure lies in its third stage, which is devoted to the non-pecuniary component. This stage simply amounts to estimating a linear instrumental model. The difference with a standard IV approach is that both the dependent variable and one of the regressors have to be estimated, this involving in particular a nonparametric regression on generated covariates. We show that the corresponding estimator is root-n consistent and asymptotically normal.

Eventually, we apply our estimation procedure to the context of higher education atten-

[2]Bayer et al. (2011) alternatively prove identification assuming independence between the potential wages. We do not make this assumption in the paper.

dance decisions in France over the nineties. We estimate semiparametrically a model a la Willis & Rosen (1979), which is extended to account for non-pecuniary factors driving the attendance decision. We use respectively the local average incomes for high school and higher education graduates as sector-specific regressors, this yielding identification of the covariate effects on earnings. As could be expected, we cannot reject (at the 10% level) the hypothesis that the local average income for high school graduates only affects the probability of attendance through the *ex ante* returns to higher education. This allows us to apply a constrained version of our estimator, leading to substantial gains of precision. Consistent with the recent evidence on this question, our results suggest that non-pecuniary factors are a key determinant of the decision to attend higher education. We find in particular that 10% of the individuals attending higher education choose to do so in spite of negative *ex ante* monetary returns to education. Besides, it follows from our estimates that the higher education attendance rate would fall from 83.1% to 72% if non-pecuniary factors did not exist. This decrease is eight time larger than the one associated with a 10% permanent decrease in labor market earnings of higher education attendees.

The remainder of the paper is organized as follows. Section 2 presents the extended Roy model which is considered throughout the paper, derives our key identification results for the non-pecuniary component and the distribution of the *ex ante* returns before discussing the identification of the covariate effects on earnings. Section 3 develops a semiparametric estimation procedure for this model, and proves the root-n consistency and asymptotic normality of the proposed estimators. Section 4 applies the preceding estimators to investigate the influence of non-pecuniary factors on higher education attendance decision in France. Finally, Section 5 concludes. The online appendix collects Monte Carlo simulations, the proofs of our results as well as additional details on the application.

2 Identification

2.1 The setting

We consider an extension of the Roy model which is obtained by including *ex ante* uncertainty as well as non-pecuniary factors in the seminal Roy's model (1951) of occupational choice. Suppose that there are two sectors 0 and 1 in the economy, and let $Y_k,\ k \in \{0,1\}$, denote the individual's potential earnings in sector k. These earnings

are not perfectly observed by the individual at the time of her decision. Instead, she can only compute the expectation $E(Y_k|X, \eta_0, \eta_1)$, where $X \in \mathbb{R}^p$ are covariates observed by the econometrician and (η_0, η_1) are sector-specific productivity terms known by the agent at the time of the choice but unobserved by the econometrician. We maintain the following assumption throughout the paper.

Assumption 2.1 *(Additive decomposition) We have, for* $k \in \{0, 1\}$, $E(Y_k|X, \eta_0, \eta_1) = E(Y_k|X, \eta_k) = \psi_k(X) + \eta_k$. *Moreover,* $X \perp\!\!\!\perp (\eta_0, \eta_1)$.

The independence assumption ($X \perp\!\!\!\perp (\eta_0, \eta_1)$) is commonly made when studying sample selection models (see, e.g., Powell, 1994) or Roy models (see, e.g., Heckman & Honoré, 1990, for the standard Roy model and French & Taber, 2011, for generalized Roy models). We shall discuss further in the paper how this assumption could be weakened.

We let hereafter $\nu_k = Y_k - E(Y_k|X, \eta_0, \eta_1)$ denote the unexpected shock on Y_k and $\varepsilon_k = \eta_k + \nu_k$ denote the sector-specific residual.[3] Noteworthy, apart from the independence assumption, we do not impose any restriction on $(\eta_0, \eta_1, \nu_0, \nu_1)$, thus departing from, e.g., Carneiro et al. (2003) who posit a factor structure on the unobservables. Such a restriction is useful to identify the joint distribution of $(\eta_0, \eta_1, \nu_0, \nu_1)$, and thus to test for comparative advantage or to assess the importance of unobserved heterogeneity (see Cunha & Heckman, 2007). We do not consider these issues here.

Unlike Roy's original model, we do not suppose that the sectoral choice is based only on income maximization. Instead, we suppose that each individual chooses to enter the sector which yields the highest expected utility, with the expected utility in sector k writing as $\mathcal{U}_k = E(Y_k|X, \eta_0, \eta_1) + G_k(X)$. $\mathcal{U}_k$ is assumed to be given by the sum of sector-specific expected earnings $E(Y_k|X, \eta_0, \eta_1)$ and the non-pecuniary component associated with sector k, $G_k(X)$, which is supposed to depend on the covariates X. Assuming additive separability between the expected earnings and the non-pecuniary component of utility is standard for the generalizations of the Roy model considered in the literature.[4] This separability assumption, which is required to obtain an additive separable form between X and η_Δ in the selection index, is key for our identification strategy.[5] Along with the covariates X, the econometrician observes the chosen sector

[3]Part of the residual ν_k may correspond to a measurement error rather than an unexpected shock. We use the latter interpretation throughout the paper for convenience of exposition only.

[4]Note that $-G_k(X)$ can be seen as a cost of entry into sector k. This interpretation is put forward in the treatment effect literature relying on generalized Roy models.

[5]This echoes the fact that additive separability in the selection index is crucial for the identification results obtained in the Marginal Treatment Effects literature (see, e.g., Heckman & Vytlacil,2005).

D, which satisfies

$$\begin{aligned} D &= \mathbb{1}\{\mathcal{U}_1 > \mathcal{U}_0\} \\ &= \mathbb{1}\{\eta_\Delta > \psi_0(X) - \psi_1(X) + G(X)\}, \end{aligned} \tag{2.1}$$

where $G(X) = (G_0 - G_1)(X)$ and $\eta_\Delta = \eta_1 - \eta_0$. Finally, the econometrician also observes the earnings in the chosen sector, that is

$$Y = DY_1 + (1 - D)Y_0.$$

This model is known in the literature as the extended Roy model, whose identification is also considered, in a version without *ex ante* uncertainty, by Heckman & Vytlacil (2007). Bayer et al. (2011) examine the identification of an extended Roy model without *ex ante* uncertainty as well, which allows for more than two sectors and includes a non-pecuniary intercept for each sector. In a recent paper, Fox & Gandhi (2011) extend this model by allowing for random functions in the selection equation.[6] The model presented above can be applied to various economic settings, including sectoral choice in the labor market, immigration or higher education attendance decisions (see our application in Section 5). A central contribution of this paper is to show that, by making the most of the extended Roy structure, the identification of the covariate effects on earnings directly entails the identification of the non-pecuniary component. In particular, unlike in Heckman & Vytlacil (2007) and Fox & Gandhi (2011), no exclusion restriction between G and (ψ_0, ψ_1) is needed.

2.2 Identification of the non-pecuniary component

Since our main contribution relates to the identification of the non-pecuniary component, we first discuss this issue, and suppose for now that the covariate effects on earnings (ψ_0, ψ_1) are known. Discussion of the identification of (ψ_0, ψ_1) is deferred to Subsection 2.4.[7] Our identification strategy for the non-pecuniary component fully relies on the detailed structure of the model, and in particular on the link between the residuals in the outcome equations and the one in the selection equation. We first suppose that conditional on the other components of X, at least one component X_j, say

[6] However, as is the case for the model we consider, Fox & Gandhi (2011) rule out the existence of additive errors for the non-pecuniary components entering the selection model.

[7] What we mean by identification throughout the paper is that these functions are uniquely defined almost everywhere by the model and the data generating process. "Almost everywhere" can be replaced by "everywhere" under for instance continuity conditions.

X_1, is continuous, and we let $X = (X_1, X_{-1})$ (and we let similarly $x = (x_1, x_{-1})$). We also impose a mild regularity condition on $T = \psi_0 - \psi_1$, G and the error terms of the outcome equation. Assumption 2.3 below is a technical condition which is usual in Roy or competing risks models (see, e.g., Heckman & Honoré, 1990, or Lee, 2006).

