Centre de Recherche en Économie et Statistique
CREST

Série des Documents de Travail

n° 2013-17

Endogenous Attrition in Panels

L. DAVEZIES[1]
X. D'HAULTFOEUILLE[2]

October 2013

1 CREST. Email : laurent.davezies@ensae.fr
2 CREST. Email : xavier.dhaultfoeuille@ensae.fr

Endogenous Attrition in Panels*

Laurent Davezies† Xavier D'Haultfœuille‡

First version: February 2011
This version: October 2013

Abstract

We consider endogenous attrition in panels where the probability of attrition may depend on current and past outcomes. We show that this probability is nonparametrically identified provided that instruments affecting the outcomes but not directly attrition, and whose distribution is identified, are available. We thus complement Hirano et al. (2001)'s framework, which does not rely on such instruments. Contrary to their approach, neither a refreshment sample nor an additive decomposition on the probability of attrition are needed. We also show that the exclusion restriction has testable implications. We propose an efficient estimation and a test of the exclusion restriction when the outcome and instruments are discrete. The continuous case, which shares some similar features with nonparametric instrumental variable additive models, is also investigated. Finally, we apply our results to the French labor force survey, and provide evidence that attrition is related to transitions on the employment status.

Keywords: panel data, endogenous attrition, instrumental variables.

JEL classification numbers: C14, C21 and C25.

*We would like to thank Pierre Cahuc, Andres Santos and the participants of the Cemmap workshop "Recent developments in nonparametric instrumental variable methods", CREST seminar, ESEM at Oslo and the panel data conference in Montreal for their useful comments. We also thank Karine Burricand for providing us with the SILC data.

†CREST, laurent.davezies@ensae.fr

‡CREST, xavier.dhaultfoeuille@ensae.fr

1 Introduction

Panel data are very useful to distinguish between state dependence and unobserved heterogeneity (see, e.g., Heckman, 2001), to analyze the dynamics of variables such as income (see, e.g., Hall & Mishkin, 1982) or spells in duration analysis (see, e.g., Lancaster, 1990). However, these advantages may be counterbalanced by attrition, which can be especially severe when units are observed over a long period of time. Besides, attrition is often considered more problematic than standard nonresponse, because the reasons of attrition are often related to the outcomes of interest, or variations in these outcomes. Several solutions have been considered in the literature to handle this issue. A first is to suppose that attrition is exogenous, i.e. depends on lagged values that are observed by the econometrician (see, e.g., Little & Rubin, 1987). This, however, rules out a dependence between attrition and current outcomes, and is thus likely to fail in many cases. A second model takes the opposite point of view by assuming attrition to depend on contemporaneous values only (see Hausman & Wise, 1979). To handle more complex attrition patterns, Hirano et al. (2001) generalize the two previous models by allowing attrition to depend both on contemporaneous and lagged values. This generalization is made possible when a refreshment sample, i.e. a sample of new units surveyed at each period, is available. Hirano et al. (2001) also impose that the probability of attrition depends on past and current outcomes through a binary model excluding any interaction between these two variables.

In this paper, we consider still another approach, based on instruments. Contrary to Hirano et al. (2001), we do not impose any functional restrictions on the probability of attrition conditional on lagged and contemporaneous values. Refreshment sample are not needed either. On the other hand, an instrument independent of attrition conditional on past and contemporaneous outcomes is supposed to be available. A rank condition between the instrument and the contemporaneous outcome, which can be stated in terms of completeness, is also needed. Hence, the instrument is typically a lagged variable that affects the contemporaneous outcome but not directly attrition. We can use for instance past outcomes obtained from a retrospective questionnaire. We show that under a nonlinear fixed effect model, such a variable is likely to meet the nonparametric rank condition, and satisfies also the conditional independence condition if attrition only depends on transitions on the outcome.

An advantage of our method is that even if no more instruments than outcomes are available, we can test for implications of the conditional independence assumption. Another way of testing this assumption is to use refreshment samples, even though they are unneces-

sary in our setting. With such samples, the marginal distribution of the contemporaneous outcome is directly identified. We can then compare this distribution with the one obtained under our identifying restriction.

We also conduct inference under such an attrition process. In the case of discrete outcomes and instruments, the model is parametric and a straightforward constrained maximum likelihood estimation procedure is proposed. In the continuous case, the model is semiparametric and estimation is more involved. We show that our setting is closely related to the one of additive, nonparametric, instrumental variable models. Similarly to Severini & Tripathi (2012), we provide a necessary and sufficient condition for the semiparametric efficiency bound to be finite, and derive the bound in this case. We also adapt, under this condition, an estimator recently proposed by Santos (2011) for nonparametric, instrumental variable models.

Finally, we apply our results to study transitions on the French labor market, using the labor force survey of the French national institute of statistics (INSEE). This survey, which interviews people in the same housings during eighteen months, is one of the most important one of INSEE. An important issue however is that the survey does not follow individuals but housings. Thus, attrition is closely related to moving of individuals. We provide evidence that these movings are themselves related to transitions on the labor market in a way that violates the additive restriction considered by Hirano et al. (2001). With either the test described above or the refreshment sample, we do not reject the conditional independence assumption with past employment status used as an instrument. Our estimates confirm that attrition is highly related to transitions in the labor market. We show that this has important implications for the estimation of the probabilities of transition on the labor market.

The paper is organized as follows. In the second section, we study identification and testability under endogenous attrition, and compare our model with the existing literature. In the third section, we develop inference for both discrete or continuous outcomes. The fourth section is devoted to our application. Finally, the fifth section concludes. All proofs are gathered in the appendix.

2 Identification

2.1 The setting and main result

For simplicity, we consider a panel dataset with two dates $t = 1, 2$, and also suppose that there is no or ignorable nonresponse at date 1. We let $D = 1$ if the unit is observed at date 2, $D = 0$ otherwise. We let Y_t denote the outcome at t and $Y = (Y_1, Y_2)$. We also consider an instrument Z_1 whose role will be explained below, and let $Z = (Y_1, Z_1)$. For the sake of simplicity, we do not introduce covariates here, though the extension with covariates would be straightforward. We focus hereafter on the identification of either the joint distribution of (D, Y, Z) or on a parameter $\beta_0 = E(g(Y, Z))$. Our first assumption states the observational problem.

Assumption 1 *The distribution of* (D, DY_2, Z) *is identified.*

To satisfy this requirement, Z_1 can be observed at the first period, or at the second period if some information on nonrespondents is available at the second period. It also holds if Z_1 (together with Y) is observed only when $D = 1$, provided that the distribution of (Z_1, Y_1) is identified for instance through another dataset. Of course, to achieve full identification of the distribution of (D, Y, Z), restrictions are needed. If attrition directly depends on the outcome Y, the usual assumption of exogenous selection fails, and it may be difficult to find an instrument that affects the selection variable but not the outcome. On the other hand, a variable Z_1 related to Y but not directly to D may be available in this case. We thus assume the following:

Assumption 2 $D \perp\!\!\!\perp Z_1 | Y$.

This assumption is identical to the one considered by D'Haultfœuille (2010) in the case of endogenous selection. It was also considered by Chen (2001), Tang et al. (2003) and Ramalho & Smith (2013) in a nonresponse framework. Intuitively, it states that the attrition equation depends on Y_1 and Y_2 but not on Z_1. If Y_2 was endogenous (but always observed) in this equation, we could instrument it by Z_1 to identify the causal effect of Y_2 on D. Here the problem is actually slightly different: Y_2 is observed only when $D = 1$. The identification strategy is similar, however, as we use the instrument to recover the conditional distribution of attrition.

Let $P(Y) = \Pr(D = 1|Y)$. Because identification is based on inverse probability weighted moment conditions, we assume the following:

Assumption 3 *$P(Y) > 0$ almost surely.*

This assumption is similar to the common support condition in the treatment effects literature. It does not hold if D is a deterministic function of Y, as in simple truncation models where $D = \mathbb{1}\{g(Y) \geq y_0\}$, y_0 denoting a fixed threshold.

Before stating our main result, let us introduce some notations. For any random variable U and $p > 0$, let $L^p(U)$ (respectively $L^p(U|D=1)$) denote the space of functions q satisfying $E(|q(U)|^p) < +\infty$ (respectively $E(|q(U)|^p|D=1) < +\infty$). Note that $1/P \in L^1(Y|D=1)$ because $E(1/P(Y)|D=1) = 1/E(D)$. For any set $A \subset L^1(U|D=1)$, let also

$$A^{\perp} = \{q \in L^1(U|D=1) : \forall a \in A, E(|q(U)a(U)||D=1) < \infty,\ E(q(U)a(U)|D=1) = 0\}.$$

The following operator will be important for identification issues:

$$\begin{array}{rccl} T: & L^1(Y|D=1) & \to & L^1(Z|D=1) \\ & q & \mapsto & (z \mapsto E(q(Y)|D=1, Z=z)). \end{array} \tag{2.1}$$

Because Y is observed when $D=1$, T is identified. Besides, and as indicated previously, identification hinges upon dependence conditions between Y_2 and Z, which are actually related to the null space $\text{Ker}\,(T)$ of T. Let $\mathcal{F} = \{q \in L^1(Y|D=1) : q(Y) \geq 1 - 1/P(Y) \text{ a.s.}\}$ and for $f \in L^1(Y,Z)$,

$$\mathcal{F}_f = \left\{q \in L^1(Y|D=1) : q(Y) \geq 1 - 1/P(Y) \text{ a.s. and } E(|q(Y)f(Y,Z)||D=1) < \infty\right\}.$$

Finally, in the case where $g \in L^1(Y,Z)$ we denote $\beta(Y) = E[g(Y,Z)|Y]$. Our main result is the following.

Theorem 2.1 *If assumptions 1-3 hold, then:*

1. *The distribution of (D,Y,Z) is identified if and only if $Ker\,(T) \cap \mathcal{F} = \{0\}$.*

Moreover, if $g \in L^1(Y,Z)$,

2. *The set of identification of β_0 is $\{\beta_0 + E(D)E\left[\beta(Y)h(Y)|D=1\right] : h \in Ker\,(T) \cap \mathcal{F}_g\}$.*
3. *β_0 is identified if and only if $\beta(.) \in (Ker\,(T) \cap \mathcal{F}_g)^{\perp}$.*

Let us provide the intuition for the easiest result, i.e. the "if" part of the first statement. We rely on the fact that under Assumptions 2 and 3, it is sufficient to identify $P(Y)$ to recover the whole distribution of (D,Y,Z). Besides, we show that this function satisfies

$$T\left(\frac{1}{P}\right) = w, \tag{2.2}$$

where $w(Z) = 1/\Pr(D = 1|Z)$. Because T and w are identified, P is identified if there is a unique solution in $(0, 1]$ of this equation. This uniqueness can be established if $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$.

The identifying condition $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$ is related to various completeness conditions considered in the literature (see, e.g., Newey & Powell, 2003, Severini & Tripathi, 2006, Blundell et al., 2007, D'Haultfœuille, 2011, Andrews, 2011 and Hu & Shiu, 2013). Our condition is intermediate between the stronger "standard" completeness condition $\text{Ker}\,(T) = \{0\}$ and the bounded completeness condition $\text{Ker}\,(T) \cap \mathcal{B} = \{0\}$, where $\mathcal{B}$ is the set of bounded functions. When Y and Z have a finite support (respectively by $(1, ..., I)$ and $(1, ..., J)$), this assumption is satisfied if $\text{rank}(M) = I$, where M is the matrix of typical element $\Pr(Y = i|D = 1, Z = j)$ (see Newey & Powell (2003)).[1] Hence, the support of Z must be at least as rich as the one of Y ($J \geq I$) and the dependence between the two variables must be strong enough for I linearly independent conditional distributions to exist. Because the matrix M is identified, it is straightforward to test for this condition, using for instance the determinant of MM' (see Subsection 3.1 below). When Y and Z are continuous, it is far more difficult to characterize them. Conditions have been provided by Newey & Powell (2003), D'Haultfœuille (2011), Andrews (2011) and Hu & Shiu (2013). We consider below another example, related to our panel framework, where the restriction $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$ is satisfied.

