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

n° 2008-23

A New Instrumental Method for Dealing with Endogenous Selection

X. d'HAULTFOEUILLE[1]

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

[1] CREST-INSEE. *Email address* : xavier.dhaultfoeuille@ensae.fr

A New Instrumental Method for Dealing with Endogenous Selection[*]

Xavier d'Haultfoeuille[†]

July 2008

Abstract

This paper develops a new method for dealing with endogenous selection. When selection is directly driven by the dependent variable, the usual instrumental strategy based on the independence between the outcome and the instruments is likely to fail. Instead, the article suggests to rely on independence between the instruments and the selection variable, conditional on the outcome. This may be particularly suitable for nonignorable nonresponse, binary models with missing covariates or Roy models with unobserved sector. Nonparametric identification of the joint distribution of variables is obtained under completeness, a rank condition which has been used recently in several nonparametric instrumental problems. Even if the conditional independence between the instrument and the selection fails, the approach provides sharp bounds on parameters of interest under weaker monotonicity conditions. Apart from identification, nonparametric and parametric estimation is also considered. Eventually, the method is applied to estimate the effect of grade retention in French primary schools.

Keywords: endogenous selection, instrumental variables, nonparametric identification, completeness, inverse problems.

Résumé

Ce papier développe une nouvelle méthode pour traiter de la sélection endogène. Quand la sélection dépend directement de la variable d'intérêt, il peut être difficile d'exhiber un instrument influant la sélection mais pas directement la variable d'intérêt. Une autre stratégie instrumentale, basée sur l'indépendance entre les instruments et la sélection, est considérée ici. Cette stratégie peut être particulièrement utile pour résoudre les problèmes de non-réponse non-ignorable, de modèles binaires avec covariables inobservées ou les modèles de Roy avec secteur inobservé. L'identification non-paramétrique est obtenue sous l'hypothèse que la variable dépendante est une statistique complète pour l'instrument, une condition de rang récemment utilisée dans plusieurs problèmes non-paramétriques instrumentaux. Même si la relation d'exclusion est violée, l'approche développée ici permet d'obtenir des bornes optimales sur des paramètres d'intérêt, sous des conditions plus faibles de monotonicité. Au-delà des résultats d'identification, l'estimation paramétrique et non-paramétrique est également considérée. Enfin, la méthode est appliquée à l'évaluation des effets du redoublement à l'école primaire en France.

Mots clés: sélection endogène, variables instrumentales, identification non-paramétrique, statistique complète, problèmes inverses.

[*] I would like to thank especially Jean-Claude Deville, for inspiring me this project, and Stéphane Bonhomme, for his fruitful suggestions. I also acknowledge Marine Carrasco, Bruno Crépon, Laurent Davezies, Philippe Février, Jean-Pierre Florens, Thierry Magnac, Charles Manski, Arnaud Maurel, Jean-Marc Robin and the participants of the ESEM and of the CEMMAP seminar for their comments.

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

1 Introduction

Missing observations are very common in micro data, either because of selection, nonresponse or simply because variables such as counterfactual cannot be observed. Ignoring this issue by making inference on the observed population generally leads to inconsistent estimators. Moreover, without additional assumptions, only bounds on the parameters of interest can be identified (see e.g. Manski, 2003). Several approaches have been followed to retrieve point identification. The first is to suppose independence between response and variables of interest conditional on observed covariates. This is the so-called missing at random hypothesis (see e.g. Little and Rubin, 1987), or the unconfoundedness assumption in the treatment effect literature (see for instance Imbens, 2004). However, this assumption is often considered too stringent because it rules out any correlation between the selection and outcome variables. When such endogenous selection arises, the common practice is to use instruments which determines selection but not outcomes (see e.g. Heckman, 1974, on tobit models, Angrist et al., 1996, or Heckman and Vytlacil, 2005 on treatment effects). However, this assumption does not point identify the distribution of the outcome in general (see Manski, 2003). Moreover, it may be difficult to find such instruments. When selection depends heavily on the dependent variable, in particular, the assumption of conditional independence is difficult to maintain. A third approach relies on functional restrictions rather than exclusion restrictions. For instance, Chamberlain (1986) obtains identification at the infinity by imposing a linear structure.[1] Lastly, using an appealing composite strategy, Lewbel (2007) obtained identification under the existence of a special regressor which is strongly exogenous (i.e., conditionally independent of the errors of the selection model), a large support condition and restrictions on the probability of selection.[2]