Assumption 2.2 *For all x_{-1} in the support of X_{-1}, the distribution of X_1 conditional on $X_{-1} = x_{-1}$ is continuous and $T(., x_{-1})$ and $G(., x_{-1})$ are differentiable on the support of X_1 conditional on X_{-1}.*

Assumption 2.3 *(Restrictions on the errors, 1) $E(|\varepsilon_k|) < \infty$ for $k \in \{0, 1\}$. The distribution of η_Δ admits a continuous density f_{η_Δ} with respect to the Lebesgue measure and for all $u \in \mathbb{R}$, $f_{\eta_\Delta}(u) > 0$.*

We start from the following observations:

$$\begin{aligned} E[D\eta_\Delta|X] &= E\left[\mathbb{1}\{\eta_\Delta \geq T(X) + G(X)\}\eta_\Delta|X\right] = \int_{T(X)+G(X)}^{\infty} u f_{\eta_\Delta}(u)du, \quad (2.2) \\ E[D|X] &= \int_{T(X)+G(X)}^{\infty} f_{\eta_\Delta}(u)du. \quad (2.3) \end{aligned}$$

By the fundamental theorem of calculus and Assumptions 2.2-2.3, the functions $q_0(x) = E(D|X = x)$ and $E[D\eta_\Delta|X = x]$ are continuously differentiable with respect to x_1, and

$$\frac{\partial E[D\eta_\Delta|X = x]}{\partial x_1} = (T(x) + G(x))\frac{\partial q_0}{\partial x_1}(x). \quad (2.4)$$

for almost all x_{-1} and all x_1 in the support of X_1 conditional on $X_{-1} = x_{-1}$. Because T and q_0 are identified, this equation shows that, provided that $\partial q_0/\partial x_1(x) \neq 0$, identification of $G(x)$ amounts to recovering $\partial E[D\eta_\Delta|X = x]/\partial x_1$. The key idea, for that purpose, is to relate this term with the residual ε of the (realized) outcome equation. Observe that by definition of ν_i and the law of iterated expectations, $E(\nu_k|D = k, X) = 0$. As a result, letting $\varepsilon = D\varepsilon_1 + (1 - D)\varepsilon_0$, we get

$$\begin{aligned} E(\varepsilon|X) &= E[D\varepsilon_1 + (1 - D)\varepsilon_0|X] \\ &= E[D\eta_1 + (1 - D)\eta_0|X] \\ &= E\left[D\eta_\Delta|X\right] + E[\eta_0]. \quad (2.5) \end{aligned}$$

Thus, letting $g_0(x) = E(\varepsilon|X = x)$, we obtain

$$\frac{\partial g_0}{\partial x_1}(x) = (T(x) + G(x))\,\frac{\partial q_0}{\partial x_1}(x). \quad (2.6)$$

Since $\varepsilon = Y - \psi_D(X)$ is identified (where we let $\psi_D = D\psi_1 + (1 - D)\psi_0$), g_0 and q_0 are identified and we can use Equation (2.6) to recover G. The only exception is actually

when $\frac{\partial q_0}{\partial x_1}$ is identically equal to zero, a case which is ruled out by Assumptions 2.3 and 2.4 below. Theorem 2.1 shows that, under these conditions, G is point identified.[8]

Assumption 2.4 *For all x_{-1} in the support of X_{-1}, the set $\{x_1 : \frac{\partial(T+G)}{\partial x_1}(x_1, x_{-1}) \neq 0\}$ is not empty.*

Theorem 2.1 *Suppose that T is identified and Assumptions 2.1-2.4 hold. Then G is identified.*

The independence condition between X and (η_0, η_1) plays an important role in the derivation above. However, this assumption could be weakened to the conditional independence condition $X_1 \perp\!\!\!\perp (\eta_0, \eta_1)|X_{-1}$, without affecting the identification result. We maintain the stronger independence assumption here for the sake of notational simplicity.

Now consider the case where no component of X is continuous, so that X has a discrete distribution. Suppose that it takes $M < \infty$ values $x_1, ..., x_M$. Then one cannot take the derivative of g_0 and q_0 anymore. However, the strategy above can be adapted to yield bounds on G, replacing derivatives with finite differences. First, note that $P(D = 0|X = x) = F_{\eta_\Delta}(T(x) + G(x))$, with F_{η_Δ} denoting the cumulative distribution function of η_Δ. This equality implies that we can sort the x_i's so that $T(x_1) + G(x_1) < ... < T(x_M) + G(x_M)$.[9] This provides a first set of inequalities on $(G(x_1), ..., G(x_M))$. Besides, letting $i < j$, we have,

$$\begin{aligned} & \sum_{k=i}^{j-1} \left[T(x_{k+1}) + G(x_{k+1})\right] \left[q_0(x_{k+1}) - q_0(x_k)\right] \\ \leq \; & g_0(x_j) - g_0(x_i) = -\int_{T(x_i)+G(x_i)}^{T(x_j)+G(x_j)} u f_{\eta_\Delta}(u) du \\ \leq \; & \sum_{k=i}^{j-1} \left[T(x_k) + G(x_k)\right] \left[q_0(x_{k+1}) - q_0(x_k)\right]. \end{aligned}$$

These inequalities provide supplementary conditions for $(G(x_1), ..., G(x_M))$. Note that we only get an upper bound for $G(x_1)$ and a lower bound for $G(x_M)$, but both for $G(x_2), ..., G(x_{M-1})$.

[8] If Assumption 2.4 fails to hold, $\frac{\partial G}{\partial x_1}$ is still identified (but not G), as it is equal to $-\frac{\partial T}{\partial x_1}$ in this case. Besides, since Assumption 2.4 implies that $\frac{\partial q_0}{\partial x_1}$ is not identically equal to zero, this restriction can be tested in the data.

[9] This is without loss of generality. In case of ties between $T(x_i) + G(x_i)$ and $T(x_{i+1}) + G(x_{i+1})$, one may remove x_{i+1} from the set of x's. Then the bounds on $G(x_{i+1})$ follow directly from those on $G(x_i)$.

When deriving our estimation procedure in Section 3, consistent with the framework of our application, we will maintain the assumptions ensuring that the non-pecuniary component G is point identified. We leave in particular the analysis of set-estimation of G for further research.

2.3 Distribution of *ex ante* returns

We now turn to the identification of the distribution of the *ex ante* returns, $\Delta = E(Y_1 - Y_0|X, \eta_0, \eta_1)$. The *ex ante* return is meaningful since it corresponds to what agents act on (see Cunha & Heckman, 2007). Besides, it corresponds to the *ex post* return if (i) agents perfectly observe or anticipate their potential outcomes (in which case $\nu_0 = \nu_1 = 0$) or if (ii) the idiosyncratic shocks are equal across sectors ($\nu_0 = \nu_1$), as postulated in standard regression models. Although we have remained completely agnostic on the information set of the agents, it is possible to point or partially identify the distribution of Δ. The intuition behind is similar to that underlying the identification of G. Δ depends on η_Δ, which is also the residual of the selection equation. Thus, the observed choice of sector directly provides information on these *ex ante* returns. To see this, first recall that

$$P(D = 0|X) = F_{\eta_\Delta}(T(X) + G(X)).$$

This shows that F_{η_Δ} is identified over the support of $T(X)+G(X)$. Now, the cumulative distribution function of Δ (F_Δ) satisfies

$$\begin{aligned} F_\Delta(u) &= E\left[P(\eta_\Delta \leq u + T(X)|X)\right] \\ &= E\left[F_{\eta_\Delta}(u + T(X))\right]. \end{aligned}$$

Hence, we can identify $F_\Delta(u)$ for all u such that the support of $u + T(X)$ is included in the support of $T(X) + G(X)$. In particular, the complete distribution of the *ex ante* returns Δ is identified as soon as $T(X) + G(X)$ has a large support. In that case, one can recover standard treatment effect parameters such as the average treatment effect or the average treatment on the treated (i.e. for the individuals such that $D = 1$ here), by integrating the *ex ante* returns over the distribution of η_Δ. Even if this large support condition fails, it is still possible to point identify a subset of the distribution of the *ex ante* returns, and bound $F_\Delta(u)$ for the rest of the distribution.[10] Indeed, letting $[\underline{M}, \overline{M}]$

[10]Heckman & Vytlacil (2007) also obtain bounds on the average returns without assuming large support on the selection probability, in the context of an extended Roy model. Their strategy hinges on an exclusion restriction between the selection equation and the potential outcomes.