The third statement of Theorem 2.1 shows that when we consider only one parameter rather than on the full distribution of (D, Y, Z), identification is achieved under weaker restrictions. To see this, note that $\mathcal{F}_g \subset \mathcal{F}$ and then $(\text{Ker}\,(T) \cap \mathcal{F})^{\perp} \subset (\text{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$. Thus, $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$ implies that $\beta(.) \in (\text{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$. On the other hand, we may have $\beta(.) \in (\text{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$ and $\text{Ker}\,(T) \cap \mathcal{F} \neq \{0\}$. This result is closely related to Lemma 2.1 of Severini & Tripathi (2012), who consider identification of linear functionals related to a nonparametric instrumental regression. Finally, the second statement of Theorem 2.1 describes the identification set of β_0 in general.

As an illustration of Theorem 2.1 with continuous outcomes, suppose that we observe at the first date a past outcome Y_0, thanks to a retrospective questionnaire. This will be the case in the application considered in Section 4. Suppose also that the outcomes satisfy the following nonlinear fixed effect model:

$$\Lambda(Y_t) = U + \varepsilon_t, \tag{2.3}$$

[1] It is not equivalent to this full rank conditions because of the inequality constraints induced by $\mathcal{F}$. One can show however that both are equivalent when $P(Y) < 1$.

where $\Lambda(.)$ is a strictly increasing real function and $(U, \varepsilon_0, \varepsilon_1, \varepsilon_2)$ are independent. Such a model generalizes standard linear fixed effect model $Y_t = U + \varepsilon_t$ and is close to the accelerated failure time model in duration analysis. Note that we do not introduce covariates here for simplicity, but our result can be extended to the more realistic model considered by Evdokimov (2011), namely $\Lambda(Y_t, X_t) = \psi(U, X_t) + \varepsilon_t$ with Λ strictly increasing in Y_t, provided that the covariates X_t are always observed at each period. We also suppose that attrition only depends on current outcomes and transitions:

$$D = g(Y_1, Y_2, \eta), \quad \eta \perp\!\!\!\perp (Y_0, Y_1, Y_2). \tag{2.4}$$

Finally, we impose the following technical restriction on U, ε_0 and ε_2. For any random variable V, we let Ψ_V denote its characteristic function.

Assumption 4 *U admits a density with respect to the Lebesgue measure, whose support is the real line. Ψ_{ε_0} vanishes only on isolated points. The distribution of ε_2 admits a continuous density f_{ε_2} with respect to the Lebesgue measure. Moreover, $f_{\varepsilon_2}(0) > 0$ and there exists $\alpha > 2$ such that $t \mapsto t^\alpha f_{\varepsilon_2}(t)$ is bounded. Lastly, Ψ_{ε_2} does not vanish and is infinitely often differentiable in $\mathbb{R}\backslash A$ for some finite set A.*

The assumption imposed on the characteristic function of ε_0 is very mild and satisfied by all standard distributions. The conditions on ε_2 are more restrictive but hold for many distributions such as the normal, the Student with degrees of freedom greater than one[2] and the stable distributions with characteristic exponent greater than one. The following proposition shows that under these conditions, the model is fully identified using Y_0 as the instrument.

Proposition 2.2 *Let $Z = (Y_0, Y_1)$, and suppose that Assumptions 3, 4, Equations (2.3) and (2.4) hold. Then Assumption 2 holds and $Ker\,(T) \cap \mathcal{F} = \{0\}$. Thus, the distribution of (D, Y, Z) is identified.*

2.2 Partial identification and testability

Apart from point identification under various completeness conditions, our attrition model displays two interesting features. First, Assumption 2 is refutable, contrary to the ignorable attrition assumption $D \perp\!\!\!\perp Y_2 | Y_1$ discussed below. Second, we can obtain bounds on

[2]See e.g. Mattner (1992) for a proof that the conditions on the characteristic function of Student distributions are indeed satisfied.

parameters of interest when the model is underidentified, i.e. when the above completeness condition fails to hold. Both are due to the fact that solutions to Equation (2.2) must lie in $[0, 1]$. These inequality constraints can be used both for testing and bounding parameters of interest.

To see this, consider the case where (Y, Z) has a finite support. If Y and Z take respectively I and J distinct values, then (2.2) can be written as a linear system of J equations with I unknown parameters and the inequality constraints:

$$\Pr(D = 0, Z = j) = \sum_{i=1}^{I} b_i \Pr(D = 1, Y = i, Z = j),\ b_i \geq 0.$$

Of course, the model is overidentified and thus testable when $I > J$, but we can also test for the inequality constraints when $I \leq J$. We derive a formal statistical test of this condition in Subsection 3.1 below. We can also partially identify parameters of interest in the underidentified case $I < J$, still using the fact that the $(b_i)_{i=1...I}$ are positive.

Finally, a stronger test of the conditional independence assumption can be derived if a refreshment sample is available, as in Hirano et al. (2001). In this case, the marginal distribution of Y_2 is identified. Then we can reject the conditional independence assumption if for all Q satisfying $T(1/Q) = w$, there exists t such that

$$E\left[\frac{D\mathbb{1}\{Y_2 \leq t\}}{Q(Y)}\right] \neq \Pr(Y_2 \leq t).$$

2.3 Comparison with the literature

We compare our approach with the most usual models of attrition.

2.3.1 Missing at random attrition

This model, which has been considered by, e.g., Rubin (1976) and Abowd et al. (1999), posits that D only depends on Y_1:

$$D \perp\!\!\!\perp Y_2 | Y_1. \tag{2.5}$$

Identification of the joint distribution of (Y_1, Y_2) follows directly from the fact that, letting f_{D,Y_1,Y_2} denote the density of (D, Y_1, Y_2) with respect to an appropriate measure,

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{f_{D,Y_1,Y_2}(1, y_1, y_2)}{\Pr(D = 1 | Y_1 = y_1)}.$$

Condition (2.5) is the equivalent, in a panel setting, of the so-called missing at random assumption (see, e.g., Little & Rubin, 1987) or the unconfoundedness assumption in the

treatment effect literature (see for instance Imbens, 2004). Because it rules out any dependence between attrition and current outcomes, it is likely to fail in many cases. In a labor force survey, for instance, house moving is a common source of attrition, and is itself related to changes in employment and/or earnings.

2.3.2 Dependence on current values

Compared to the first, the logic of this model is the opposite, as attrition is related to current values only:

$$D \perp\!\!\!\perp Y_1 | Y_2. \tag{2.6}$$

This assumption has been considered by Hausman & Wise (1979) in a parametric model. This assumption takes into account nonignorable attrition, but in a special way. It rules out in particular the possibility that transitions (namely, functions of Y_1 and Y_2) explain attrition. Abstracting from the parametric restrictions of Hausman & Wise (1979), identification can be proved along the same lines as previously. It suffices to solve in g the functional equation

$$E\left[g(Y_2)|D=1, Y_1\right] = 1/\Pr(D=1|Y_1).$$

Under completeness conditions similar to the one above, this equation admits a unique solution in g, namely $1/\Pr(D=1|Y_2=.)$.

2.3.3 Standard instrumental strategy

Attrition can be considered a particular selection problem, and thus be treated using the same tools. A classical solution (see, e.g., Heckman, 1974, Angrist et al., 1996, or Heckman & Vytlacil, 2005) is to use an instrument Z that affects attrition but not directly the current outcome:

$$Y_2 \perp\!\!\!\perp Z | Y_1.$$

Such an exclusion restriction may be credible if for some exogenous reasons, some individuals were less likely to be interviewed at the second period. However, as pointed out by Manski (2003), an important drawback of this assumption is that it is not sufficient in general to point identify the distribution of Y_2. Basically, this can be achieved only if there exists some z such that the probability of attrition $\Pr(D=1|Z=z, Y_1=y_1)$ is equal to zero, or is arbitrarily close to zero under continuity conditions. With limited variations in this probability, the distribution of Y_2 can only be set identified.

2.3.4 Additive restriction on the probability of attrition

Hirano et al. (2001) propose a two-period framework that generalize both previous examples in the sense that D may depend on both Y_1 and Y_2. This generalization is possible when a refreshment sample, which allows one to identify directly the distribution of Y_2, is available. Note that because, the distribution of Y_1 is also identified from the panel at date 1, the problem reduces to recover the copula of (Y_1, Y_2). For that purpose, Hirano et al. (2001) also suppose that

$$1/\Pr(D = 1|Y_1, Y_2) = g(k_1(Y_1) + k_2(Y_2)), \tag{2.7}$$

where g is a known function while $k_1(.)$ and $k_2(.)$ are unknown. They show that $k_1(.)$ and $k_2(.)$ are identified by the knowledge of the marginal distributions of Y_1 and Y_2. This allows them to recover the joint distribution of (Y_1, Y_2), since, by Bayes' rule,

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1,Y_2|D=1}(y_1, y_2) \Pr(D = 1) g(k_1(y_1) + k_2(y_2)).$$

Compared to our approach, Hirano et al. (2001) do not rely on any exclusion restriction. This comes at the cost of imposing the additive restriction on $\Pr(D = 1|Y_1, Y_2)$, which may be restrictive (see below), and having a refreshment sample, which is not needed in our case.

Though the identification proof of Hirano et al. (2001) is much different from ours, the two frameworks are actually related. As shown by Bhattacharya (2008), identification in this additive model can be directly obtained from the functional equations

$$E\left[g(k_1(Y_1) + k_2(Y_2))|D = 1, Y_i\right] = 1/\Pr(D = 1|Y_i).$$

Thus, identification is actually achieved along similar lines as in our case, the instrument Z being equal to (Y_1, Y_2). The difference here is that only the marginal distributions of the instrument is identified. This is the reason why they have to impose Model (2.7) to the attrition process. Such a restriction is not innocuous. If attrition depends on transitions, then their restriction is likely fails to hold. If, as in our application, attrition occurs for individuals who move, and that moving itself occurs with a large probability when employment status changes, then $\Pr(D = 1|Y_1, Y_2)$ depends on $\mathbb{1}\{Y_1 = Y_2\}$. Model (2.7) cannot handle such attrition processes.

2.3.5 Attrition with unobserved heterogeneity

Finally, Sasaki (2012) proposes a very different approach where attrition at time t depends on Y_t and a constant unobserved heterogeneity term U that also affects the dynamics of

Y_t. Such a model is attractive if individual fixed effects affect both the dynamics of Y_t and the decision to respond to the panel. He shows that the dynamics of Y_t, the attrition rule and the initial conditions (the joint distribution of Y_1 and U) are identified under, basically, four restrictions. First, Y_t should follow a Markov model of order 1, conditional on U. Second, attrition at date t should be independent of past outcomes, conditional on (U, Y_t). Third, both the law of dynamics and the attrition rule should be time invariant. Fourth, the number of periods of observations should be at least three, and a proxy of U, independent of other variables conditional on U, should be available. If such a proxy does not exist, the length of the panel should be at least six.

Even if attractive, his approach is more demanding than ours in terms of data, since it requires at least three periods. Besides, he also relies on exclusion restrictions, and contrary to our approach, it is not clear whether these restrictions are testable or not.

3 Estimation

We now turn to inference within our framework of endogenous attrition. As previously, we focus on the estimation of the distribution of (D, Y, Z), but also on the parameter $\beta_0 = E(g(Y, Z))$, which can be estimated under restrictions detailed before. We first posit an i.i.d. sample of n observations.

Assumption 5 *We observe an iid sample of size n of* (D, DY_2, Z).

We consider two cases subsequently. The first one, in line with our application, assume that the support of (Y, Z) is finite. In this setting, we derive a simple and efficient estimator and a test of the rank condition and exclusion restriction. We then turn to the continuous case, where we investigate conditions for root-n estimability of β_0, derive the semiparametric efficiency bound when it exists and propose an estimator under this condition.

3.1 The discrete case

We denote the support of Y_t and Z_1 by respectively $\{1, ..., I\}$ and $\{1, ..., J\}$, with $I \leq J$. In this case, the data (D, DY_2, Z) are distributed according to a multinomial distribution. To obtain asymptotically efficient estimators, we consider constrained maximum likelihood estimation hereafter.