In this paper, another instrumental strategy for solving endogenous selection is considered. Nonparametric identification is based on independence between the instruments and the selection variable, conditional on the outcome and possible on other explanatory variables. This assumption has been also used in the framework of nonignorable nonresponse by Chen (2001), Tang et al. (2003), Hemvanich (2004) and Ramalho and Smith (2007).[3] Apart from

[1] The distribution of categorical data can also be recovered under nonignorable nonresponse, through restrictions in a log-linear model between the dependent and explanatory variables (see e.g. Baker and Laird, 1988 and Park and Brown, 1994). Similarly, endogenous attrition in panel data can be handled by imposing semiparametric restrictions (see, among others, Scharfstein et al., 1999, or Hirano et al., 2001).

[2] This probability must tend to zero or one when the special regressor tends to infinity.

[3] The difference with these papers is that they focus mainly on parametric and semiparametric estimation issues, whereas the emphasis is put on nonparametric identification here. Chen (2001) and Tang et al.

nonresponse, this assumption may be particularly suitable when selection is directly driven by the dependent variable. Consider for instance a variable which is observed only if it exceeds an unobserved truncation. Finding an instrument which only affects selection is impossible if this truncation variable is purely random. Instead, any variable which affects the dependent variable will satisfy the exclusion restriction considered here. Other examples where this assumption can be useful include Roy models with unobserved sector, one stratum response based samples or truncated count data models. As in usual instrumental regressions, a rank condition between instruments and outcomes is also required to achieve identification. This condition is stated in terms of completeness, and was already considered in several nonparametric instrumental problems (see, among others, Newey and Powell, 2003, Florens et al., 2003 and Hu and Schennach, 2008). This hypothesis, together with the conditional independence assumption, enables to recover the joint distribution of the data nonparametrically.

If only some moments of the instrument are used, and not its full distribution, the distribution of the data can still be recovered under a parametric restriction on the selection model. This result may be useful when only aggregated information on the instruments are available, or for the ease of estimation. The idea of using moments of instruments to deal with nonresponse has also been applied in survey sampling (see Deville, 2002). It is also related to the literature on auxiliary information, which has been developed either for efficiency reasons (see Imbens and Lancaster, 1994, Hellerstein and Imbens, 1999) or, as here, to provide identification (see Hellerstein and Imbens, 1999, and Nevo, 2002). Our parametric framework extends Nevo's result to the case of endogenous selection.

The fact that the identification strategy relies on an exclusion restriction may seem restrictive in some applications, and is not needed in Lewbel's framework for instance.[4] However, and contrary to the missing at random assumption for instance, this condition is testable. Furthermore, the method appears to be fruitful even if the exclusion restriction fails. The intuition behind is that this condition is the extreme opposite of unconfoundedness. Indeed, selection depends only on the outcome in the first case and only on covariates in the second one. In between, if selection depends monotonically on both the outcome and

(2003) propose sufficient conditions for identification in parametric models, and Hemvanich (2004) studies identification when the support of the outcome is finite. We extend his result to a general situation here.

[4] On the other hand, the existence of a special regressor, which may be difficult to find in practice, is not needed here. Indeed, the instrument may be continuous or discrete, the completeness condition only implying that its support has the same number of or more elements than the one of the outcome. Moreover, no restriction is imposed on the conditional probability of selection, except, as usual, that it should be positive everywhere.

a given instrument, I show that the identifying equations underlying the two assumptions provide sharp and finite bounds on parameters of the outcome. Thus, even if the dependent variable is unbounded, one can obtain compact interval on parameters of interest. This result is similar to the one of Manski and Pepper (2000) (see their proposition 2, corollary 2) but within a slightly different framework and under other assumptions. Instead of their monotone treatment response condition, which states that outcomes increase with the treatment, the result relies on the existence of an instrument which affects selection in a monotonic way. Such a condition is weak and is likely to be satisfied in many contexts, including the use of data with nonignorable nonresponse and treatment effects estimation. In this latter case in particular, the result should be of practical importance as it enables to go beyond the standard routine of computing matching estimators as point estimates of these effects.