(resp. $[\underline{P}, \overline{P}]$) denote the support of $T(X)+G(X)$ (resp. of $P(D=0|X)$), we have, by the monotonicity of F_{η_Δ}, $F_\Delta(u) \in [\underline{F_\Delta}(u), \overline{F_\Delta}(u)]$, where

$$\begin{aligned}
\underline{F_\Delta}(u) &= E\left(F_{\eta_\Delta}\left(u+T(X)\right)\mathbb{1}\{u+T(X)\in[\underline{M},\overline{M}]\}\right) \\
&\quad +\overline{P}\times P(u+T(X)>\overline{M})+0\times P(u+T(X)\leq \underline{M}), \qquad (2.7)\\
\overline{F_\Delta}(u) &= E\left(F_{\eta_\Delta}\left(u+T(X)\right)\mathbb{1}\{u+T(X)\in[\underline{M},\overline{M}]\}\right) \\
&\quad +1\times P(u+T(X)>\overline{M})+\underline{P}\times P(u+T(X)\leq \underline{M}). \qquad (2.8)
\end{aligned}$$

The distribution of the *ex ante* treatment effect on the treated can be identified in a similar way, with

$$F_{\Delta|D=1}(u)=\frac{E\{(F_{\eta_\Delta}\left(u+T(X)\right)-P(D=0|X))\times\mathbb{1}\{G(X)\leq u\}\}}{P(D=1)}. \qquad (2.9)$$

In our setting, the *ex ante* return Δ is closely related to the marginal treatment effect Δ^{MTE} (Heckman & Vytlacil (2005)). Indeed, denoting by S_{η_Δ} the survival function of η_Δ, we have, under Assumption 2.3,

$$\begin{aligned}
\Delta^{MTE}(x,u) &= E(Y_1-Y_0|X=x, S_{\eta_\Delta}(\eta_\Delta)=u) \\
&= \psi_1(x)-\psi_0(x)+S^{-1}_{\eta_\Delta}(u)
\end{aligned}$$

Thus, $\Delta=(\psi_1-\psi_0)(X)+\eta_\Delta$ coincides with $\Delta^{MTE}(X, S_{\eta_\Delta}(\eta_\Delta))$. Besides, one is able to identify $\Delta^{MTE}(x,u)$ for all u in the support of $P(D=1|X)$, since in that case there exists $\widetilde{x}$ in the support of X such that $S^{-1}_{\eta_\Delta}(u)=(\psi_0-\psi_1+G)(\widetilde{x})$.

2.4 Identification of the covariate effects on earnings

We now relax the assumption that the covariate effects on earnings are known, and discuss in this subsection two alternative strategies to identify (ψ_0,ψ_1). In both strategies, we impose the following normalization, which is innocuous since adding a constant to ψ_k and subtracting it to η_k does not modify the model.

Assumption 2.5 *(Normalization) There exists x^* in the support of X such that $\psi_0(x^*)=\psi_1(x^*)=0$.*

The first and standard approach we focus on is based on exclusion restrictions, in the same spirit as, e.g., Das et al. (2003). The second hinges on a nonstandard identification at infinity, with the advantage of not requiring any exclusion restriction. The first strategy relies on the following assumption.

Assumption 2.6 *(Exclusion restrictions) ψ_0 (resp. ψ_1) depends only on $\widetilde{X}_0 \subset X$ (resp. on $\widetilde{X}_1 \subset X$). Moreover, $\widetilde{X}_0$ (resp. $\widetilde{X}_1$) and $P(D = 1|X)$ are measurably separated, that is, any function of $\widetilde{X}_0$ (resp. of $\widetilde{X}_1$) almost surely equal to a function of $P(D = 1|X)$ is almost surely constant.*

The first part of Assumption 2.6 covers two rather different situations. The first one is when $X = (\widetilde{X}_0, Z)$ and $\widetilde{X}_1 = \widetilde{X}_0$. This corresponds to the standard instrumental setting in sample selection models, where the instrument Z affects the probability of selection but not the potential outcomes. In our framework, Z would be a determinant of the non-pecuniary component but not of the potential earnings. The second situation corresponds to the case where $X = (X_0, X_1, X_c)$, $\widetilde{X}_0 = (X_0, X_c)$ and $\widetilde{X}_1 = (X_1, X_c)$. This occurs in the presence of sector-specific regressors. In this case, no exclusion restriction between the non-pecuniary factors and the potential earnings is required. This kind of exclusion restrictions was previously used in particular by Heckman & Sedlacek (1985, 1990) when estimating parametrically a multiple-sector Roy model of self-selection in the labor market. We also use sector-specific regressors in our application.

Intuitively, the measurable separation requirement[11] of Assumption 2.6 ensures that $\psi_0(X)$ (or $\psi_1(X)$) and $P(D = 1|X)$ can vary in a sufficiently independent way. This assumption, also made by Das et al. (2003), is weak when, considering the two cases above, Z or (X_0, X_1) is continuous (see Florens et al., 2008, for sufficient conditions in this case). However, it may not hold when Z (or (X_0, X_1)) is discrete. As an illustration, consider a standard instrumental setting where $\widetilde{X}_0$ and Z are binary and let $P_{ij} = P(D = 1|\widetilde{X}_0 = i, Z = j)$ for $i, j \in \{0, 1\}$. Then, provided that P_{10} and P_{11} do not belong to $\{P_{00}, P_{01}\}$, there exists a function h such that $h(P_{00}) = h(P_{01})$ and $h(P_{10}) = h(P_{11})$ but $h(P_{00}) \neq h(P_{10})$. In this case, the function g defined by $g(0) = h(P_{00})$ and $g(1) = h(P_{10})$ is not constant. As a result, $\widetilde{X}_0$ and $P(D = 1|X)$ are not measurably separated.

Given the preceding exclusion restrictions and the additive decomposition assumption, it is possible to identify ψ_0 and ψ_1 up to location parameters. Then full identification stems from the normalization of Assumption 2.5. Similarly to Das et al. (2003), Proposition 2.2 below does not provide any result on the location parameters. In general, such parameters are identified only at infinity under a large support condition, i.e. when $P(D = 1|X)$ can be arbitrarily close to zero and one (see Heckman, 1990).

Proposition 2.2 *Suppose that Assumptions 2.1, 2.3, 2.5 and 2.6 hold. Then ψ_0 and ψ_1 are identified.*

[11] We adopt here the terminology of Florens et al. (2008) (see their Assumption A4).

Proposition 2.2 is similar to Theorem 2.1 of Das et al. (2003), but identification is shown here without assuming that the regressors are continuous nor that ψ_0 and ψ_1 are continuously differentiable. The idea behind the proof of Proposition 2.2 is that $E(\varepsilon_k|D = k, X)$ only depends on $P(D = 1|X)$. We can then rely on the measurable separability condition of Assumption 2.6 to prove the result. Because identification is based on $P(D = 1|X)$ only, the structure imposed on the selection equation is not needed at this stage, whereas, as stressed above, it is crucial to identify the non-pecuniary component and the distribution of the *ex ante* returns. Finally, following Das et al. (2003), one could actually relax the independence condition $X \perp\!\!\!\perp (\eta_0, \eta_1)$ and allow for endogenous covariates X while still identifying ψ_0 and ψ_1 up to location. However, it is not clear in this case how to recover the non-pecuniary component G.

We now also show, using a result from a companion paper (D'Haultfoeuille & Maurel, 2012), that ψ_0 and ψ_1 can be identified at the limit without any exclusion restriction, under the following restrictions on the error terms.

Assumption 2.7 *(Restrictions on the errors, 2) (i) $X \perp\!\!\!\perp (\varepsilon_0, \varepsilon_1)$, (ii) for $k \in \{0, 1\}$, the supremum of the support of ε_k is infinite and there exists $b_k > 0$ such that $E(\exp(b_k\varepsilon_k)) < \infty$, (iii) for all $u \in \mathbb{R}$,*

$$\lim_{v\to\infty} P(\eta_k - \eta_{1-k} > u|\eta_k + \nu_k = v) = 1, \quad k \in \{0, 1\}.$$

The first restriction reinforces the condition that $X \perp\!\!\!\perp (\eta_0, \eta_1)$, by ruling out in particular heteroskedasticity of the shocks (ν_0, ν_1). The second restriction is a light tail condition, which is in practice fairly mild. If we consider the example of log-wages $Y_k = \ln W_k$, the assumption is satisfied provided that there exists $b_k > 0$ such that $E(W_k^{b_k}) < \infty$. Hence, it holds even if wages have fat tails, Pareto-like for instance. The last restriction can be interpreted as a moderate dependence condition between η_0 and η_1, which is not very restrictive either. When $(\eta_0, \eta_1, \nu_0, \nu_1)$ is gaussian for instance, one can show that it is equivalent to $\text{cov}(\eta_0, \eta_1) < \min(V(\eta_0), V(\eta_1))$. In particular, when $V(\eta_0) = V(\eta_1)$, this condition is automatically satisfied, except in the degenerate case where $\eta_0 = \eta_1$.

Proposition 2.3 *Suppose that Assumptions 2.1, 2.5 and 2.7 hold. Then ψ_0 and ψ_1 are identified.*

This result does not follow from the typical identification at infinity strategy for sample selection models, which relies on the fact that the selection probability tends to zero or

one when one of the regressors takes arbitrarily large values. Rather, the intuition can be described as follows. First, the moderate dependence restriction on (η_0, η_1), together with the extended Roy structure of the selection equation, ensures that

$$\lim_{y\to\infty} P(D = k|X = x, Y_k = y) = 1, \text{ for all } x \text{ and } k \in \{0, 1\}. \tag{2.10}$$

In other words, individuals whose potential outcome in one sector tends to infinity will choose this sector with a probability approaching one, whatever their observed characteristics $X = x$. This is because these individuals will have, with a large probability, a smaller potential outcome in the other sector, even though the latter may also be large on average.