For a fixed y, let $p_{1ij} = \Pr(D = 1, Y_2 = i, Z_1 = j | Y_1 = y)$ and $p_{0.j} = \Pr(D = 0, Z_1 = j | Y_1 = y)$ denote the probabilities corresponding to the observations, and define $p_1 = (p_{111}, ..., p_{1IJ})$, $p_{0.} = (p_{0.1}, ..., p_{0.J})$ and $p = (p_1, p_{0.})$. Note that we let the dependence in y implicit hereafter. p is the natural parameter of the statistical model here, as it fully describes the distribution of (D, DY_2, Z_1) conditional on Y_1. However, it does not directly allow us to recover the whole distribution of (D, Y_2, Z_1). This is why we also introduce $p_{0ij} = \Pr(D = 0, Y_2 = i, Z_1 = j | Y_1 = y)$, and $p_0 = (p_{011}, ..., p_{0IJ})$ as p_1. Then any parameter θ_0 of the distribution of (D, Y_2, Z_1) is a function of (p_0, p_1), and we write $\theta_0 = g(p_0, p_1)$. We thus consider here implicitly parameters that depend on the distribution of (D, Y_2, Z_1) conditional on Y_1. Unconditional parameters depend on all the different (p_0, p_1) corresponding to different values of Y_1, and on the marginal distribution of Y_1. We can estimate them similarly, using the empirical distribution of Y_1. Because Assumption 2 does not impose any restriction on the distribution of Y_1, such estimators are also asymptotically efficient.

Finally, we adopt the same notations for the constrained maximum likelihood estimator $\widehat{p}$ as for p. We let $n_{1ij} = \sum_{k:Y_{1k}=y} D_k \mathbb{1}\{Y_{2k} = i\}\mathbb{1}\{Z_{1k} = j\}$ and $n_{0.j} = \sum_{k:Y_{1k}=y}(1-D_k)\mathbb{1}\{Z_{1k} = j\}$. The following proposition shows how to compute $\widehat{p}$ and an efficient estimator of θ_0 in our attrition model.

Proposition 3.1 *Suppose that Assumptions 1-3 hold. Then the maximum likelihood estimator $\widehat{p}$ satisfies*

$$(\widehat{p}, \widehat{b}) = \arg \max_{(q,b)\in[0,1]^{(I+1)J}\times\mathbb{R}^I} \sum_{j=1}^{J} \left[n_{0.j} \ln q_{0.j} + \sum_{i=1}^{I} n_{1ij} \ln q_{1ij} \right]$$
$$s.t. \left| \begin{array}{ll} \sum_{j=1}^{J} \left[q_{0.j} + \sum_{i=1}^{I} q_{1ij} \right] = 1, & \\ b_i \geq 0 & i = 1, ..., I, \\ \sum_{i=1}^{I} q_{1ij} b_i = q_{0.j} & j = 1, ..., J. \end{array} \right. \quad (C)$$

Suppose moreover that the matrix P_1 of typical element p_{1ij} has rank I, $(P_1 P_1')^{-1} P_1 p_0 > 0$ (where the inequality should be understood componentwise) and g is differentiable. Then θ_0 is identifiable and $\widehat{\theta} = g(\widehat{p}_0, \widehat{p}_1)$, with $\widehat{p}_0 = (\widehat{p}_{011}, ..., \widehat{p}_{0IJ})$ and for all (i, j), $\widehat{p}_{0ij} = \widehat{b}_i \widehat{p}_{1ij}$, is asymptotically normal and efficient.

Proposition 3.1 establishes that the maximum likelihood of p can be obtained by a constrained maximization with quite simple (although nonlinear) constraints. It also shows how to compute an asymptotically efficient estimator of θ_0. The idea behind the introduc-

tion of the $(b_i)_{1\leq i\leq I}$ is that, by Bayes' rule and Assumption 2,

$$p_{0ij} = \frac{\Pr(D=0|Y_1=y,Y_2=i)}{\Pr(D=1|Y_1=y,Y_2=i)} p_{1ij},$$

and b_i represents the odds $\Pr(D=0|Y_1=y,Y_2=i)/\Pr(D=1|Y_1=y,Y_2=i)$. The inequality constraints $b_i \geq 0$ then ensure that $\Pr(D=1|Y_1=y,Y_2=i)$ is indeed a probability, while the equality constraints are a rewriting of Equation (2.2) in this discrete context (see the proof of Proposition 3.1 in the appendix).

The condition $\text{rank}(P_1) = I$ implies $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$, and is thus sufficient for the identification of θ_0 by Theorem 2.1. It can be easily tested in the data because under the null hypothesis that $\text{rank}(P_1) < I$, we have $\mu_0 \equiv \det(P_1 P_1') = 0$. Then, letting $\widehat{\mu} = \det(\widehat{P}_1 \widehat{P}_1')$, $\sqrt{n}\widehat{\mu}$ tends to a zero mean normal variable under the null by the delta method. We use this result to test for the rank condition in our application (see Section 4 below).

$\widehat{\theta}$ is asymptotically normal and efficient when $(P_1P_1')^{-1}P_1 p_0 > 0$. When $(P_1P_1')^{-1}P_1 p_0 = 0$, the true parameters lie at the boundary of the parameter space. $\widehat{\theta}$ is still a root-n consistent estimator in this case. However, it is not asymptotically normal anymore (see, e.g., Andrews, 1999, for a thorough study of such cases). Moreover, the standard bootstrap typically fails to be valid (see Andrews, 2000, for an illustration). Subsampling remains valid, on the other hand. We use it in the application when the estimator is at the boundary or close to it.

Finally, as noted before, we can test Assumption 2 by two ways. The first and standard one is that the equality constraints in (C) may not hold when $J > I$, because there is no $(b_i)_{1\leq i\leq I}$ such that $\sum_{i=1}^{I} b_i p_{1ij} = p_{0.j}$. Basically, this arises when the different values of Z are not "compatible", as with the Sargan test in linear IV models. The second is that the $(b_i)_{1\leq i\leq I}$ satisfying these equality constraints must be nonnegative. This may not hold in general, even when $I = J$. To test for both conditions simultaneously, we use the same Wald statistic as the one considered by Kodde & Palm (1986). In our framework, the unconstrained model where Assumption 2 does not necessarily hold is simply the multinomial model on (D, DY_2, Z) parameterized by p, and the maximum likelihood estimator $\widehat{p}^U$ simply corresponds to the sample proportions. The constraints (C) corresponding to Assumption 2 hold if and only if there exists $b \geq 0$ (understood componentwise) such that $P_1'b = p_{0.}$. If P_1 is full rank, the latter equation has a unique solution, the least square solution $(P_1P_1')^{-1}P_1 p_{0.}$. Therefore, if $\text{rank}(P_1) = I$, Assumption 2 is equivalent to

$$\left[P_1'(P_1P_1')^{-1}P_1 - \text{Id}\right] p_{0.} = 0, \quad (P_1P_1')^{-1}P_1 p_{0.} \geq 0, \tag{3.1}$$

where Id is the identity matrix. The idea, therefore, is to see whether $\left[P_1'(\widehat{P}_1^U P_1')^{-1}\widehat{P}_1^U - \mathrm{Id}\right]\widehat{p}_{0.}^U$ is close to zero and $(\widehat{P}_1^U P_1')^{-1}\widehat{P}_1^U\widehat{p}_{0.}^U$ is positive componentwise, where $\widehat{P}_1^U$ and $\widehat{p}_{0.}^U$ are the estimators of P_1 and $p_{0.}$ obtained from $\widehat{p}^U$.

Let us rewrite the two constraints of (3.1) as $h_1(p) = 0$ and $h_2(p) \geq 0$, and let $h(p) = (h_1(p), h_2(p))$. Let also $\mathcal{H}_0 = \{0\}^J \times \mathbb{R}^{+I}$ denote the set of $h = (h_1, h_2)$ satisfying these constraints. Denote by Σ_{ii} (resp. Σ_{12}) the asymptotic variance of $\widehat{h}_i \equiv h_i(\widehat{p}^U)$ (resp. covariance of $h_1(\widehat{p}^U)$ and $h_2(\widehat{p}^U)$), and by Σ the asymptotic variance of $\widehat{h} \equiv h(\widehat{p}^U)$. Finally, let $\widehat{\Sigma}$ denote a consistent estimator of Σ. The test statistic W_n is then defined as

$$W_n = n \min_{h \in \mathcal{H}_0} \left(h - \widehat{h}\right)' \widehat{\Sigma}^- \left(h - \widehat{h}\right), \tag{3.2}$$

where $\widehat{\Sigma}^-$ denotes the Moore-Penrose inverse of $\widehat{\Sigma}$. $\widehat{\Sigma}$ is not full rank because the rank of Σ_{11} is $J - I$, while $h_1(p) \in \mathbb{R}^J$. This is logical, since we only have $J - I$ overidentifying equality constraints here. Computing W_n is straightforward as it corresponds to a quadratic programming problem.

We now indicate how to compute critical values that are asymptotically valid under the null. We do not rely on the asymptotic result of Kodde & Palm (1986) here as they only compute the critical value corresponding to the least favorable case of the null hypothesis. Namely, they compute c such that $\sup_{h \in \mathcal{H}_0} \lim_{n\to\infty} \Pr_h(W_n \geq c) = \alpha$. This leads to a conservative test, and therefore to low power if the null hypothesis is violated. By contrast, we compute here a critical value corresponding to the most plausible DGP satisfying the null hypothesis, given the data. Therefore, our test is not conservative for a whole range of DGP satisfying the null hypothesis (see Proposition 3.2 below).

Let $(h_{10}, h_{20}) = h_0 = h(p_1, p_{0.})$ denote the true parameter. The asymptotic distribution of W_n depends on whether the components $(h_{20i})_{1 \leq i \leq I}$ are equal to zero or not. Let $\mathcal{R}_j$ be equal to $\mathbb{R}^+$ if $h_{20j} = 0$, and to $\mathbb{R}$ otherwise. Then let

$$\mathcal{H}(h_0) = \{0\}^J \times \mathcal{R}_1 \times ... \times \mathcal{R}_I.$$

We show in the proof of Proposition 3.2 below that

$$\lim_{n\to\infty} \Pr(W_n > w) = \Pr\left(\min_{h \in \mathcal{H}(h_0)} (h - U)' \Sigma^- (h - U) > w\right) \tag{3.3}$$

where $U \sim \mathcal{N}(0, \Sigma)$. To compute the level of the test based on this asymptotic distribution, we need to estimate $\mathcal{H}(h_0)$. Following, e.g., Rosen (2008) or Andrews & Soares (2010), we consider a sequence $(c_n)_{n \in \mathbb{N}}$ such that $c_n \to \infty$ and $c_n/\sqrt{n} \to 0$. We let $\widehat{\mathcal{R}}_j$ be equal to $\mathbb{R}^+$ if $\widehat{h}_{2j} \leq c_n/\sqrt{n}$, and to $\mathbb{R}$ otherwise, and

$$\widehat{\mathcal{H}(h_0)} = \{0\}^J \times \widehat{\mathcal{R}}_1 \times ... \times \widehat{\mathcal{R}}_I.$$

Finally, let $\widehat{c}_\alpha$ satisfy

$$\widehat{c}_\alpha = \inf\left\{c > 0 : \Pr\left(\min_{h \in \widehat{\mathcal{H}(h_0)}} \left(h - \widehat{U}\right)' \widehat{\Sigma}^- \left(h - \widehat{U}\right) > c\right) \leq \alpha\right\}, \tag{3.4}$$

where $\widehat{U} \sim \mathcal{N}(0, \widehat{\Sigma})$. $\widehat{c}_\alpha$ or, similarly, the p-value of the test, can be obtained easily by simulations.

Proposition 3.2 *For any $\alpha \leq 1/2$, the test defined by the critical region $\{W_n > \widehat{c}_\alpha\}$ is consistent. Its asymptotic level is α if $J > I$ or $\mathcal{R}_i = \mathbb{R}^+$ for some $i \in \{1, ..., I\}$, and 0 otherwise.*

Note that the asymptotic distribution of W_n is degenerated when $I = J$ and $\mathcal{R}_i = \mathbb{R}$ for all i, which is logical since there is no overidentifying equality constraints and the inequality constraints are not binding. In this case, the asymptotic level of the test will be 0 rather than α, as could be expected. In all other cases, the test has a non-degenerated distribution and its asymptotic level is exactly α. Following the analysis of Kodde & Palm (1986), it is also possible to express this asymptotic distribution as a mixture of chi-square. The corresponding weights, however, do not have a closed form in general, so that it is actually easier to approximate the asymptotic distribution using (3.3). We use such simulations to compute our p-values in the application below.