Apart from identification issues, estimation of the model is also considered. Standard GMM can be used in the parametric case or in the nonparametric one with a discrete outcome. In a nonparametric setting with a continuous dependent variable, the parameter is functional and must be estimated through an infinite number of moment conditions. Estimation is based on a Tikhonov regularization method, as in Hall and Horowitz (2005) or Carrasco et al. (2006). The estimator of the conditional probability of selection is shown to be consistent. This estimator enables in turn to make valid inference on the whole population, by an inverse probability weighting procedure, in a similar fashion to Horvitz and Thompson (1952), Hellerstein and Imbens (1999), Nevo (2002) or Wooldridge (2007).

Lastly, the method is used to estimate the effect in terms of test achievement of grade retention in fifth grade in France. Besides the usual counterfactual problem, identification of this effect is complicated by the fact that French students only take standardized tests at the beginning of the third and sixth grades. Thus, the ability at the end of the fifth grade, which is one of the main factor of grade retention, is observed for promoted students, thanks to the sixth test, but not for retained students. Consequently, the problem fits within our framework. Using the third grade test score as an instrument, sharp bounds on the effects of grade retention are computed. Overall, the short term impact of grade retention seems more likely to be positive. This result is in line with the one of Jacob and Lefgren (2004) for third graders in Chicago.

The rest of the paper is structured as follows. Section two is devoted to identification issues. The estimation is discussed in section three. Lastly, the application to grade retention is presented in section four.

2 Identification

2.1 The setting and main result

Let Y denote the dependent variable and D denote the binary variable of selection. The covariates are denoted by X, the instruments by Z. We also define $\mathbf{Y} = (X, Y)$ and $\mathbf{Z} = (X, Z)$. The first assumptions set the selection problem.

Assumption 1 *We observe D and (X, Y, Z) when $D = 1$. Y is not observed when $D = 0$.*

Assumption 2 *The distribution of $\mathbf{Z}$ is identified.*

Assumptions 1 and 2 are satisfied when Y alone is missing, as in selection problems or item nonresponse. It also covers unit nonresponse where (X, Y, Z) are missing when $D = 0$. In this latter situation, auxiliary information on $\mathbf{Z}$ is needed to satisfy assumption 2. This information typically stems from a refreshment sample, censuses or administrative data. In these two latter cases, supposing the identifiability of the whole distribution of $\mathbf{Z}$ may be overly strong, and we will see in subsection 2.3 that it can be weakened to the knowledge of moments of $\mathbf{Z}$, at the price of imposing parametric restrictions.

Assumptions 1 and 2 alone do not enable to point identify the distribution of (D, X, Y, Z). More structure on the dependence between these variables is needed. If selection directly depends on Y, the usual assumption of exogenous selection will fail, and it may be difficult to find an instrument which affects selection but not the outcome. On the other hand, we may find variables which are related to Y but not to D. More precisely, we will assume here the following condition:

Assumption 3 *$D \perp\!\!\!\perp Z \mid \boldsymbol{Y}$.*

This assumption has also been made by Chen (2001), Tang et al. (2003), Hemvanich (2004) and Ramalho and Smith (2007) in the framework of nonresponse. It is also a particular case of assumption (41) of Manski (1994). The condition can be interpreted as follows. The selection equation depends on Y, which is missing when $D = 0$, and thus cannot be identified with the data alone. On the other hand, if an instrument which affects Y but not directly D is available, one can identify this selection equation, in a similar fashion to usual instrumental regressions. For instance, suppose that (D, X, Y, Z) follow the nonparametric system

$$\left\{ \begin{array}{lcl} Y & = & g\,(X, Z, \varepsilon) \\ D & = & h(X, Y, \eta). \end{array} \right. \tag{2.1}$$

In this setting, the independence condition 3 holds if η is independent of (Z, ε) (see lemma 5.2 in appendix A2 for a formal proof). By letting h denote the conditional quantile function of D, we can suppose without loss of generality that η is independent of Y, conditional on X.[5] The exclusion restriction amounts to reinforce this into a conditional independence between η and (Y, Z). Note that in system (2.1), the usual instrumental strategy cannot be applied, since there is no exclusion restriction in the selection equation.