In turn, (2.10) implies that the right tails of the observed and potential outcomes are similar. Formally, one can show that as $y \to \infty$,

$$\lim_{y\to\infty} \frac{P(Y \geq y, D = k|X = x)}{S_{\varepsilon_k}(y - \psi_k(x))} = 1.$$

As a result,

$$\lim_{y\to\infty} \frac{P(Y \geq y - \psi_k(x^*), D = k|X = x)}{P(Y \geq y - \psi_k(x), D = k|X = x^*)} = 1.$$

It follows from the location normalization imposed in Assumption 2.5 ($\psi_k(x^*) = 0$) that

$$u = \psi_k(x) \Rightarrow \lim_{y\to\infty} \frac{P(Y \geq y, D = k|X = x)}{P(Y \geq y - u, D = k|X = x^*)} = 1. \tag{2.11}$$

Because the function $y \mapsto P(Y \geq y, D = k|X = x)$ is identified for each x, $\psi_k(x)$ is identified provided that the converse of (2.11) also holds. The latter implication is ensured by Assumption 2.7 (ii).

This type of identification at infinity is similar to the one used by Heckman & Honoré (1989) and Abbring & van den Berg (2003) in the related competing risks model. Nevertheless, their results cannot be used here because their strategies break down when turning to extended Roy models.[12] An appealing feature of Condition (2.10) is that it is testable (see D'Haultfoeuille & Maurel, 2012). Besides, this identification strategy does not rely on any support condition on X. In particular, it may be applied even if X is discrete.[13] On the other hand, estimators corresponding to this setting have not been derived yet. We therefore restrict in the estimation part (Section 3) to the case where exclusion restrictions are available.

[12]Lee (2006) and Lee & Lewbel (2012) obtain identification of competing risks models without using arguments at the limit. Their strategy cannot be extended easily to extended Roy models either.

[13]If one of the covariates has large support, one can use alternatively the results of Lewbel (2007) which also yield identification of the covariate effects on earnings without any instrument for selection.

3 Semiparametric estimation

Although our identification results hold in a nonparametric setting, we focus here on semiparametric estimation in order to provide root-n consistent and asymptotically normal estimators of ψ_0, ψ_1 and G. More precisely, we consider a class of extended Roy models with a linear index structure of the form:

$$\left\{ \begin{array}{rcl} Y_0 &=& X'\beta_0 + \varepsilon_0 \\ Y_1 &=& X'\beta_1 + \varepsilon_1 \\ D &=& \mathbb{1}\{-\delta_0 + X'(\beta_1 - \beta_0 - \gamma_0) + \eta_\Delta > 0\}. \end{array} \right. \tag{3.1}$$

Here, our normalization on $\psi_0(.)$ and $\psi_1(.)$ is that $\psi_0(0) = \psi_1(0) = 0$.[14] In our setting, the non-pecuniary component $G(X)$ is of the form $\delta_0 + X'\gamma_0$. Let γ_{0j} (resp. β_{0j}, β_{1j}) denote the j-th component of γ_0 (resp. β_0, β_1). We impose the following conditions.

Assumption 3.1 *(Exclusion restrictions) There exists j_1 and j_2 such that $\beta_{0j_1} = \beta_{1j_2} = 0$, $\gamma_{0j_1} \neq \beta_{1j_1}$ and $\gamma_{0j_2} \neq -\beta_{0j_2}$.*

Assumption 3.2 *(Regularity of X) The support of X is bounded and not contained in a proper subset of $\mathbb{R}^p$. For all x_{-1} in the support of X_{-1}, the distribution of X_1 conditional on $X_{-1} = x_{-1}$ admits a continuously differentiable and positive density on its support, which is a compact interval independent of x_{-1}. Besides, $\beta_{11} - \beta_{01} - \gamma_{01} \neq 0$. Moreover, the support of $X'(\beta_1 - \beta_0 - \gamma_0)$ is an interval. Finally, for all j, $t \mapsto E(X_j | X'(\beta_1 - \beta_0 - \gamma_0) = t)$ is continuously differentiable.*

Assumption 3.3 *(i.i.d. sample) We observe a sample $(Y_i, X_i, D_i)_{1 \leq i \leq n}$ of i.i.d. copies of (Y, X, D).*

Assumption 3.1 corresponds, in this semiparametric framework, to Assumption 2.6. The case where $j_1 = j_2$ corresponds to the standard instrumental variable setting of sample selection models, while $j_1 \neq j_2$ applies when some covariates are sector-specific. Assumption 3.2 corresponds to Assumptions 2.2 and 2.4. It ensures that at least one covariate is continuous and has a nonzero effect on D (because $\beta_{11} - \beta_{01} - \gamma_{01} \neq 0$). As shown in Theorem 2.1, this condition is sufficient to provide point identification of G. We also require the support of $X'(\beta_1 - \beta_0 - \gamma_0)$ to be an interval. This condition, together with the requirement that the X_j are not colinear, is sufficient to point identify

[14] Thus, it may differ from Assumption 2.5 if zero does not belong to the support of X. Yet, this is still without loss of generality since we do not constraint the expectations of ε_0 and ε_1 to be zero.

the single index model on D (see, e.g., Horowitz, 1998) that corresponds to our first step estimator described below.

Let us assume, without loss of generality, that $\beta_{11} - \beta_{01} - \gamma_{01}$ is strictly positive. We define $\zeta_0 = (\beta_1 - \beta_0 - \gamma_0)/(\beta_{11} - \beta_{01} - \gamma_{01})$ (so that $\zeta_{01} = 1$) and $\widetilde{\eta}_\Delta = (\eta_\Delta - \delta_0)/(\beta_{11} - \beta_{01} - \gamma_{01})$. We propose a three-stage estimation procedure of the model, where we estimate first ζ_0, then (β_0, β_1) and finally (δ_0, γ_0). The first and second stages of our procedure are not new, and rely on the fact that we can rewrite the model in the following reduced form:

$$\begin{aligned} D &= \mathbb{1}\{X'\zeta_0 + \widetilde{\eta}_\Delta > 0\} \\ Y_k &= X'\beta_k + \varepsilon_k, \quad k \in \{0, 1\}, \end{aligned} \tag{3.2}$$

where Y_k is observed when $D = k$, $\widetilde{\eta}_\Delta$ is independent of X and $E(\varepsilon_k | D = k, X)$ only depends on $X'\zeta_0$.[15] Besides, by Assumption 3.1, X_{j_1} (resp. X_{j_2}) affects selection since $\zeta_{0j_1} \neq 0$ (resp. $\zeta_{0j_2} \neq 0$) but not the potential earnings Y_0 (resp. Y_1). Hence, Equations (3.2) correspond to Newey (2009)'s selection model and we follow his approach here. First, we estimate ζ_0 by a single index estimator $\widehat{\zeta}$, for which we suppose Assumption 3.4 below to be satisfied. This is the case of many semiparametric estimators, such as the one of Klein & Spady (1993) or Ichimura (1993). Secondly, we estimate β_0 and β_1 by series estimator, and we suppose that they satisfy Assumption 3.5. This condition can be obtained under more primitive assumptions (see Newey, 2009, p. S227).

Assumption 3.4 *(Regularity of the first stage estimator) There exists $(\chi_i)_{1 \leq i \leq n}$, i.i.d. random variables such that $E(\chi_i) = 0$, $E(\chi_i \chi_i')$ exists and is non singular and*

$$\widehat{\zeta} - \zeta_0 = \frac{1}{n} \sum_{i=1}^{n} \chi_i + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Assumption 3.5 *(Regularity of the second stage estimators) Let $k \in \{0, 1\}$, there exists $(\chi_{ki})_{1 \leq i \leq n}$, i.i.d. random variables such that $E(\chi_{ki}) = 0$, $E(\chi_{ki}\chi_{ki}')$ exists and is non singular and*

$$\widehat{\beta_k} - \beta_k = \frac{1}{n} \sum_{i=1}^{n} \chi_{ki} + o_P\left(\frac{1}{\sqrt{n}}\right).$$

[15] Indeed, $\varepsilon_k = \eta_k + \nu_k$ with $E(\nu_k | D = k, X) = 0$ by definition and $E(\eta_1 | D = 1, X = x) = E(\eta_1 | \widetilde{\eta}_\Delta > -x'\zeta_0)$ (and similarly for $k = 0$). Note that in general, ε_k is not independent of X because ν_k is not.