3.2 The continuous case

The situation is more involved when (Y, Z) is continuous, because the distribution of (Y, Z) depends on the nonparametric function $P(.)$ that is identified through an integral equation. We mostly focus on the estimation of $\beta_0 = E[g(Y, Z)]$ here. The key insight is that this problem is closely related to the estimation of linear functionals in additive, nonparametric instrumental variables (IV) models. Recall that such models satisfy

$$Y = m(X) + \varepsilon,\ E(\varepsilon|Z) = 0.$$

These models have been investigated by, among others, Newey & Powell (2003), Hall & Horowitz (2005), Santos (2011) and Severini & Tripathi (2012). m is identified through the integral equation $E(Y - m(Z)|X) = 0$. This identifying equation is similar to ours, namely $E(1 - D/P(Y)|Z) = 0$. Rather than m itself, one may be interested in linear functionals of m, $\theta_0 = E[\phi(X)m(X)]$, where ϕ is known. In our context, we also have to estimate a linear functional of $1/P$, since $\beta_0 = E[D\beta(Y)/P(Y)]$. Given these analogies, it

is not surprising that a similar methodology can be applied to our setting. An overview of the relationship between the two problems is given by Table 1.

Table 1: Analogy with additive nonparametric IV problems

	Endogeneous Attrition	Additive nonparametric IV
Observed variables	(D, DY, Z)	(Y, X, Z)
Unknown function	$1/P(.)$	$m(.)$
Exclusion restriction	$E\left(1 - \frac{D}{P(Y)}|Z\right) = 0$	$E(Y - m(X)|Z) = 0$
Operator T	$T(f) = E(f(Y)|D = 1, Z = .)$	$T(f) = E(f(X)|Z = .)$
Operator $T^\star$	$T^\star(f) = E(f(Z)|D = 1, Y = .)$	$T^\star(f) = E(f(Z)|X = .)$
Parameter of interest	$\beta_0 = E\left(\frac{Dg(Y)}{P(Y)}\right)$	$\theta_0 = E\left(\phi(X)m(X)\right)$
Root-n estimability condition: $\exists q \in L^2(Z)$ s.t.	$T^\star(q) = \beta(.)$	$T^\star(q) = \phi(.)$
Estimating equation	$\beta_0 = E(q(Z))$	$\theta_0 = E(Yq(Z))$
Estimator	$\widehat{\beta} = \widehat{E}\left(\widehat{q}(Z)\right)$	$\widehat{\theta} = \widehat{E}(Y\widehat{q}(Z))$

The first issue we investigate is the root-n estimability of β_0, that is to say, the existence of regular estimators converging at the root-n rate to β_0 (see, e.g., van der Vaart, 2000, Chapter 25). Our results are closely related to those of Severini & Tripathi (2012) in the classical IV framework. Let T^* be the adjoint operator of T, defined in (2.1):[3]

$$\begin{array}{rccl} T^*: & L^2(Z|D=1) & \to & L^2(Y|D=1) \\ & q & \mapsto & (y \mapsto E(q(Z)|D=1, Y=y)). \end{array}$$

Actually, we only need considering the restriction $T^*_{\mathcal{Y}_0}(q)$ of $T^*(q)$ on $\mathcal{Y}_0 = \text{Supp}(Y|D=0)$, which is included in $\text{Supp}(Y|D=1)$ under Assumption 3. By Assumptions 2 and 3, $E(q(Z)|D=1, Y) = E(q(Z)|D=0, Y)$ $P^{Y|D=0}$- almost surely. This allows us to extend $T^*_{\mathcal{Y}_0}(q)$ on $L^2(Z|D=0)$. By a slight abuse of notation, this extension, as well as the restriction of $\beta(.)$ on $\mathcal{Y}_0$, are also denoted by T^* and $\beta(.)$. The condition for root-n estimability is the following.

Assumption 6 *$g \in L^2(Y, Z)$ and there exists $q \in L^2(Z|D=0)$ such that $T^*(q) = \beta(.)$ and*

$$E\left[\frac{1 - P(Y)}{P(Y)}\left(q(Z) - g(Y, Z)\right)\left(q(Z) - g(Y, Z)\right)'\right] < \infty.$$

[3]We define here our operators on L^2 rather than on L^1, as in Section 2. This is not really a restriction since square integrability is required for root-n consistency in the first place.

The condition $g \in L^2(Y,Z)$ is standard to derive the asymptotic distribution of $\frac{1}{n}\sum_{i=1}^n g(Y_i,Z_i)$, even when Y is always observed. The second condition is similar to the one considered by Severini & Tripathi (2012), namely the existence of q satisfying $T^\star(q) = \phi$ in their context. If the standard completeness condition holds, then $\text{Ker}\,(T) = \{0\}$ and $\overline{\mathcal{R}\,(T^*)} = L^2(Y|D=1)$, where $\mathcal{R}\,(T^*)$ denotes the range of T^* and $\overline{A}$ denotes the closure of A. As a consequence if $g \in L^2(Y,Z)$, then $\beta(.)$ lies in $\overline{\mathcal{R}\,(T^*)}$. However, when Y is continuous, $\mathcal{R}\,(T^*)$ is not closed in general, so that even if the standard completeness holds, it may happen that $\beta(.) \notin \mathcal{R}\,(T^*)$. In such a case, the following theorem states that β_0 can not be consistently estimated at the root-n rate, as in the additive nonparametric IV problem. We also provide the semiparametric efficiency bound under Assumption 6.

Theorem 3.3 *Suppose that Assumptions 1-3 hold, and* $\beta(.) \in (Ker(T) \cap \mathcal{F}_g)^\perp$. *Then a regular root-n estimator of* β_0 *exists only if Assumption 6 holds and in this case the semi-parametric efficiency bound of* θ_0 *is:*

$$V^* = V\left(g(Y,Z)\right) + \min_{q(.)\in T^{*-1}(\{\beta(.)\})} E\left[\frac{1-P(Y)}{P(Y)}\left(q(Z)-g(Y,Z)\right)\left(q(Z)-g(Y,Z)\right)'\right]. \quad (3.5)$$

The second part of the theorem shows that the asymptotic efficiency bound comprises two terms. The first corresponds to the standard estimation of β_0 without any attrition, i.e. when $D=1$. The second accounts for attrition, and is indeed, loosely speaking, increasing with $P(Y)$. It is also related to the quality of the approximation of $g(Y,Z)$ by functions of Z. If g only depends on Z, this term disappears, which makes sense because we can estimate directly β_0 by the sample average of $g(Z)$. On the other hand, if $g(Y,Z)$ only depends on Y, the expectation on the right-hand side of (3.5) can be rewritten as

$$E\left[\frac{1-P(Y)}{P(Y)} V\left(q(Z)|Y\right)\right].$$

Hence, if there is a strong dependence between Y and Z, we may expect this second term to be small.

Turning to inference, a key observation for estimating β_0 under Assumption 6 is that

$$\beta_0 = E\left[\beta(Y)\right] = E\left[E[q(Z)|D=1,Y]\right] = E\left[E[q(Z)|Y]\right] = E[q(Z)],$$

where the third equality follows by conditional independence. Once more, a similar estimating equation arises in nonparametric IV models, since we have $\theta_0 = E[Yq(Z)]$. In a similar way as Santos (2011), the idea is to estimate q first, and then estimate β_0 by taking

the sample average of the $(\widehat{q}(Z_i))_{i=1...n}$. A difficulty is that the q satisfying Equation 6 may not be unique. Santos (2011) proposes to choose the one with the smallest norm. Adapting his idea to our context, we consider

$$\widehat{q} = \arg\min_{q \in \Theta_n} \sum_{i=1}^{n} q(Z_i)^2 \quad \text{s.t.} \quad \frac{a_n}{n} \sum_{i=1}^{n} \widehat{E}\left([\beta(Y) - q(Z)]\widehat{f}_{Y|D=1}(Y)|Y = y_i, D = 1\right)^2 \leq b_n.$$

Θ_n denotes a sieve space, i.e. a subset of $L^2(Z)$ such that $\Theta_n \subset \Theta_{n+1}$ and $q \in \overline{\cup\Theta_n}$. $\widehat{f}_X$ is a kernel estimator of f_X and $\widehat{E}[U|V = v]$ is a linear sieve estimator of $E[U|V = v]$, for any random variables (U, V).[4] The constraint of the program defines the set of functions $q \in \Theta_n$ that approximately satisfy $E(q(Z)|Y) = \beta(Y)$. Among those functions, $\widehat{q}$ is the one with the smallest norm. Santos (2011) shows that under technical conditions and with appropriate smoothing parameters, the corresponding estimator of θ_0 is root-n consistent and asymptotically normal. It is unclear, on the other hand, whether this estimator reaches the semiparametric efficiency bound.

4 Application

4.1 Introduction

In this section, we apply the previous results to estimate transitions on employment status in the French labor market. Beyond the unemployment rate, measuring such transitions is important to assess, for instance, the importance of short and long-term unemployment. We use for that purpose the Labor Force Survey (LFS) conducted by the French national institute of statistics (INSEE). This survey is probably the best tool to measure such transitions in France. Compared to administrative data or other surveys, it properly measures unemployment with respect to the standard ILO definition, has a comprehensive coverage of the population and has a large sample size. Since 2003, the French LFS is a rotating panel with approximately 5,900 new households each quarter. Each household is interviewed during six waves. On the first and sixth wave, interviews are face to face, while on the others they are conducted by telephone. It has been argued that the use of phone may introduce specific measurement errors (see, e.g., Biemer, 2001), so we focus on the first and last interrogations hereafter. We also restrict ourselves to people between 15 and 65 and pool together all labor force surveys on the period 2003-2005.

[4] Here, we have supposed that $\beta(Y)$ is known, which is the case in the common situation where $g(Y, Z)$ does not depend on Z. Otherwise, $\beta(Y)$ should also be estimated with a nonparametric estimator of $E(g(Y, Z)|D = 1, Y)$.

Table 2: Summary statistics on the French LFS.

Statistics	All	Men	Women
Main sample:			
Number of individuals	107,031	52,245	54,786
Attrition rate on last waves	21.78%	22.26%	21.31%
Participation rate on first waves	68.17%	73.91%	62.69%
(Uncorrected) participation rate on last waves	67.38%	72.75%	62.32%
Unemployment rate on first waves	9.68%	9.05%	10.39%
(Uncorrected) unemployment rate on last waves	8.02%	7.22%	8.90%
Refreshment sample for last waves:			
Number of observations	109,404	53,337	56,067
Participation rate on the refreshment sample	67.92%	73.31%	62.78%
Unemployment rate on the refreshment sample	9.97%	9.43%	10.57%

Sources: French LFS, first waves between 2003 and 2005, individuals between 15 and 65 year old.

Table 2 provides some summary statistics on our dataset, which emphasize that attrition may be problematic in the LFS survey. This is especially striking when we compare the (uncorrected) participation and unemployment rate on last waves and the one on the refreshment sample, which corresponds to entrants interviewed at the same time. We observe differences around 1.5 percent points on participation rates, and around 2 percent points on unemployment rates. To understand these differences, recall that in the French LFS, moving households are not followed by interviewers, who stick instead on housings which were selected in the first waves. This is likely to affect activity rates and transition estimates on the labor market, because transitions are very different for moving and non-moving households.

This latter fact can be illustrated using the French sample of the European Survey on Income Living Conditions (SILC). Contrary to the LFS, this panel follows individuals even if they move. It is therefore possible to estimate the difference in the transition matrix for those who have moved and the others. Note, on the other hand, that it is difficult to use its results as a benchmark, for several reasons. First, and most importantly, the status on the labor market is not obtained with the same questions as in the LFS, and it is well-known that this matters much for defining in particular unemployment (for evidence

on this issue in France, see, e.g., Guillemot, 1996, and Gonzalez-Demichel & Nauze-Fichet, 2003). Second, still around 40% of the individuals in the French sample of SILC that move from one year to another are lost, so the bias stemming from such nonrespondents may still be substantial. That said, Table 3 shows that the difference in the transitions on the labor market between individuals who have moved and the others are substantial. In particular, the diagonal of the transition matrix is much smaller for individuals who move. The difference reaches around 30 percentage points for inactive people. This suggests that the MAR and HIRR methods may overestimate the diagonal of the transition matrix.