As indicated previously, a dependence condition between $\mathbf{Y}$ and $\mathbf{Z}$ is required to achieve identification of the model. I rely in the sequel on a completeness condition. Let $\mathcal{B}$ denotes the set of real functions h such that $h(\mathbf{Y})$ is bounded below almost surely and $h \in L^1_Y$, where, for any random variable T and any $q > 0$, L^q_T is the space of functions g satisfying $E(|g(T)|^q) < +\infty$.

Assumption 4 ***Y*** *is* $\mathcal{B}$*-complete for* ***Z****, that is for all* $h \in \mathcal{B}$*,*

$$\left(E(h(\boldsymbol{Y})|\boldsymbol{Z}) = 0 \quad a.s.\right) \Longrightarrow \left(h(\boldsymbol{Y}) = 0 \quad a.s.\right). \tag{2.2}$$

Assumption 4 is weaker than the usual completeness condition, for which condition (2.2) must hold for any $h \in L^1_Y$, but stronger than bounded completeness, for which condition (2.2) must hold for bounded functions h only (see e.g. Mattner, 1993, for a discussion on the difference between completeness and bounded completeness). The standard completeness condition has been used in the study of nonparametric instrumental regression under additive separability (see Newey and Powell, 2003, Darolles et al., 2007), local instrumental variables (see Florens et al., 2003) and nonclassical measurement error problems (see Chen and Hu, 2006 and Hu and Schennach, 2008),[6] while the bounded completeness condition has been used for instance by Blundell et al. (2007).

Completeness can be easily characterized when Y and Z have finite supports. Indeed, letting $(y_1, ..., y_s)$ and $(z_1, ..., z_t)$ denote these supports, this assumption amounts to

$$P(\text{rank}(M(X)) = s) = 1, \tag{2.3}$$

where $M(X)$ is the random matrix of typical element $P(Y = y_i | Z = z_j, X)$ (see Newey and Powell, 2003). Hence, the support of Z must be at least as rich as the one of Y ($t \geq s$) and the dependence between the two variables must be strong enough for s distinct conditional

[5] In this case, h is not necessarily structural.

[6] Indeed, assumption 2.4 of Chen and Hu (2006) and assumption 2 of Hu and Schennach (2008) are equivalent, under technical conditions, to a completeness condition.

distributions $P(Y=.|Z=z_j,X)$ to exist. In this case, completeness is equivalent to bounded completeness. Completeness or bounded completeness are much more difficult to characterize when the support of Y or Z is infinite, and only sufficient conditions have been obtained until now. Both hold when the density of Y conditional on Z belongs to an exponential family (see Newey and Powell, 2003). These conditions can also be obtained if Y and $\mathbf{Z}$ are related through a nonparametric model with an additive decomposition and a large support assumption (see d'Haultfœuille, 2008). In lemma 5.3 of appendix A2, this last result is adapted to obtain sufficient conditions for $\mathcal{B}$-completeness in the example of system (2.1).[7] Interestingly, the regularity conditions imposed to obtain $\mathcal{B}-$completeness are hardly stronger than the one used by d'Haultfœuille (2008) to yield bounded completeness. Hence, in this framework at least, the two conditions appear to be quite close.

Because identification is based on inverse probability weighted moment conditions, we also suppose that the conditional probability $P(\mathbf{Y})\equiv P(D=1|\mathbf{Y})$ is positive 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 $\mathbf{Y}$, as in simple truncation models for instance.