The originality of the estimation procedure lies in its third stage, which is devoted to the estimation of (δ_0, γ_0). Actually, it suffices to estimate δ_0 and $\alpha_0 \equiv \beta_{01} - \beta_{11} + \gamma_{01}$, since $\gamma_0 = \beta_1 - \beta_0 + \alpha_0\zeta_0$. Equations (2.2), (2.3) and (2.5) applied to the current index model show that $E(D|X)$ and $E(\varepsilon|X)$ only depend on $U \equiv X'\zeta_0$. Letting, with a slight abuse of notation, $q_0(u) = E(D|U = u)$ and $g_0(u) = E(\varepsilon|U = u)$, we have, similarly to Equation (2.6),

$$g_0'(U) = q_0'(U)(\delta_0 + \alpha_0 U). \tag{3.3}$$

Integrating (3.3) between u_0 in the support of U and U, we obtain:

$$g_0(U) = \widetilde{\lambda_0} + q_0(U)\delta_0 + \left[\int_{u_0}^{U} uq_0'(u)du\right]\alpha_0,$$

where $\widetilde{\lambda_0}$ is the constant of integration. An integration by part yields

$$g_0(U) = \lambda_0 + q_0(U)\delta_0 + \left[q_0(U)U - \int_{u_0}^{U} q_0(u)du\right]\alpha_0, \tag{3.4}$$

where $\lambda_0 = \widetilde{\lambda}_0 - u_0 q_0(u_0)\alpha_0$. In other terms,

$$\varepsilon = \lambda_0 + D\delta_0 + \left[DU - \int_{u_0}^{U} q_0(u)du\right]\alpha_0 + \xi, \quad E(\xi|X) = E(\xi|U) = 0. \tag{3.5}$$

Let $\theta_0 = (\lambda_0, \delta_0, \alpha_0)'$, $V = DU - \int_{u_0}^{U} q_0(u)du$ and $W = (1, D, V)'$, so that $\varepsilon = W'\theta_0 + \xi$. The regressors D and V are endogenous since selection D depends both on U and $\widetilde{\eta}_\Delta$. We therefore use an IV estimator of θ_0 with functions of the index U as instruments for D and V. To avoid boundary effects, we include some trimming by considering feasible versions of the instruments $Z = \mathbb{1}\{X \in \mathcal{X}\}h(U)$, where $h(U) = (1, h_1(U), h_2(U))' \in \mathbb{R}^3$ and $\mathcal{X}$ is a set included in the support of X and such that $\{x'\zeta_0, x \in \mathcal{X}\}$ is a closed interval strictly included in the interior of the support of U.[16] Then $\theta_0 = E(ZW')^{-1}E(Z\varepsilon)$, and we estimate it by

$$\widehat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{Z}_i\widehat{W}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{Z}_i\widehat{\varepsilon}_i\right),$$

where $\widehat{\varepsilon}_i = Y_i - X_i'(D_i\widehat{\beta}_1 + (1 - D_i)\widehat{\beta}_0)$, $\widehat{W}_i = (1, D_i, \widehat{V}_i)'$ and

$$\begin{aligned}
\widehat{V}_i &= D_i\widehat{U}_i - \int_{u_0}^{\widehat{U}_i} \widehat{q}(u, \widehat{\zeta})du, \\
\widehat{Z}_i &= \mathbb{1}\{X_i \in \mathcal{X}\}h\left(\widehat{U}_i\right).
\end{aligned}$$

Finally, $\widehat{U}_i = X_i'\widehat{\zeta}$ and

$$\widehat{q}(u, \zeta) = \frac{\sum_{i=1}^{n} D_i K\left(\frac{u - X_i'\zeta}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{u - X_i'\zeta}{h_n}\right)}. \tag{3.6}$$

[16] This trimming procedure ensures uniform consistency of kernel estimators over $\{x'\zeta_0, x \in \mathcal{X}\}$.

where $K(.)$ is a kernel function and h_n the bandwidth parameter. The result on the third step estimator $\widehat{\theta}$ relies on the following conditions on $h(.)$ and $K(.)$.

Assumption 3.6 *(Restrictions on the kernel) $K(.)$ is nonnegative, zero outside a compact set, continuously twice differentiable on this compact set and satisfies $\int K(v)dv = 1$ and $\int vK(v)dv = 0$. Moreover, $K(.)$ and $K'(.)$ are zero on the boundary of this compact set.*

Assumption 3.7 *(Regular instruments) $h_k(.)$ is twice differentiable and $|h_k''|$ is bounded for $k \in \{1, 2\}$.*

Assumption 3.6 is satisfied for instance by the quartic kernel $K(v) = (15/16)(1 - v^2)^2\mathbb{1}_{[-1,1]}(v)$. Assumption 3.7 is imposed to ensure that $\widehat{Z}_i - Z_i$ is small for large sample sizes, and behaves regularly.

Theorem 3.1 *Suppose that $nh_n^6 \to \infty$, $nh_n^8 \to 0$ and that Assumptions 2.1, 2.3, 2.4, 3.1-3.7 hold. Then*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0,\, E(ZW')^{-1}V(Z\xi + \Omega_{11} + \Omega_{21})E(WZ')^{-1}\right),$$

where Ω_{11} is defined by Equation (3.7) in the online appendix and

$$\Omega_{21} = \alpha_0 Z(1 - F_0(U))\mathbb{1}\{U \geq u_0\}(D - q_0(U))/f_0(U),$$

$F_0(.)$ and $f_0(.)$ denoting respectively the cumulative distribution function and the density of U.

Theorem 3.1 establishes the root-n consistency and asymptotic normality of $\widehat{\theta}$. We prove the result by first remarking that $\widehat{\theta}$ is a two-step GMM estimator with a nonparametric first step estimator $(\widehat{q})$. We then follow Newey & McFadden (1994)'s outline for establishing asymptotic normality. Some differences arise however because $\widehat{q}$ also depends on the estimator $\widehat{\zeta}$. Theorem 3.1 also shows that the asymptotic variance of $\widehat{\theta}$ depends on the three variables Ω_{11}, Ω_{21} and $Z\xi$. The first one corresponds to the contribution of the estimators of the first and second steps. The second one arises because of the nonparametric estimation of $q_0(.)$ in $\widehat{V}_i$. The third one corresponds to the moment estimation of the linear instrumental model (3.5) in the last step.

As the proof of the theorem shows, $\widehat{\theta}$ can be linearized. Thus, by Assumptions 3.4 and 3.5, the estimator of γ_0, $\widehat{\gamma} = \widehat{\beta}_1 - \widehat{\beta}_0 + \widehat{\alpha}\widehat{\zeta}$, is also root-n consistent and asymptotically normal.

Although δ_0 and γ_0 are identified without any exclusion restriction, imposing restrictions on γ_0 may still be useful to improve the accuracy of the estimators. Suppose that, e.g., X_1 is excluded from the non-pecuniary component, so that $\gamma_{01} = 0$. In this case, we get from the second stage $\alpha_0 = \beta_{01} - \beta_{11}$. Hence, $\gamma_0 = \beta_1 - \beta_0 + \alpha_0\zeta_0$ can be estimated using only the first two steps, resulting in general in accuracy gains (see our Monte Carlo simulations and application below for evidence on this point). The third stage then boils down to estimating δ_0 only, through the instrumental linear model

$$\varepsilon - \left[DU - \int_{u_0}^{U} q_0(u)du\right] \alpha_0 = \lambda_0 + D\delta_0 + \xi, \quad E(\xi|X) = E(\xi|U) = 0, \tag{3.7}$$

where α_0 in the left hand side can now be estimated by $\widehat{\beta}_{01} - \widehat{\beta}_{11}$. One can show that the corresponding estimator is also asymptotically normal.[17]

Once δ_0 and α_0 have been estimated, we can also estimate bounds on the distribution of the *ex ante* returns Δ, namely $F_\Delta(u) = E[F_{\eta_\Delta}(u + X'(\beta_0 - \beta_1))]$. For that purpose, remark that, by (3.1),

$$P(D = 0|X) = F_{\eta_\Delta}\left(\delta_0 + X'\alpha_0\zeta_0\right).$$

Therefore, we can obtain an estimator $\widehat{F}_{\eta_\Delta}(.)$ on $[\underline{\widehat{M}}, \overline{\widehat{M}}]$, the estimated support of $\delta_0 + X'\alpha_0\zeta_0$, by regressing nonparametrically $1-D$ on the index $\widehat{\delta} + X'\widehat{\alpha}\widehat{\zeta}$. On $[\overline{\widehat{M}}, +\infty)$ (resp. $(-\infty, \underline{\widehat{M}}]$), we simply set estimate $F_{\eta_\Delta}(.)$ by $[\overline{\widehat{P}}, 1]$ (resp. $[0, \underline{\widehat{P}}]$), where $\overline{\widehat{P}}$ (resp. $\underline{\widehat{P}}$) is the supremum (resp. infimum) of $\widehat{F}_{\eta_\Delta}(.)$ on $[\underline{\widehat{M}}, \overline{\widehat{M}}]$. Finally, we can estimate $\underline{F_\Delta}(u)$ and $\overline{F_\Delta}(u)$ with the empirical analogs of (2.7) and (2.8). Bounds on the distribution of the *ex ante* returns for the treated can be estimated similarly, using (2.9).