Table 3: Comparison moving and non moving people in SILC

	Non moving			Moving		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV-MAR						
Y_1 = Empl.	92.86	3.50	3.64	90.53	3.99	5.48
	(0.34)	(0.27)	(0.23)	(1.34)	(0.95)	(0.99)
Y_1 = Unempl.	30.48	51.31	18.21	38.32	41.55	20.13
	(1.84)	(2.05)	(1.60)	(6.33)	(6.60)	(0.52)
Y_1 = Out L.F.	7.40	5.83	86.77	34.64	8.05	57.30
	(0.48)	(0.45)	(0.64)	(3.67)	(2.13)	(3.77)

Sources: French sample of SILC 2004/2005, individuals between 15 and 65.

Notes: standard error in parentheses.

As suggested in Section 2, we propose to correct for potentially endogenous attrition by using past employment status, measured by a retrospective question asked on the first waves. The underlying assumption is that attrition depends on the current transition on this outcome, but not on previous ones. This assumption is plausible if most of the endogeneity in attrition stems from the moving of households. The instrument Z we use is employment status six months before the first wave. We choose to divide this variable in three categories (unemployed, employed, and out of labour force), in the same way as our outcome, which is contemporary employment status.

4.2 The results

We first check the rank condition between Z_1 and Y_2 conditional on gender and Y_1, relying on the determinant test proposed in Subsection 3.1. Results are displayed in Table 4. The p-value of the rank test associated to any state Y_1 are always smaller than 10% for

both men and women. We also implement the test developed in the Proposition 3.2, using $c_n = \ln(n)$. Though some inequality constraints are binding with $Y_1 =$ Unempl., we do not reject the independence assumption $Z \perp\!\!\!\perp D|Y_1, Y_2$ here, the p-value being close to 0.50. The p-values equal to one that we obtain correspond to situations where the inequality constraints are not binding. In such a case, $W_n = 0$ and we accept the null hypothesis at any level.

Table 4: Rank test between Z and Y_2 conditional on gender and Y_1.

	P-value (Men)	P-value (Women)
$Y_1 =$ Empl.	0.004	0.001
$Y_1 =$ Unempl.	0.077	0.057
$Y_1 =$ Out L.F.	0.059	0.091

Sources: French LFS (2003-2005).

Notes: the p-values are obtained by bootstrap with 1,000 bootstrap samples.

Table 5: Test of $Z \perp\!\!\!\perp D|Y_1, Y_2$ by gender.

	P-value (Men)	P-value (Women)
$Y_1 =$ Empl.	1	1
$Y_1 =$ Unempl.	0.491	0.488
$Y_1 =$ Out L.F.	1	1

Sources: French LFS (2003-2005).

Notes: we use the test based on W_n and $\widehat{c}_\alpha$ defined in (3.2) and (3.4).

Second, we estimate the probabilities of attrition (or non-attrition) conditional on (Y_1, Y_2). Our results, displayed in Table 6, confirm that attrition is related to transitions on employment status. People who remain stable on the labor market have always a significant larger probability to respond in the second wave than people who change. In particular, we observe a large attrition for those who move from employment to unemployment or

inactivity whereas attrition seems negligible for those who remain unemployed at both periods. As suggested above, such transitions are likely to be related to house movings. For instance, transitions from inactivity to employment or unemployment mostly correspond to students who enter the labor market and move at the same time. Such features cannot be captured under the missing at random (MAR) scheme $D \perp\!\!\!\perp Y_2|Y_1$, or the additive model of Hirano et al. (2001). In particular, they tend to underestimate the probability of attrition for people whose status change on the labor market, and to overestimate them for stable trajectories (see Table 7 for the tests on the difference between our IV models and the two others). Note also that we estimate the probability of attrition to be zero for people who remain unemployed. This indicates that for those people, the inequality constraint $b_i \geq 0$ is binding. This could suggest that the exclusion restriction is violated. However, the test conducted previously shows that the unconstrained estimator, if negative, is actually close to zero, and we cannot reject at standard levels that the true value is actually positive. That $\widehat{b}_i = 0$ for individuals initially unemployed also indicates that the true value of b_i may be equal to zero, in which case the estimator is not asymptotically normal. We therefore use subsampling rather than the bootstrap or the normal approximation for inference on this subpopulation.

Table 6: Estimation of $P(D = 1|Y_1, Y_2)$ by gender under various assumptions.

	Men			Women		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV						
Y_1 = Empl.	84.33 (0.83)	34.33 (3.93)	46.31 (7.11)	85.72 (0.92)	46.45 (6.17)	44.57 (4.42)
Y_1 = Unempl.	55.56 (4.73)	100 (4.6)	51.01 (7.13)	52.46 (4.88)	100 (4.45)	76.38 (6.15)
Y_1 = Out L.F.	54.83 (10.91)	55.85 (11.15)	85.72 (2.01)	56.43 (11.88)	67.1 (17.22)	84.34 (1.11)
MAR						
Y_1 = Empl.	78.22 (0.22)	78.22 (0.22)	78.22 (0.22)	79.00 (0.22)	79.00 (0.22)	79.00 (0.22)
Y_1 = Empl.	65.9 (0.81)	65.9 (0.81)	65.9 (0.81)	69.77 (0.78)	69.77 (0.78)	69.77 (0.78)
Y_1 = Empl.	79.52 (0.35)	79.52 (0.35)	79.52 (0.35)	79.77 (0.28)	79.77 (0.28)	79.77 (0.28)
HIRR						
Y_1 = Empl.	79.01 (0.29)	59.98 (2.13)	76.44 (2.17)	79.57 (0.3)	66.59 (2.18)	77.57 (2.01)
Y_1 = Empl.	75.84 (1.4)	55.55 (1.25)	73.01 (2.06)	76.04 (1.45)	61.89 (1.37)	73.81 (1.75)
Y_1 = Empl.	82.41 (1.68)	65.09 (2.33)	80.15 (0.45)	82.01 (1.68)	69.99 (2.01)	80.18 (0.37)

Sources: French LFS (2003-2005).

Notes: the standard errors, in parentheses, are computed with the bootstrap except for Y_1 = Unempl. in the IV case, where we use subsampling.

Table 7: Comparison between our method and other ones on $\widehat{P}(D = 1|Y_1, Y_2)$

	Men			Women		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV-MAR						
Y_1 = Empl.	6.11 (<0.001)	−43.89 (<0.001)	−31.91 (<0.001)	6.72 (<0.001)	−32.55 (<0.001)	−34.43 (<0.001)
Y_1 = Unempl.	−10.34 (0.018)	34.1 (<0.001)	−14.9 (0.009)	−17.31 (<0.001)	30.23 (<0.001)	6.61 (0.322)
Y_1 = Out L.F.	−24.69 (0.024)	−23.67 (0.034)	6.2 (0.002)	−23.33 (0.049)	−12.66 (0.462)	4.58 (<0.001)
IV-HIRR						
Y_1 = Empl.	5.32 (<0.001)	−25.64 (<0.001)	−30.13 (<0.001)	6.15 (<0.001)	−20.14 (0.002)	−33 (<0.001)
Y_1 = Unempl.	−20.28 (<0.001)	44.45 (<0.001)	−22.01 (<0.001)	−23.58 (<0.001)	38.11 (<0.001)	2.57 (0.862)
Y_1 = Out L.F.	−27.58 (0.012)	−9.24 (0.415)	5.57 (0.005)	−25.57 (0.033)	−2.89 (0.867)	4.16 (<0.001)

Sources: French LFS (2003-2005).

Notes: the p-values, in parentheses, are computed with the bootstrap (Y_1 = Empl. and Out L.F.) and subsampling (Y_1 = Unempl.).

Before presenting our results on transitions, we estimate the distribution of Y_2 with our IV method and compare it with the one of the refreshment sample. We also estimate this distribution supposing that data are missing at random (MAR), i.e. $D \perp\!\!\!\perp Y_2|Y_1$. Table 8 shows that on the five statistics related to the distribution of Y_2, our estimator is close, and not statistically significant at usual levels, to the one based on the refreshment sample. Those based on the MAR assumptions, on the other hand, do differ significantly for several features of Y_2. In other words, we can reject, using the refreshment sample, the hypothesis that attrition only depends on past outcomes, while our independence condition is not rejected in the data. Note that we cannot use the refreshment sample to properly compare our method with the one of Hirano et al. (2001) because by construction, their estimator exactly matches the distribution of Y_2 on the refreshment sample.

Table 8: Comparison of the methods with the refreshment sample

	Men			Women		
	REF.	MAR	IV	REF.	MAR	IV
$P(Y_2 = \text{Empl.})$	66.4	67.47 (<0.0001)	64.59 (0.055)	56.15	56.81 (0.0054)	55.07 (0.159)
$P(Y_2 = \text{Unempl.})$	6.92	5.62 (<0.0001)	7.53 (0.235)	6.63	5.78 (<0.0001)	6.51 (0.801)
$P(Y_2 = \text{Out L.F.})$	26.69	26.92 (0.2641)	27.88 (0.159)	37.22	37.4 (0.4243)	38.42 (0.127)
Participation rate	73.31	73.08 (0.2641)	72.12 (0.159)	62.78	62.6 (0.4243)	61.58 (0.127)
Unemployment rate	9.43	7.68 (<0.0001)	10.44 (0.146)	10.57	9.24 (<0.0001)	10.58 (0.982)

Sources: French LFS (2003-2005).

Notes: the p-values of the difference with the refreshment sample, in parentheses, are obtained using the bootstrap (MAR) and subsampling (IV).

Finally, we compute transitions on the labor market using our IV method, the MAR assumption and the additive method of Hirano et al. (2001) (see Table 9). Not surprisingly given the discrepancies on the probabilities of attrition, our results differ significantly from those obtained by the other methods. Other methods lead in particular to a higher stability on the labor market. This is not surprising, given the assumptions underlying these methods. Table 3 suggests that there is a specific effect of being in the diagonal on the transition matrix on attrition, but neither the MAR assumption nor the additivity condition of Hirano et al. (2001) can incorporate such effects. The final results suggest that this could lead to important biases on the estimation of transitions.

Table 9: Estimated probability of transitions by gender under various assumptions

	Men			Women		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV						
Y_1 = Empl.	85.86 (0.83)	6.12 (0.69)	8.02 (1.11)	83.46 (0.9)	4.73 (0.62)	11.81 (1.15)
Y_1 = Unempl.	49 (2.97)	25.85 (2.33)	25.15 (2.86)	52.61 (2.9)	25.3 (2.25)	22.08 (2.6)
Y_1 = Out L.F.	13.81 (2.59)	6.48 (1.15)	79.72 (1.81)	12.74 (2.24)	5.92 (1.49)	81.34 (1.07)
MAR						
Y_1 = Empl.	92.56 (0.16)	2.69 (0.1)	4.75 (0.13)	90.56 (0.18)	2.78 (0.1)	6.66 (0.16)
Y_1 = Empl.	41.31 (1.03)	39.23 (0.97)	19.46 (0.82)	39.56 (0.99)	36.27 (0.94)	24.18 (0.86)
Y_1 = Empl.	9.52 (0.28)	4.55 (0.21)	85.93 (0.34)	9.02 (0.22)	4.98 (0.17)	86 (0.27)
HIRR						
Y_1 = Empl.	91.64 (0.2)	3.5 (0.12)	4.86 (0.14)	89.92 (0.22)	3.3 (0.12)	6.79 (0.18)
Y_1 = Empl.	35.9 (0.93)	46.54 (0.92)	17.57 (0.79)	36.29 (0.94)	40.87 (0.88)	22.85 (0.84)
Y_1 = Empl.	9.19 (0.28)	5.56 (0.23)	85.26 (0.36)	8.77 (0.22)	5.68 (0.17)	85.55 (0.29)

Sources: French LFS (2003-2005).

Notes: the standard errors, in parentheses, are computed with the bootstrap except for Y_1 = Unempl. in the IV case, where we use subsampling.

5 Conclusion

In this paper, we develop an alternative method to correct for endogenous attrition in panel. We allow for both dependence on current and past outcomes and, thanks to the availability of an instrument, do not need to impose functional restrictions on the probability of attrition. The application suggests that our method may do a good job for handling attrition processes which mostly depend on transitions.