Assumption 5 $P(\boldsymbol{Y})>0$ *almost surely.*

Theorem 2.1 *Suppose that assumptions 1-5 hold. Then the distribution of* (D,X,Y,Z) *is identified.*

The proof of theorem 2.1 is deferred to appendix A1. Basically, the result stems from the fact that under assumption 3 and 4, the equation in $Q(.)$

$$E\left(\frac{D}{Q(\mathbf{Y})}\middle|\mathbf{Z}\right)=1 \tag{2.4}$$

admits a unique solution, $P(.)$. Identification of $P(.)$ follows because the left term is identified for any given $Q(.)$. Then it is easy to show that the knowledge of $P(.)$ enables to identify the distribution of (D,X,Y,Z). We now present several potential applications of this framework.

Example 1: nonignorable nonresponse. In this case, Y is observed only if the individual answers to the survey or to a given question in the questionnaire ($D=1$). Nonresponse depends directly on Y (and possibly on covariates X). For instance, accepting to answer questions on sensitive topics such as drug use, religious or sexual preferences is likely to

[7] Of course, these conditions are only sufficient. The large support assumption, especially, is not necessary when Y is discrete.

depend on the answer itself. The method can be applied if an instrument affects Y but not directly D. For instance, local drug prices affect drug use but are unlikely to play directly on response on drug use.

Example 2: Roy model with an unobserved sector. Let for instance Y (resp. W_r) denote the wage an individual can obtain in sector 1 (resp. in sector 0). The individual chooses the sector that provides him with the better wage. Y is observed if sector 1 is chosen but W_r is never observed. For instance, Y may represent the potential wage of an individual, which is observed only if the person enters the labor market, while W_r denotes his reservation wage. The usual exclusion restriction requires the existence a variable which affects W_r but not Y. On the other hand, the strategy above can be applied if there is an instrument Z which affects the wage but not directly the reservation wage, so that W_r is independent of Z conditional on $\mathbf{Y}$. A possible example of such an instrument is the local unemployment rate (see Haurin and Sridhar, 2003, for evidence that the local unemployment rate does not affect the reservation wage).

Example 3: sample from one response stratum. Suppose that a researcher seeks to study the effects of Y on a binary variable D but has only accessed to data from the subpopulation for whom $D = 1$ or, at least, no data on Y is available for the stratum $D = 0$. Examples include the study of prevalence of a disease using hospital records, or the evaluation of a welfare program using different sources for the participants and untreated individuals (see Manski, 2003). Our instrumental strategy relies on the existence of an instrument Z which affects Y but not D directly, and whose distribution is identified. Suppose for instance that one wants to study the efficiency of vaccination in a developing country, but data on ill people only are available, and the vaccination rate in the population is unknown. If there has been an important vaccination campaign after a given date, one can use the dummy of being born after this date as an instrument.[8]

Example 4: truncated count data models. We seek to modelize an integer valued variable Q as a function of covariates Y but only observe (Q, Y) when $Q > 0$. An application is the use of retail data to study the demand for a good.[9] Suppose indeed that we observe the quantities sold and the sales, but not the prices. Then these prices can be deduced only when the quantities sold are positive. The framework can be applied if there is an

[8] If age is a factor of the disease a well, one can use only individuals born just before and just after the beginning of the campaign, as in the regression-discontinuity approach.

[9] As discussed by Grogger and Carson (1991), truncated counts arise more generally with data from surveys which ask participants about their number of participations, or administrative records where inclusion in the database is predicated on having engaged in the activity of interest.

instrument whose distribution is identified and which affects the prices but not directly the demand. Production costs or prices of the inputs, for instance, may be good candidates for that.

2.2 Testability and set identification without conditional independence

In some contexts, the conditional independence assumption 3 may seem overly strong. An interesting feature of this assumption, yet, is that it is refutable, contrary to the usual missing at random assumption. Firstly, equation (2.4) may have no solution. This is especially clear when (X, Y, Z) has a finite support. If indeed $\mathbf{Y}$ and $\mathbf{Z}$ take respectively m_1 and m_2 distinct values, with $m_2 > m_1$, (2.4) can be written as a system of m_2 linear equations with m_1 unknown parameters, so that the model is overidentified.