4 Application to the decision to attend higher education

4.1 The model and data

In this section, we apply our method to estimate the relative importance of non-pecuniary factors and monetary returns to education in the decision to attend higher education in France. We consider here a generalization of the Willis & Rosen's model (1979) which accounts for the non-pecuniary consumption value of schooling, in a semi-parametric setting. After completing secondary education, individuals decide either to

[17] The proof is very close to the one of Theorem 3.1 and is available from the authors upon request.

enter directly the labor market with a high school degree ($k = 0$) or to attend higher education ($k = 1$).[18] They are supposed to make their decision $D \in \{0, 1\}$ by comparing the expected utility $\mathcal{U}_k$ of each schooling alternative k, given by

$$\mathcal{U}_k = E(Y_k^*|X, \eta_0, \eta_1) + G_k(X) = \psi_k(X) + \eta_k + G_k(X),$$

where Y_k^* and $G_k(X)$ denote respectively the stream of log-earnings and the consumption value associated with the schooling alternative k. As above, η_k is an individual productivity term known by the individual at the time of her decision but unobserved by the econometrician. Thus, the selection equation corresponds exactly to Equation (2.1).

As opposed in particular to the U.S., tuition fees are very low in most of the French higher education institutions (on average around 200 euros per year over the period of interest). This suggests that $G_0 - G_1$, which would in principle also account for the direct costs of post-secondary schooling, can be interpreted in this context as a truly non-pecuniary component, including taste for schooling and preferences for future non-wage job attributes (as those may depend on higher education attendance).

We use pooled data from the French *Generation 1992* and *Generation 1998* surveys in order to estimate our schooling choice model. These surveys collect information on individuals who left the French educational system in 1992 and 1998. They both record educational and labor market histories over the first five years following the exit from the educational system. The surveys also provide a set of individual covariates used as controls in our estimation procedure. Our subsample of interest comprises respondents having at least passed the national high school final examination. Because the labor market participation rate for this subsample is above 90% over the period of interest for both genders, we keep both males and females in our final sample. We drop individuals who only worked as temporary workers or were out of the labor force during the observation length, as we do not observe any wage for them. This finally leaves us with a sample of 24,225 individuals.[19] Working with many observations is especially important for the semiparametric estimation procedure to perform well.

Apart from a set of common regressors, including high school track, age in 6th grade, school leaving year, dummies for being born abroad (same for parents) and living in the Paris region, gender, parental profession, we include sector-specific variables, by supposing that the average local log-earnings of high school (resp. higher education)

[18] The French higher education system includes universities, which do not impose any entry selection, as well as the *Grandes Ecoles* and specialized technical colleges, which are selective.

[19] Descriptives are reported in the online appendix.

graduates affects $\psi_0(.)$ (resp. $\psi_1(.)$) alone. These variables, computed from the French Labor Force Surveys (1990-2000), are used as proxies for local labor market conditions (at the level of the French *departements*, which roughly correspond to U.S. counties) for high school and higher education graduates. Migration costs imply that labor market conditions in the places where individuals live while studying are likely to be correlated with the earnings perceived when entering the labor market.

As already mentioned, we only observe incomes during the first five years in the labor market, so that we cannot compute the discounted streams of log-earnings $Y^* = DY_1^* + (1-D)Y_0^*$. To cope with this issue, we estimate a dynamic wage model with sector-specific returns to experience. Even if we cannot recover Y^* with this model because of uncertainty on future wages, we show in the online appendix that we can identify a proxy Y satisfying $E(Y_k|X,\eta_0,\eta_1) = E(Y_k^*|X,\eta_0,\eta_1)$ (with $Y = DY_1 + (1-D)Y_0)$). The model may then be written in terms of Y_k instead of Y_k^*, and our identification strategy applies with Y instead of Y^*.

We estimate the model relying on the three-stage semiparametric procedure detailed in Section 3. Identification is secured here through the use of the average local log-earnings of high school and higher education graduates as sector-specific regressors. We use for the first step a mixture of probit (see Coppejans, 2001) with $K_1 = 3$ mixture components. The second step is performed with Newey (2009)'s series estimator, with $K_2 = 9$ approximating terms. We use for the last step the same specifications as in the Monte Carlo simulations (see the online appendix for details). Finally, to estimate he bounds on the distribution of the *ex ante* returns, we consider a kernel estimator of F_{η_Δ} with a gaussian kernel, and a bandwidth $\tilde{h}_n = 1.6\sigma(\widehat{U})n^{-1/5}$.

4.2 Results

We focus hereafter on the estimates of the non-pecuniary components and *ex ante* returns. The first step estimates of $(\zeta, \beta_0, \beta_1)$ are discussed in the online appendix. The first column of Table 1 below reports the parameter estimates relative to the non-pecuniary component G which are obtained with the unconstrained specification. The coefficients corresponding to the local average income of higher education and high school graduates are both not significant at the 10% level. This supports the idea that, as proxies for local labor market conditions, these variables have no clear reason to enter the non-pecuniary factors and should therefore only affect the probability of attendance through the *ex ante* returns. It also indicates that the data is consistent

with a constrained specification where the coefficient related to the local average income of high school graduates is set equal to zero.[20] As the estimators are more accurate when using an exclusion restriction on G, we focus on the constrained specification hereafter.

Several patterns emerge from the constrained estimates of G displayed in the second column of Table 1. The results suggest that individuals attending a general secondary schooling track (namely L for Humanities, ES for Economics and Social Sciences and S for Sciences), relative to a technical or vocational secondary schooling track, value positively higher education attendance, with the related coefficients being significant at the 1% level.[21] This pattern is consistent with the fact that the courses which are given in vocational secondary schooling tracks and, to a lesser extent, in technical tracks, are much more oriented towards the labor market than they are in general tracks. The positive effect of entering the labor market in 1998 probably reflects the enlargement of access to higher education which took place in France during the nineties. Individuals living in the Paris region also have a higher probability to attend higher education through these non-pecuniary factors, reflecting the large supply of post-secondary institutions in this area. Parental profession, in particular that of the father, has also a significant influence on the non-pecuniary determinants of the decision to attend higher education. For instance, for a given *ex ante* return to higher education, individuals whose father is employed, relative to a white collar position, as an executive, a tradesman or in an intermediate occupation have a higher propensity to enroll in higher education. This pattern suggests that part of the intergenerational transmission of human capital acts through non-pecuniary factors affecting the higher education attendance decision. Interestingly also, for a given level of expected monetary returns, males have a significantly higher probability of attending higher education (with a parameter significant at the 1% level), possibly reflecting higher educational aspirations for males than for females. Age in 6th grade, which is used as a proxy for schooling ability, also affects the attendance decision through non-pecuniary factors. Individuals who were less than 10 when entering junior high school have for instance a significantly higher probability to get some post-secondary education. These results may stem from a positive correlation between schooling ability and taste (or motivation) for schooling.

[20] We choose to impose the nullity of the coefficient associated with the local average income of high school graduates rather than the one of higher education graduates since (i) its point estimate in the unconstrained setting is much smaller and (ii) the latter coefficient is close to the 10% significance level.

[21] Recall that $G = G_0 - G_1$, so that a negative sign for a given coefficient of G implies a positive valuation of higher education compared to high school graduation.