The paper raises challenging issues, related to our main conditional independence assumption. The first is whether the refreshment sample could be used to weaken this assumption,

rather than to test for it. This may be useful in settings where this condition is considered too stringent. The second is whether one can build bounds on parameters of interest if the conditional independence assumption is replaced by weaker conditions such as monotonicity ones. Finally, an issue that also arises for nonparametric additive IV models would be to obtain efficient estimators for linear functionals under Assumption 6, and consistent estimators without such an assumption.

References

Abowd, J. M., Crépon, B. & Kramarz, F. (1999), 'Moment estimation with attrition: An application to economic models', *Journal of the American Statistical Association* **96**, 1223–1231.

Andrews, D. K. (1999), 'Estimation when a parameter is on a boundary', *Econometrica* **67**, 1341–1383.

Andrews, D. K. (2000), 'Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space', *Econometrica* **68**, 399–405.

Andrews, D. K. (2011), Examples of l2-complete and boundedly-complete distributions. Working Paper.

Andrews, D. K. & Soares, G. (2010), 'Inference for parameters defined by moment inequalities using generalized moment selection', *Econometrica* **78**, 119–157.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**, 444–455.

Bhattacharya, D. (2008), 'Inference in panel data models under attrition caused by unobservables', *Journal of Econometrics* **144**, 430–446.

Biemer, P. P. (2001), 'Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing', *Journal of Official Statistics* **17**, 295–320.

Bierens, H. J. (1982), 'Consistent model specification tests', *Journal of Econometrics* **20**, 105–134.

Blundell, R., Chen, X. & Kristensen, D. (2007), 'Nonparametric iv estimation of shape-invariant engel curves', *Econometrica* **75**, 1613–1669.

Carter, M. (2001), *Foundations of Mathematical Economics*, MIT Press.

Chen, C. (2001), 'Parametric models for response-biased sampling', *Journal of the Royal Statistical Society, Series B* **63**, 775–789.

D'Haultfœuille, X. (2010), 'A new instrumental method for dealing with endogenous selection', *Journal of Econometrics* **154**, 1–15.

D'Haultfœuille, X. (2011), 'On the completeness condition in nonparametric instrumental regression', *Econometric Theory* **27**, 460–471.

Evdokimov, K. (2011), Nonparametric identification of a nonlinear panel model with application to duration analysis with multiple spells. Working paper.

Gonzalez-Demichel, C. & Nauze-Fichet, E. (2003), 'Les contours de la population active : aux frontières de l'emploi, du chômage et de l'inactivité', *Economie et Statistique* **362**, 85–103.

Guillemot, D. (1996), 'La population active : une catégorie statistique difficile à cerner', *Economie et Statistique* **300**, 39–53.

Hall, P. & Horowitz, J. L. (2005), 'Nonparametric methods for inference in the presence of instrumental variables', *Annals of Statistics* **33**, 2904–2929.

Hall, R. & Mishkin, F. S. (1982), 'The sensitivity of consumption to transitory income: estimates from panel data on households', *Econometrica* **50**, 461–481.

Hausman, J. A. & Wise, D. A. (1979), 'Attrition bias in experimental and panel data: the Gary income maintenance experiment', *Econometrica* **47**, 455–473.

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

Heckman, J. J. (2001), 'Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture', *Journal of Political Economy* **109**, 673–748.

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

Hirano, K., Imbens, G. W., Ridder, G. & Rubin, D. B. (2001), 'Combining panel data sets with attrition and refreshment samples', *Econometrica* **69**, 1645–1659.

Hu, Y. & Shiu, J. (2013), Nonparametric identification using instrumental variables: sufficient conditions for completeness. Working Paper.

Imbens, G. (2004), 'Nonparametric estimation of average treatment effects under exogeneity: a review', *The Review of Economics and Statistics* **86**, 4–29.

Kodde, D. A. & Palm, F. C. (1986), 'Wald criteria for jointly testing equality and inequality restrictions', *Econometrica* **54**, 1243–1248.

Lancaster, T. (1990), *The Econometric Analysis of Transition Data*, Cambridge University Press.

Little, R. & Rubin, D. B. (1987), *Statistical analysis with Missing Data*, John Wiley & Sons, New York.

Manski, C. F. (2003), *Partial Identification of Probability Distribution*, Springer.

Mattner, L. (1992), 'Completeness of location families, translated moments, and uniqueness of charges', *Probability Theory and Related Fields* **92**, 137–149.

Newey, W. & Powell, J. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**, 1565–1578.

Ramalho, E. A. & Smith, R. J. (2013), 'Discrete choice nonresponse', *Review of Economic Studies* **80**, 343–364.

Rosen, A. (2008), 'Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities', *Journal of Econometrics* **146**, 107–117.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592.

Santos, A. (2011), 'Instrumental variables methods for recovering continuous linear functionals', *Journal of Econometrics* **161**, 129–146.

Sasaki, Y. (2012), Heterogeneity and selection in dynamic panel data. Working paper.

Severini, T. & Tripathi, G. (2006), 'Some identification issues in nonparametric linear models with endogenous regressors', *Econometric Theory* **22**, 258–278.

Severini, T. & Tripathi, G. (2012), 'Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors', *Journal of Econometrics* **170**, 491–498.

Stewart, G. W. (1969), 'On the continuity of the generalized inverse', *SIAM Journal on Applied Mathematics* **17**, pp. 33–45.

Tang, G., Little, R. J. A. & Raghunathan, T. E. (2003), 'Analysis of multivariate missing data with nonignorable nonresponse', *Biometrika* **90**, 747–764.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge University Press.

A Appendix: proofs

Theorem 2.1

The distribution of (Y, Z, D) is identified if and only if the distribution of $Y|Z, D = 0$ is identified. We have:

$$\begin{aligned} f_{Y|Z=z,D=0}(y) &= \frac{f_{D,Y,Z}(0,y,z)}{f_{D,Z}(0,z)} \\ &= \frac{1}{f_{D,Z}(0,z)} P(D=0|Y=y,Z=z) f_{Y,Z}(y,z) \\ &= \frac{1}{f_{D,Z}(0,z)} \frac{P(D=0|Y=y,Z=z)}{P(D=1|Y=y,Z=z)} f_{Y,Z|D=1}(y,z) \\ &= \frac{1}{f_{D,Z}(0,z)} \frac{1-P(y)}{P(y)} f_{Y,Z|D=1}(y,z). \end{aligned}$$

Then we deduce that $Y|Z, D = 0$ is identified if and only if P is identified. Under assumptions 2 and 3 (i), the function P is such that $T(1/P) = w$ and $1/P \geq 1$. Reciprocally, let Q a function such that $1/Q \in L^1(Y|D = 1)$, $T(1/Q) = w$ and $1/Q \geq 1$. If the unobserved distribution of $Y|Z, D = 0$ is such that

$$f_{Y|Z=z,D=0}(y) = \frac{1}{f_{D,Z}(0,z)} \frac{1-Q(y)}{Q(y)} f_{Y,Z|D=1}(y,z),$$

we have $P(D = 1|Y) = Q(Y)$ and $D \perp\!\!\!\perp Z|Y$. So the set of identification of P is

$$\{Q : 1/Q \in L_1(Y|D = 1), T(1/Q) = w, 1/Q \geq 1\},$$

which is reduced to a point if and only if $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$. This proves the first point of Theorem 2.1.

For the second and the third points, let Q be such that $T(1/Q) = w$, $1/Q \geq 1$ and $E(|g(Y, Z)|/Q(Y) \ |D = 1) < \infty$. Choosing $f_{Y|Z,D=0}$ as above, we can rationalize that $P(D = 1|Y) = Q(Y)$, $D \perp\!\!\!\perp Z|Y$ and $g \in L_1(Y, Z)$. So the set of identification of $1/P$ is

$$\left\{1/Q : 1/Q \in L^1(Y|D = 1), T(1/Q) = w, 1/Q \geq 1, E(|g(Y, Z)|/Q(Y)|D = 1) < \infty\right\}$$

or, equivalently,

$$\{1/P + h : h \in \text{Ker}\,(T) \cap \mathcal{F}, E(|gh||D = 1) < \infty\} = 1/P + \text{Ker}\,(T) \cap \mathcal{F}_g.$$

For all $h \in \mathcal{F}_g$ the quantities $E(|g(Y, Z)h(Y)||D = 1)$, $E(g(Y, Z)h(Y)|D = 1)$, $E(|\beta(Y)h(Y)||D = 1)$, $E(\beta(Y)h(Y)|D = 1)$ are well defined and finite. Then the set identification of β_0 is

$$\{\beta_0 + E(\beta(Y)h(Y)|D = 1)E(D) : h \in \text{Ker}\,(T) \cap \mathcal{F}_g\},$$

which reduces to a point if and only if $E(\beta(Y)h(Y)|D=1)=0$ for every $h \in \text{Ker}\,(T) \cap \mathcal{F}_g$. Hence, β_0 is identified if and only if $\beta(.) \in (\text{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$ □

Proposition 2.2

First, remark that if $\eta \perp\!\!\!\perp (Y_0, Y_1, Y_2)$, then $\eta \perp\!\!\!\perp Y_0|Y_1, Y_2$. As a result, $D \perp\!\!\!\perp Y_0|Y$ and Assumption 2 holds with $Z_1 = Y_0$. Now, suppose that $T(h) = 0$ for $h \in \mathcal{F}$, and let us prove that $h = 0$. First, $T(h) = 0$ rewrites as

$$0 = E(Dh(Y_1, Y_2)|Y_0, Y_1) = E(\widetilde{h}(Y_1, \widetilde{Y}_2)|Y_0, Y_1),$$

with $\widetilde{Y}_t = \Lambda(Y_t)$ and $\widetilde{h}(y_1, y_2) = h(y_1, \Lambda(y_2)) \times P(y_1, y_2)$. As a result, for all $t \in \mathbb{R}$,

$$E(\widetilde{h}(Y_1, \widetilde{Y}_2)e^{it\widetilde{Y}_0}|Y_1) = 0$$

Because $\varepsilon_0 \perp\!\!\!\perp (U, Y_1, Y_2)$,

$$E(\widetilde{h}(Y_1, \widetilde{Y}_2)e^{itU}|Y_1)\Psi_{\varepsilon_0}(t) = 0.$$

Thus, by assumption, $t \mapsto E(\widetilde{h}(Y_1, \widetilde{Y}_2)e^{itU}|Y_1)$ is equal to zero except perhaps on a set of isolated points. Because this function is continuous by dominated convergence, it is actually equal to zero on the whole line. This implies (see e. g. Bierens, 1982, Theorem 1)

$$E(\widetilde{h}(Y_1, \widetilde{Y}_2)|Y_1, U) = 0.$$

Now, ε_2 is independent of (Y_1, U) and U admits a density with respect to the Lebesgue measure. Thus, for almost all y_1 and almost every u,

$$\int \widetilde{h}(y_1, u - v) f_{-\varepsilon_2}(v) dv = 0. \tag{A.1}$$

Fix y_1 so that Equation (A.1) holds for almost every u. Because $h \geq 1 - 1/P$ by assumption, $\widetilde{h}$ is bounded below by -1. Letting $g = \widetilde{h}(y_1, .)$ and $\star$ denote the convolution product, we have $(g + 1) \star f_{-\varepsilon_2} = 1$ almost everywhere. Besides, by the first step of the proof of Proposition 2.2 of D'Haultfœuille (2010), there exist positive c_1, c_2 and $0 < \alpha' < \alpha - 2$ such that

$$c_1 \leq (f_{\varepsilon_2} \star f_{\alpha'})(x) \times (1 + |x|)^{\alpha'+1} \leq c_2, \tag{A.2}$$

where $f_{\alpha'}$ denotes the density of an α'-stable distribution of characteristic function $\exp(-|t|^{\alpha'})$. Moreover, because $g + 1, f_{-\varepsilon_2}$ and $f_{\alpha'}$ are nonnegative, we can apply Fubini's theorem, so that

$$(g + 1) \star (f_{-\varepsilon_2} \star f_{\alpha'}) = ((g + 1) \star f_{-\varepsilon_2}) \star f_{\alpha'} = 1 \star f_{\alpha'} = 1.$$

Thus, $g \star \phi = 0$, with $\phi = f_{-\varepsilon_2} \star f_{\alpha'}$. In other words, we have a similar result as (A.5) in the proof of D'Haultfœuille (2010). Applying the third step of this proof shows that the location family generated by ϕ is complete. Thus $g = 0$ almost everywhere. Because this reasoning holds for almost all y_1, $h(Y_1, Y_2) = 0$ almost surely. Hence, $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$, and the result follows by Theorem 2.1 □

Proposition 3.1

For simplicity, we keep hereafter the conditioning on $Y_1 = y$ implicit. We first prove that $D \perp\!\!\!\perp Z_1 | Y_2$ is equivalent to the existence of $b_i \geq 0$, for $i = 1, ..., I$ such that $\forall (i,j), p_{0ij} = b_i p_{1ij}$. For $i \in \{1, ...I\}$, let $A(i)$ the set of j such that $\Pr(Y_2 = i, Z_1 = j) > 0$. Suppose first that $Z_1 \perp\!\!\!\perp D | Y_2$. Then for all i and all $j \in A(i)$,

$$\left[1 - \Pr\left(D = 1 | Y_2 = i, Z_1 = j\right)\right] / \Pr\left(D = 1 | Y_2 = i, Z_1 = j\right)$$

does not depend on j. Thus, there exists $b_i \geq 0$ such that for all $j \in A(i)$,

$$\Pr\left(D = 0 | Y_2 = i, Z_1 = j\right) = b_i \Pr\left(D = 1 | Y_2 = i, Z_1 = j\right).$$

Multiplying both sides by $\Pr(Y_2 = i, Z_1 = j)$ and remarking that both sides are equal to 0 when $j \notin A(i)$, we get, for all j, $p_{0ij} = b_i p_{1ij}$. This proves the "only if" part.

Conversely, suppose that there exists $b_i \geq 0$ such that $p_{0ij} = b_i p_{1ij}$. Because $\sum_j p_{1ij} = P(D = 1, Y_2 = i) > 0$ by Assumption 3, $p_{0ij} = b_i p_{1ij}$ implies that

$$p_{0ij} = \frac{\sum_j p_{0ij}}{\sum_j p_{1ij}} p_{1ij} = \frac{\Pr(D = 0, Y_2 = i)}{\Pr(D = 1, Y_2 = i)} p_{1ij}.$$

In other words, $P(Z_1 = j | D = 0, Y_2 = i) = P(Z_1 = j | D = 1, Y_2 = i)$ for all j, implying that $Z_1 \perp\!\!\!\perp D | Y_2 = i$.

We deduce that the distribution of (D, DY_2, Z_1) is compatible with $D \perp\!\!\!\perp Z_1 | Y_2$ if and only if

$$\forall (i,j), \exists p_{0ij} \geq 0, \exists b_i \geq 0 : \sum_i p_{0ij} = p_{0.j} \text{ and } p_{0ij} = b_i p_{1ij}.$$

This condition is also equivalent to the existence, for all i, of $b_i \geq 0$ satisfying $p_{0.j} = \sum_i b_i p_{1ij}$ for all j. If such $(b_i)_{i=1,...,I}$ exist, indeed, p_{0ij} can always be defined by $p_{0ij} = b_i p_{1ij}$.

As a result, the maximum likelihood estimator $\widehat{p}$ of p under Assumptions 2, 3 and 5 is defined by:

$$(\widehat{p},\widehat{b}) = \arg \max_{(q,b)\in[0,1]^{(I+1)J}\times\mathbb{R}^{+I}} \sum_{j=1}^{J}\left[n_{0.j}\ln q_{0.j} + \sum_{i=1}^{I} n_{1ij}\ln q_{1ij}\right]$$
$$\text{s.t.} \left| \begin{array}{l} \sum_{j=1}^{J}\left[q_{0.j} + \sum_{i=1}^{I} q_{1ij}\right] = 1, \\ \sum_{i=1}^{I} q_{1ij} b_i = q_{0.j} \;\; j = 1, ..., J. \end{array} \right.$$

To complete the proposition, remark that $\widehat{p}$ is an asymptotically efficient estimator of p. The Fisher information matrix of p is singular because $\sum_{j,i} p_{1ij} + \sum_j p_{0.j} = 1$. However the parameter u defined by the $IJ + J - 1$ first components of p has a nonsingular Fisher information matrix. If matrix P_1 has rank I then $p_0 = P_1'(P_1P_1')^{-1}P_1 p_{0.}$, and then $p_0 = l(u)$ and $\theta = g(p_0, p_1) = k(u)$ with l and k being differentiable at u. Because $\widehat{P}_1$ has rank i with probability tending to one, $\widehat{b}$ is equal to $(\widehat{P}_1\widehat{P}_1')^{-1}\widehat{P}_1\widehat{p}_{0.}$ with probability tending to one and then $\widehat{b} = m(\widehat{u})$ with m differentiable with probability tending to one. It follows that $\widehat{b}$, $\widehat{p}_0$ and $\widehat{\theta}$ are efficient estimators of $(P_1P_1')^{-1}P_1 p_{0.}$, p_0 and θ (see for instance van der Vaart, 2000, Section 8.9) □

Proposition 3.2

The proof proceeds in three steps. We first establish that the set $\widehat{\mathcal{H}(h_0)}$ approximates well the set $\mathcal{H}(h_0)$. Then we establish the asymptotic distribution of W_n under the null. Thirdly, we compute the asymptotic level of the test, and show that it is consistent.

Step 1. $\Pr\left(\widehat{\mathcal{H}(h_0)} = \mathcal{H}(h_0)\right) \to 1$**.**

It suffices to show that

$$\Pr\left(\widehat{h}_{2j} \le c_n/\sqrt{n}\right) \;\to\; 1 \;\text{ when } h_{20j} = 0 \tag{A.3}$$
$$\Pr\left(\widehat{h}_{2j} > c_n/\sqrt{n}\right) \;\to\; 1 \;\text{ when } h_{20j} > 0 \tag{A.4}$$

We have

$$\sqrt{n}\left(\widehat{h}_{2j} - h_{20j}\right) \to U_j \sim \mathcal{N}(0, \sigma_j^2).$$

Fix $\varepsilon > 0$, and let c_j be such that $\Pr(U_j \le c_j) = \Pr(U_j > -c_j) > 1 - \varepsilon$. Because $c_n \to \infty$, we have $c_n \ge c_j$ for n large enough. Thus, when $h_{20j} = 0$,

$$\Pr(\widehat{h}_{2j} \le c_n/\sqrt{n}) \ge \Pr(\sqrt{n}\widehat{h}_{2j} \le c_j) \to \Pr(U_j \le c_j) > 1 - \varepsilon.$$

This establishes (A.3). When $h_{20j} > 0$, we have $c_n - \sqrt{n}h_{20j} \le -c_j$ for n large enough

because $c_n/\sqrt{n} \to 0$. Thus,

$$\begin{aligned}\Pr\left(\widehat{h}_{2j} > c_n/\sqrt{n}\right) &= \Pr\left(\sqrt{n}(\widehat{h}_{2j} - h_{20j}) > c_n - \sqrt{n}h_{20j}\right) \\ &\geq \Pr\left(\sqrt{n}(\widehat{h}_{2j} - h_{20j}) > -c_j\right) \to \Pr\left(U_j > -c_j\right) > 1 - \varepsilon.\end{aligned}$$

Hence, (A.4) also holds, ending the first step.

Step 2. Asymptotic distribution of W_n under the null.

By Step 1, we can suppose without loss of generality that $\widehat{\mathcal{H}(h_0)} = \mathcal{H}(h_0)$. We show here that (3.3) holds under the null hypothesis. Let $\widetilde{h} = \widehat{h}_2 - \widehat{\Sigma}'_{12}\widehat{\Sigma}^-_{11}\widehat{h}_1$, $U_{2n} = \sqrt{n}(\widetilde{h} - h_{20})$, $V = \Sigma_{22} - \Sigma'_{12}\Sigma^-_{11}\Sigma_{12}$ and $\widehat{V} = \widehat{\Sigma}_{22} - \widehat{\Sigma}'_{12}\widehat{\Sigma}^-_{11}\widehat{\Sigma}_{12}$. Straightforward computations show that $U_{2n} \to U_2 \sim \mathcal{N}(0, V)$. Besides, we have, following Kodde & Palm (1986),

$$\begin{aligned}W_n &= n\widehat{h}'_1\widehat{\Sigma}^-_{11}\widehat{h}_1 + n\min_{x\geq 0}\left(x - \widetilde{h}\right)'\widehat{V}^{-1}\left(x - \widetilde{h}\right) \\ &= n\widehat{h}'_1\widehat{\Sigma}^-_{11}\widehat{h}_1 + \min_{x\geq 0}\left(\sqrt{n}(x - h_{20}) - \sqrt{n}(\widetilde{h} - h_{20})\right)'\widehat{V}^{-1}\left(\sqrt{n}(x - h_{20}) - \sqrt{n}(\widetilde{h} - h_{20})\right) \\ &= n\widehat{h}'_1\widehat{\Sigma}^-_{11}\widehat{h}_1 + \min_{t\geq -\sqrt{n}h_{20}}(t - U_{2n})'\widehat{V}^{-1}(t - U_{2n}).\end{aligned}$$

Let $\mathcal{H}_2(h_0) = \mathcal{R}_1 \times ... \times \mathcal{R}_I$ and define

$$\widetilde{W}_n = n\widehat{h}'_1\widehat{\Sigma}^-_{11}\widehat{h}_1 + \min_{t\in\mathcal{H}_2(h_0)}(t - U_{2n})'\widehat{V}^{-1}(t - U_{2n}).$$

For a given ε, there exists a compact set K such that $\Pr((U_{2n}, \widehat{V}) \in K) \geq 1 - \varepsilon$ for all n large enough. Let $\pi(u, V) = \arg\min_{t\in\mathcal{H}_2(h_0)}(t - u)'V^{-1}(t - u)$. Because $\mathcal{H}_2(h_0)$ is convex, π is a function rather than simply a correspondence. Moreover, it is continuous by Berge maximum theorem (see, e.g., Carter, 2001, Theorem 2.3). Thus $\pi(K)$ is compact. As a result, for n large enough, $\pi(K)$ is included in $[-\sqrt{n}h_{201}, +\infty[\times ... \times [-\sqrt{n}h_{201}, +\infty[$. In other words, for n large enough,

$$\Pr(W_n = \widetilde{W}_n) \geq \Pr((U_{2n}, \widehat{V}) \in K) \geq 1 - \varepsilon. \tag{A.5}$$

Besides, the application $\Xi \mapsto \Xi^-$ is continuous once restricted to matrices or rank J (see, e.g., Stewart, 1969). By continuity of π and the continuous mapping theorem, we have, under the null hypothesis,

$$\widetilde{W}_n \xrightarrow{\mathcal{L}} U_1\Sigma^-_{11}U_1 + \min_{t\in\mathcal{H}_2(h_0)}(t - U_2)'V^{-1}(t - U_2) = \min_{t\in\mathcal{H}(h_0)}(t - U)'\Sigma^-(t - U), \tag{A.6}$$

where $U_1 \sim \mathcal{N}(0, \Sigma_{11})$ and $U = (U_1, U_2 - \Sigma'_{12}\Sigma^-_{11}U_1)$. Note that $U \sim \mathcal{N}(0, \Sigma)$. Besides, (A.5) implies that W_n converges to the same distribution as $\widetilde{W}_n$. Hence, (A.6) implies that (3.3) holds.

Step 3. Consistency and asymptotic level of the test.

Let us define, for any positive matrix Ξ of rank J

$$g(u, \Xi) = \min_{t \in \mathcal{H}(h_0)} (t-u)' \Xi^- (t-u).$$

Let $\widehat{U}$ be a random normal variable satisfying $\widehat{U}|\widehat{\Sigma} \sim \mathcal{N}(0, \widehat{\Sigma})$. Because $\widehat{\Sigma} \xrightarrow{P} \Sigma$, we have $(\widehat{U}, \widehat{\Sigma}) \xrightarrow{\mathcal{L}} (U, \Sigma)$. Thus, by Berge maximum theorem once more, g is continuous. As a result, by the continuous mapping theorem, $g(\widehat{U}, \widehat{\Sigma}) \xrightarrow{\mathcal{L}} g(U, \Sigma)$.