Variable	Unconstrained	Constrained
Constant (δ_0)	-0.185 (0.174)	-0.026 (0.155)
Local average income		
Higher education graduates	-0.026 (0.017)	-0.014* (0.008)
High school graduates	0.01 (0.012)	0
Secondary schooling track		
L	-0.288*** (0.087)	-0.142*** (0.054)
ES	-0.336*** (0.097)	-0.172*** (0.058)
S	-0.349*** (0.097)	-0.175*** (0.061)
Vocational	0.62** (0.248)	0.293* (0.164)
Technical	*Ref.*	*Ref.*
Born abroad	-0.084** (0.033)	-0.031 (0.021)
Father born abroad	-0.034* (0.02)	-0.005 (0.011)
Mother born abroad	0.003 (0.014)	-0.009 (0.013)
Entering the labor market in 1998 (relative to 1992)	-0.272*** (0.084)	-0.12** (0.051)
Male	-0.062*** (0.015)	-0.038*** (0.009)
Father's profession		
Farmer	-0.029 (0.02)	-0.023 (0.017)
Tradesman	-0.053*** (0.02)	-0.025** (0.011)
Executive	-0.105*** (0.034)	-0.054** (0.022)
Intermediate occupation	-0.071*** (0.025)	-0.035*** (0.011)
Blue collar	0.000 (0.012)	-0.004 (0.008)
Other	-0.036** (0.015)	-0.023** (0.011)
White collar	*Ref.*	*Ref.*
Mother's profession		
Farmer	0.091** (0.039)	0.057 (0.037)
Tradesman	0.021 (0.019)	-0.003 (0.011)
Executive	-0.056*** (0.02)	-0.023* (0.014)
Intermediate occupation	-0.018 (0.013)	-0.019* (0.011)
Blue collar	0.076*** (0.027)	0.019* (0.01)
Other	0.012 (0.014)	-0.01 (0.007)
White collar	*Ref.*	*Ref.*
Age in 6th grade		
≤ 10	-0.103*** (0.038)	-0.047** (0.024)
11	*Ref.*	*Ref.*
≥ 12	0.108*** (0.041)	0.056** (0.026)
Paris region	-0.082*** (0.025)	-0.03** (0.012)
Vocational $\times$ ...		
Entering the labor market in 1998	0.068** (0.029)	0.034 (0.024)
Male	-0.02 (0.021)	0.003 (0.014)
Paris region	0.126*** (0.048)	0.059** (0.029)

Standard errors, presented in parentheses, were computed by bootstrap with 200 sample replicates. Significance levels: *** (1%), ** (5%) and * (10%).

Table 1: Determinants of non-pecuniary factors: parameter estimates.

Consistent with the results of the unconstrained specification, the coefficient related to the local average income of higher education graduates is small, and here only significant at the 10% level. Finally, an estimation of the non-pecuniary component of each individual in the sample reveals that for 84% of them, this component is negative. Hence, we find, in line with Carneiro et al. (2003), that there is for most of the individuals what could be referred to as a psychic gain of attending higher education.

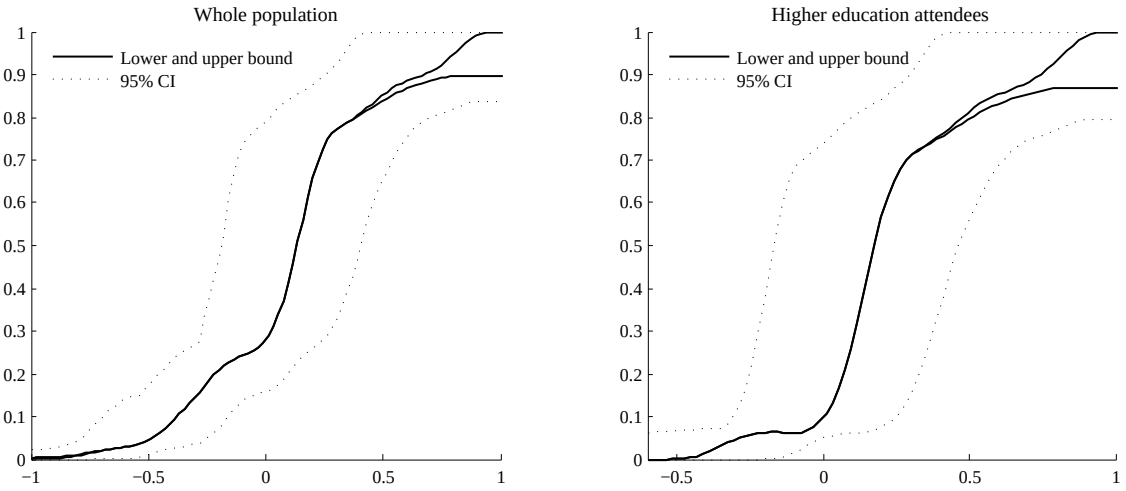

Figure 1: Distribution of the *ex ante* returns to higher education.

The estimated distributions of the *ex ante* returns to higher education are displayed in Figure 1, for the whole sample and for the subsample of higher education attendees. The streams of earnings were divided by 1,000 for scaling reasons, so that these returns must be compared to values which range from 0.7 to 2. A first striking point is that both distributions are point identified for most values. Differences between the upper and lower bounds appear only for $u \geq 0.36$, and still for these values the identifying interval remains small until $u \simeq 0.65$. The upper bound of the distribution can be used to compute a lower bound $\underline{E}$ on the average return to higher education $E(Y_1 - Y_0)$.[22] We obtain $\underline{E} \simeq 0.12$, which is quite large since it is close to one standard deviation of Y. We also observe a large heterogeneity on these returns, with a range on the *ex ante* returns $E(Y_1 - Y_0|X, \eta_0, \eta_1)$ which is similar to the one of Y. This substantial *ex ante*

[22]Indeed, an integration by part shows that

$$E(Y_1 - Y_0) = \int_{-\infty}^{\infty} [\mathbb{1}\{u \geq 0\} - F_\Delta(u)]\, du.$$

This integral can be bounded below by the corresponding integrals on $\overline{F_\Delta}$. Note that we cannot obtain a finite upper bound on $E(Y_1 - Y_0)$ here because $\lim_{u \to +\infty} \underline{\widehat{F_\Delta}}(u) < 1$.

dispersion of the returns to higher education is in line with the conclusion of Cunha & Heckman (2007, p. 887) on U.S. data.

As expected, the distribution of the *ex ante* returns is shifted towards the right for the subsample of higher education attendees, with a close to 10% probability of having a negative *ex ante* return, versus 28% for the whole sample. Hence, about 10% of the individuals attending higher education choose to do so despite a negative *ex ante* return to higher education, stressing the important role played by non-pecuniary factors in this schooling decision. Along those lines, the probability of attending higher education would fall by 11.1 percentage points (from the predicted access rate, equal to 83.1%, to the probability of having a positive *ex ante* return, 72%) if non-pecuniary factors did not exist. For comparison purposes, this decrease in higher education attendance rate is notably eight times larger than the 1.4 point decrease which is found to be associated with a 10% permanent decrease in labor market earnings of higher education attendees.

Quartile	*Ex ante* return	Non-pecuniary factors
25%	-0.069	-0.430
50%	0.133	-0.326
75%	0.267	-0.191

Table 2: Quartiles of *ex ante* returns and non-pecuniary factors.

Several other results highlight the influence of non-pecuniary factors, relative to *ex ante* monetary returns, in the decision to attend higher education. First, as shown in Table 2, the median non-pecuniary component (-0.326) is, in absolute terms, quantitatively much larger than the median *ex ante* return to higher education (0.133). Aside from their large magnitude, non-pecuniary factors also have a fairly large dispersion, with an interquartile range equal to 0.239 which is nevertheless smaller than the interquartile range for *ex ante* returns (0.336). We also compute the predicted probabilities of higher education attendance for fixed values of the non-pecuniary factors. If the non-pecuniary factors of every individual were equal to the first (resp. last) decile of the sample distribution of these factors, the attendance rate in the population would reach 95.2% (resp. 63.0%). Hence, the predicted attendance rate would fall by more than 32 points if G varied from its first to its last decile. Overall, in line with recent evidence by Carneiro et al. (2003) and Beffy et al. (2012), non-pecuniary factors appear to be a key determinant of the decision to attend higher education.

4.3 Robustness checks

4.3.1 Validity of the identification strategy

The validity of the results discussed above hinges on the exclusion restrictions between sectors. A reason why this identification strategy may not hold is that some individuals who attended higher education might face labor market conditions similar to the ones faced for those with a high school level. This might in particular be true for higher education dropouts. In order to cope with this potential concern, we run our estimates without the $3,092$ dropouts. By doing so, we focus on higher education graduation rather than attendance, in a similar spirit as in Carneiro et al. (2003). The resulting estimates of the non-pecuniary factors (see Panel 1, Table 4 in the online appendix) are very similar to previously. Secondary schooling track, gender, father's profession and year of entry into the labor market remain the main determinants of this non-pecuniary component. The distribution of the *ex ante* returns to higher education is also very similar to previously (see Figure 2 in the online appendix) and remains within the confidence intervals of that of the baseline specification. Hence, the robustness of the results to the exclusion of higher education dropouts from the sample supports our exclusion restrictions.

One might also suspect that variations across *departements* in sector-specific average incomes could be correlated with geographical variations in sector-specific labor market productivity. In an attempt to solve this issue, we add in the regressors the local proportion of individuals who graduated from high school with honours. This variable, computed from the *Panel 1989* dataset (French Ministry of Education), is used to control for differences across *departements* in productivity levels.[23] The estimates of the non-pecuniary factors as well as of the distribution of the *ex ante* returns to education (see Panel 2, Table 4 and Figure 2 in the online appendix) are robust to this alternative specification, suggesting that our estimates are likely not to be biased by this type of confounding effects.