Now, suppose first that $J > I$ or $\mathcal{R}_i = \mathbb{R}^+$ for some $i \in \{1, ..., I\}$. Then $g(U, \Sigma)$ is a mixture of chi-square distributions, and the weight of the chi-square of degree 0 is smaller than 1/2 (see, e.g., Kodde & Palm, 1986). Therefore, its quantile function is continuous on the interval $(1/2, 1)$. Combined with the convergence in distribution of $g(\widehat{U}, \widehat{\Sigma})$, this implies (see, e.g., van der Vaart, 2000, Theorem 21.2) that for any $\alpha \leq 1/2$, $\widehat{c}_\alpha \to c_\alpha$, the quantile of order $1-\alpha$ of $g(U, \Sigma)$. Because the convergence of F_n, the cdf of W_n, towards F, the cdf of $g(U, \Sigma)$, is uniform (van der Vaart, 2000, Lemma 2.11), we have $F_n(\widehat{c}_\alpha) \to F(c_\alpha) = 1 - \alpha$. Thus, the test has the asymptotic level α.

Now, if $J = I$ and $\mathcal{R}_i = \mathbb{R}$ for all i, $g(U, \Sigma) = 0$ and the previous reasoning does not apply. On the other hand, remarking that $W_n = 0$ when $\widehat{\mathcal{H}(h_0)} = \mathcal{H}(h_0)$, we have

$$\Pr(W_n > \widehat{c}_\alpha) \leq \Pr(W_n > 0) \leq \Pr\left(\widehat{\mathcal{H}(h_0)} \neq \mathcal{H}(h_0)\right) \to 0.$$

Thus, the test has asymptotic level 0 in this case.

Finally, under the alternative, $h(p) \notin \mathcal{H}_0$. Then, by the continuous mapping theorem,

$$\min_{h \in \mathcal{H}(h_0)} \left(h - \widehat{h}\right)' \widehat{\Sigma}^- \left(h - \widehat{h}\right) \xrightarrow{P} \min_{h \in \mathcal{H}(h_0)} (h - h(p))' \Sigma^- (h - h(p)) > 0.$$

This implies that $W_n \xrightarrow{P} +\infty$, proving that the test is consistent □

Proof of Theorem 3.3

The proof proceeds in two steps. In the first step we follow the approach of van der Vaart (2000, Chapter 25), who derive a necessary condition for the existence of a regular root-n estimator in semiparametric models. Intuitively, we exploit the fact that any score for the distribution of (D, DY, Z) is a projection of a score for the distribution of (D, Y, Z). This allows us to obtain a necessary condition for the existence of an influence function. In the second step, we characterize the set of such influence functions and derive the semiparametric efficiency bound of β_0.

Let us first introduce some notations. For the random variables U and V, we define $L^{20}(U)$ and $L^{20}(U|V)$ as the following sets of functions:

$$L^{20}(U) = L^2(U) \cap \{f | E(f(U)) = 0\},$$

$$L^{20}(U|V) = L^2(U,V) \cap \left\{f | E(f(U,V)^2|V) < \infty, E(f(U,V)|V) = 0 \;\; V\text{-almost surely}\right\}.$$

For any closed linear space $\mathcal{E} \subset L^2(U)$, we also let $\mathcal{P}_{\mathcal{E}}$ denote the orthogonal projection on $\mathcal{E}$.

Step 1. Assumption 6 is a necessary condition for existence of a regular root-n estimator.

Let $\mathcal{T}$ (respectively $\mathcal{S}$) denote the set of score function, for any subparametric model of the distribution of (D,Y,Z) (respectively of (D,DY,Z)). By Assumption 2, $\mathcal{T} = L^{20}(Y) + L^{20}(Z|Y) + L^{20}(D|Y) \subset L^{20}(D,Y,Z)$. Besides, because (D,DY,Z) is a function of (D,Y,Z), it follows from van der Vaart (2000, Section 25.5) that $\mathcal{S} = \{E(t|D,DY,Z) : t \in \mathcal{T}\}$. Hence, $\mathcal{T}$ and $\mathcal{S}$ are linear and closed here.

We define the score operator A by

$$\begin{array}{rccl} A: & \mathcal{T} & \to & L^{20}(D,DY,Z) \\ & h & \mapsto & [(d,u,z) \mapsto E(h(D,Y,Z)|D=d, DY=u, Z=z)]. \end{array}$$

Note that by definition, $\mathcal{R}(A) = \mathcal{S}$. The usual adjoint of A is the identity here. But, following van der Vaart (2000), we define the adjoint score operator A^* as the adjoint of A followed by the orthogonal projection onto $\mathcal{T}$:

$$\begin{array}{rccl} A^*: & L^{20}(D,DY,Z) & \to & \mathcal{T} \\ & \psi & \mapsto & \mathcal{P}_{\mathcal{T}}\psi. \end{array}$$

Because $L^{20}(Y), L^{20}(Z|Y)$ and $L^{20}(D|Y)$ are orthogonal for the usual inner product of $L^2(D,Y,Z)$, $\mathcal{P}_{\mathcal{T}}\psi = \mathcal{P}_{L^{20}(Y)}\psi + \mathcal{P}_{L^{20}(Z|Y)}\psi + \mathcal{P}_{L^{20}(D|Y)}\psi$.

Now let us consider a regular parametric submodel indexed by θ whose density with respect to an appropriate mesure is

$$f_Y(y,\theta) f_{Z|Y}(z|y,\theta)(P(y,\theta)d + (1 - P(y,\theta)(1-d)).$$

Let also θ_0 denote the parameter corresponding to the true model. Defining $\mu(\theta) = E(g(Y,Z)|\theta)$, we have

$$\mu(\theta) = \int g(y,z) f_Y(y,\theta) f_{Z|Y}(z|y,\theta) dy dz.$$

The score of the submodel in θ_0 is, with obvious notations, $s_Y(y) + s_{Z|Y}(z|y) + \frac{P'(y)(d-P(y))}{P(y)(1-P(y))}$. Then

$$\frac{\partial \mu}{\partial \theta}(\theta_0) = E\left(g(Y,Z)(s_Y(Y) + s_{Z|Y}(Z|Y))\right).$$

It follows that the set of influence function is $\{g(Y,Z)\} + L^{20}(Y,Z)^{\perp}$. We can check that the second term is actually the set of constants. Thus, the efficient influence function corresponding to the complete model where (D,Y,Z) is observed, defined as the unique influence function that belongs to $\mathcal{T}$, is $g(Y,Z) - \beta_0$.

Theorem 25.32 of van der Vaart (2000) shows that if a regular root-n estimator exists, then $g(Y,Z) - \beta_0 \in \mathcal{R}(A^*)$. Let us now prove that this condition is equivalent to Assumption 6.

Let $\psi \in L^{20}(D, DY, Z)$ be such that $A^*(\psi) = g(Y,Z) - \beta_0$. Because A^* is a projector and satisfies $A^* = \mathcal{P}_{L^{20}(Y)} + \mathcal{P}_{L^{20}(Z|Y)} + \mathcal{P}_{L^{20}(D|Y)}$, we have

(a) $\mathcal{P}_{L^{20}(Y)}(\psi) = \mathcal{P}_{L^{20}(Y)}(g - \beta_0)$ or equivalently $E(\psi|Y) = \beta(Y) - \beta_0$,

(b) $\mathcal{P}_{L^{20}(D|Y)}(\psi) = \mathcal{P}_{L^{20}(D|Y)}(g - \beta_0)$ or equivalently $E(\psi|D,Y) - E(\psi|Y) = 0$

Let $m(Y,Z) = \psi(1,Y,Z)$ and $l(Z) = \psi(0,0,Z)$, we have :

$$E(\psi|D,Y) = \beta(Y) - \beta_0 \Rightarrow DE[m(Y,Z)|Y,D] + (1-D)E[l(Z)|Y,D] = \beta(Y) - \beta_0 \quad \text{(A.7)}$$

Hence, if $g(Y,Z) - \beta_0 \in \mathcal{R}(A^*)$, there exists $l \in L^2(Z|D=0)$ such that $E[l(Z)|Y, D=0] = \beta(Y) - \beta_0$ $P^{Y|D=0}-$ almost surely and $m = 1/P(g - \beta_0 - (1-P)l) \in L^2(Y,Z|D=1)$. Equivalently, because $Pm^2 + (1-P)l^2 = (g-\beta_0)^2 + \frac{1-P}{P}(g - \beta_0 - l)^2$, there exists $l \in L^2(Z|D=0)$ such that $E[l(Z)|Y, D=1] = \beta(Y) - \beta_0$ $P^{Y|D=0}-$ almost surely and $E\left(\frac{1-P}{P}(g-\beta_0-l)(g-\beta_0-l)'\right) < \infty$ and $g \in L^2(Y,Z)$. Therefore, there exists $q \in L^2(Z|D=0)$ such that

$$T^*(q) = \beta(.) \text{ and } E\left(\frac{1-P}{P}(g-q)(g-q)'\right) < \infty.$$

Step 2. Characterization of the semiparametric efficiency bound.

First, recall that the semiparametric efficiency bound V^* satisfies

$$V^* = \min_{\psi \in \mathcal{I}} E[\psi\psi'], \quad \text{(A.8)}$$

where the minimum is understood in the partial order of symmetric nonnegative matrices and $\mathcal{I}$ is the set of influence functions, that is to say, the set of ψ satisfying, for all $s = A(\tau)$ $(\tau \in \mathcal{T})$, $E[\psi s] = E[(\beta(.) - \theta_0)\tau]$. Let us first show that

$$\mathcal{I} = \psi_0 + \mathcal{S}^{\perp}, \quad \text{(A.9)}$$

where $\psi_0 \in A^{*-1}(\{g(.,.) - \beta_0\})$. Such an element exists under Assumption 6. First, for any $u \in \mathcal{S}^\perp$,

$$E[(\psi_0 + u)s] = E[\psi_0 s] = E[\psi_0 A(\tau)] = E[A^*(\psi_0)\tau] = E[(\beta(.) - \theta_0)\tau].$$

As a result, $\psi_0 + \mathcal{S}^\perp \subset \mathcal{I}$. Now, let $\psi \in \mathcal{I}$. By definition, $E[(\psi - \psi_0)s] = 0$. Thus, $\psi \in \psi_0 + \mathcal{S}^\perp$ and (A.9) holds. Note that we can also write any $\psi \in \mathcal{I}$ as $\mathcal{P}_\mathcal{S}(\psi_0) + u$, with $u \in \mathcal{S}^\perp$. By orthogonality,

$$E[\psi\psi'] - E[\mathcal{P}_\mathcal{S}(\psi_0)\mathcal{P}_\mathcal{S}(\psi_0)']$$

is nonnegative. Thus,

$$V^* = E[\mathcal{P}_\mathcal{S}(\psi_0)\mathcal{P}_\mathcal{S}(\psi_0)']. \tag{A.10}$$

Now, remark that $\mathcal{P}_{\mathcal{S}^\perp}(\psi_0) \in \mathcal{R}(A)^\perp = \text{Ker}\,(A^*)$. Hence, because $\psi_0 = \mathcal{P}_\mathcal{S}(\psi_0) + \mathcal{P}_{\mathcal{S}^\perp}(\psi_0)$, we have $\mathcal{P}_\mathcal{S}(\psi_0) \in A^{*-1}(\{g(.,.) - \beta_0\})$. Combined with (A.10), this implies that

$$V^* \geq \min_{\psi \in A^{*-1}(\{g(.,.) - \beta_0\})} E[\psi\psi'] \geq V^*,$$

where the inequalities correspond to the partial order of symmetric matrices and the second inequality follows by (A.8) and the fact that $A^{*-1}(\{g(.,.) - \beta_0\}) \subset \mathcal{I}$.

Finally, note that ψ was associated to $l(Z)$, so that taking the minimum in ψ is equivalent to taking the minimum in $l(.) \in T^{*-1}(\{g(.,.) - \beta_0\})$, or in $q(.) \in T^{*-1}(\{g(.,.)\})$ (where $q = l + \beta_0$). Hence, because $V^* = E\left(\psi^*\psi^{*'}\right)$, we have

$$V^* = V(g(Y,Z)) + \min_{q \in T^{*-1}(\{\beta(.)\})} E\left(\frac{1 - P(Y)}{P(Y)}(q(Z) - g(Y,Z))(q(Z) - g(Y,Z))'\right) \quad \square$$