4.3.2 Misspecification bias

As already stressed, assuming that the non-pecuniary component of the selection equation varies across individuals according to observed covariates only allows to identify the model without any exclusion restrictions nor large support condition on the covariates. However, this specification may seem too restrictive relative to a generalized

[23]The *Panel 1989* is a longitudinal dataset that follows 22,000 students entering 6th grade in 1989.

Roy model where the non-pecuniary factors also vary with unobserved characteristics. We examine this issue by computing, under some distributional assumptions, the misspecification bias on the non-pecuniary component that would arise by using our estimation procedure when the true structure of the selection equation is that of a generalized Roy model, where the non-pecuniary component writes $G(X) + U$, with U unobserved. We need to impose some further restrictions to compute the misspecification bias $B(X)$ defined by the difference between the non-pecuniary component obtained with our method (denoted here by $\widetilde{G}(X)$) and the deterministic part of the true non-pecuniary component, $G(X)$. We assume that η_Δ and U are independent and normally distributed, respectively with mean m and 0. Under these assumptions, it follows from some algebra that:

$$B(X) = -(\widetilde{G}(X) - B(X) + T(X) - m) \exp\left[\frac{(\widetilde{G}(X) - B(X) + T(X) - m)^2}{2}\rho\left(1 - \frac{1}{\sigma^2_{\eta_\Delta}}\right)\right]\rho,$$

where $\rho = \sigma^2_U/(\sigma^2_U + \sigma^2_{\eta_\Delta})$, σ^2_U and $\sigma^2_{\eta_\Delta}$ denoting respectively the variance of U and η_Δ. We estimate the bias by solving numerically this equation on the support of X, after (i) replacing $(\widetilde{G}(X), T(X))$ by their estimators obtained with our semiparametric procedure, (ii) approximating m by $E(T(X)) + \text{median}(\Delta)$ and (iii) calibrating $(\sigma^2_U, \sigma^2_{\eta_\Delta})$ from the estimates provided in Carneiro et al. (2003). (ii) and (iii) are needed since we do not identify these parameters in our setting. This leads to an average bias equal to 0.065, corresponding to 40% of the estimated standard error of δ_0. Overall, this suggests that the misspecification bias is actually negligible relative to the finite sample estimation error on the non-pecuniary component. An important implication is that we will tend to understate the dispersion of the non-pecuniary component $G(X) + U$ with our method. This actually strengthens our finding of a substantial dispersion in the non-pecuniary factors.

5 Conclusion

This paper considers the identification of an extended Roy model, with a focus on the non-pecuniary component of the selection equation. Recovering this component is key to disentangle the relative importance of monetary incentives versus preferences in the context of sorting across sectors. Our main theoretical contribution is to show that we can identify these non-pecuniary factors, and provide informative bounds on the distribution of the *ex ante* monetary returns, without any exclusion restriction nor large support condition on the covariates. We also develop a three-stage semiparametric estimation procedure leading to a root-n consistent and asymptotically normal estimator

of the non-pecuniary component. We use our approach to quantify the relative importance of non-pecuniary factors and expected returns to schooling in the decision to attend higher education in France. Consistent with the recent empirical evidence on this question, our main insight is that non-pecuniary factors are a key determinant of the attendance decision. From a policy point of view, our results suggest that a moderate increase in tuition fees, which is currently discussed to help finance the French higher education system, would only have a small detrimental effect on the higher education participation rate.

Aside from applying our results to the analysis of, e.g., public versus private sector or migration decisions, another promising avenue for further research would be to build on our approach to conduct inference on the relative importance of general vs. specific human capital through the dependence between the sector-specific unobserved productivity terms. This dependence has received much attention in competing risks models (see, e.g., Peterson, 1976, van den Berg, 1997, Abbring & van den Berg, 2003), but less so in the extensions of Roy models considered in the literature. We leave these interesting questions for further research.

References

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

Arcidiacono, P. (2004), 'Ability sorting and the returns to college major', *Journal of Econometrics* **121**, 343–375.

Ashenfelter, O. & Card, D. (1985), 'Using the longitudinal structure of earnings to estimate the effect of training programs', *Review of Economics and Statistics* **67**, 648–660.

Bayer, P. J., Khan, S. & Timmins, C. (2011), 'Nonparametric identification and estimation in a roy model with common non-pecuniary returns', *Journal of Business and Economic Statistics* **29**, 201–215.

Beffy, M., Fougère, D. & Maurel, A. (2012), 'Choosing the field of study in postsecondary education: Do expected earnings matter?', *Review of Economics and Statistics* **94**, 334–347.

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

Buera, F. (2006), Non-parametric identification and testable implications of the roy model. Working Paper, Northwestern University.

Carneiro, P., Hansen, K. & 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.

Coppejans, M. (2001), 'Estimation of the binary response model using a mixture of distributions estimator (mod)', *Journal of Econometrics* **102**, 231–269.

Cunha, F. & Heckman, J. (2007), 'Identifying and estimating the distributions of ex post and ex ante returns to schooling', *Labour Economics* **14**, 870–893.

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

D'Haultfoeuille, X. & Maurel, A. (2012), 'Another look at the identification at infinity of sample selection models', *Econometric Theory* **Forthcoming**.

Dolton, P., Makepeace, G. & van der Klaauw, W. (1989), 'Occupational choice and earnings determination: The role of sample selection and non-pecuniary factors', *Oxford Economic Papers* **41**, 573–594.

Dustmann, C. & van Soest, A. (1998), 'Public and private sector wages of male workers in germany', *European Economic Review* **42**, 1417–1441.

Florens, J., Heckman, J. J., Meghir, C. & Vytlacil, E. (2008), 'Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects', *Econometrica* **76**, 1191–1206.

Fox, J. & Gandhi, A. (2011), Using selection decisions to identify the joint distribution of outcomes. Working Paper.

French, E. & Taber, C. (2011), Identification of models of the labor market, *in* O. Ashenfelter & D. Card, eds, 'Handbook of Labor Economics', Vol. 4A, Elsevier.

Ham, J. & LaLonde, R. (1996), 'The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training', *Econometrica* **64**, 175–205.

Heckman, J. J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica* **42**, 679–693.

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

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

Heckman, J. J. & Honoré, B. (1990), 'The empirical content of the Roy model', *Econometrica* **58**, 1121–1149.

Heckman, J. J. & Sedlacek, G. (1985), 'Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market', *Journal of Political Economy* **93**, 1077–1125.

Heckman, J. J. & Sedlacek, G. (1990), 'Self-selection and the distribution of hourly wages', *Journal of Labor Economics* **8**, S329–S363.

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

Heckman, J. & Vytlacil, E. (2007), Econometric evaluation of social programs, Part II, *in* J. Heckman & E. Leamer, eds, 'Handbook of Econometrics', Vol. 6B, Elsevier.

Horowitz, J. L. (1998), *Semiparametric Methods in Econometrics*, Springer-Verlag.

Ichimura, H. (1993), 'Semiparametric least squares (SLS) and weighted SLS estimation of single-index models', *Journal of Econometrics* **58**, 71–120.

Klein, R. W. & Spady, R. H. (1993), 'An efficient semiparametric estimator for binary response models', *Econometrica* **61**, 387–421.

Lee, L. (1978), 'Unionism and relative wage rates: A simultaneous equations model with qualitative and limited dependent variables', *International Economic Review* **19**, 415–433.

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

Lee, S. & Lewbel, A. (2012), 'Nonparametric identification of accelerated failure time competing risks models', *Econometric Theory* **Forthcoming**.

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

Newey, W. K. (2009), 'Two step series estimation of sample selection models', *The Econometrics Journal* **12**, S217–S229.

Newey, W. K. & McFadden, D. (1994), Large sample estimation and hypothesis testing, *in* R. Engle & D. McFadden, eds, 'Handbook of Econometrics', Vol. 4, Elsevier.

Peterson, A. (1976), 'Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks', *Proceedings of the National Academy of Science* **73**, 11–13.

Powell, J. (1994), Estimation of semiparametric models, *in* R. Engle & D. McFadden, eds, 'Handbook of Econometrics', Vol. 4, Elsevier.

Robinson, C. & Tomes, N. (1984), 'Union wage differentials in the public and private sectors: A simultaneous equations specification', *Journal of Labor Economics* **2**, 106–127.

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

van den Berg, G. (1997), 'Association measures for durations in bivariate hazard rate models', *Journal of Econometrics* **79**, 221–245.

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