But even when $m_1 = m_2$, the model is testable since the solution $Q(.)$ of equation (2.4) must be a positive probability, i.e. $Q(y) \in]0,1]$ for all y.[10] As an illustration, consider a simple case without covariates and such that $(Y, Z) \in \{0,1\}^2$. Let $p(y,z) = P(D = 1, Y = y|Z = z)$, $\alpha = 1/Q(0)$ and $\beta = 1/Q(1)$. Then, as soon as $p(0,0)p(1,1) \neq p(0,1)p(1,0)$ (that is to say under the completeness condition), equation (2.4) is equivalent to

$$\begin{aligned}\alpha &= \frac{p(1,1) - p(1,0)}{p(0,0)p(1,1) - p(0,1)p(1,0)}\\ \beta &= \frac{p(0,0) - p(0,1)}{p(0,0)p(1,1) - p(0,1)p(1,0)}.\end{aligned}$$

Hence, when $p(1,1) - p(1,0)$ and $p(0,0) - p(0,1)$ have opposite signs, assumption 3 is rejected.[11] Basically, this happens when $z \mapsto P(D = 1|Y = y, Z = z)$ varies too much compared to $z \mapsto P(Y = y|Z = z)$.

Now, when a solution $Q(.) \in]0,1]$ of equation (2.4) does exist, one can expect that assumption 3 cannot be rejected, since intuitively, this equation makes use of all the available information. Theorem 2.2 formalizes this idea.

Theorem 2.2 *Suppose that assumption 1, 2 and 5 hold. Then assumption 3 can be rejected if and only there exists no solution $Q(.)$ of equation (2.4) which belongs to $]0,1]$.*

A second interesting feature of equation (2.4) is that it provides an informative bound on parameters of interest under monotonicity conditions, which are far weaker than the

[10] If the completeness condition does not hold, $Q(.)$ may not be unique. Then at least one of the solution must belong to $]0,1]$.

[11] Of course, there are other cases where the assumption is violated.

conditional independence condition of assumption 3. In the sequel, we let $\mathbf{Z}' = (X, Z')$ denote covariates which may be different or not from $\mathbf{Z} = (X, Z)$ and whose distribution is also identified. Besides, we restrict to the case where $(Y, Z) \in \mathbb{R}^2$. We replace assumption 3 by the following ones.

Assumption 3' *Almost surely, $z \mapsto P(D = 1|\boldsymbol{Y}, Z = z)$ is increasing.*

Assumption 6 *Almost surely, $y \mapsto P(D = 1|Y = y, \boldsymbol{Z'})$ is increasing.*

Assumption 3' weakens the conditional independence between selection and instrument set in assumption 3 into a monotone dependence. It is also a variant of the usual instrumental condition which supposes that the instrument affects the probability of selection but is independent of the outcome. Here, the effect on the probability of select is restricted to be monotonic, but no independence condition between Y and Z is needed. Assumption 6 weakens the missing at random hypothesis of independence between selection and outcome into a monotone dependence. This assumption is very similar to the mean missing monotonicity assumption considered by Manski (2003, p. 28), and actually implies it, as part a) of theorem 2.3 shows.

Theorem 2.3 below provides bounds on parameters of the form $E(h(\mathbf{Y}))$ for $h \in H^1_Y$ or $h \in H^2_Y$, where we let

$$\begin{array}{rcl} H^1_T & = & \{h(x, .) \in L^1_T \text{ and } h(x, .) \text{ is increasing, for } P^X\text{-almost all } x\}, \ (T = Y \text{ or } Z) \\ H^2_Y & = & \{h \in L^1_Y / \exists \psi \in H^1_Z / h(\mathbf{Y}) = E(\psi(\mathbf{Z})|D = 1, \mathbf{Y})\}. \end{array}$$

We suppose in the following that equation (2.4) admits a solution, and, as before, we let $Q(.)$ denote such a solution. More precisely, if the constant function $P(D = 1)$ is a solution, we let $Q(\mathbf{Y}) = P(D = 1)$ but otherwise $Q(.)$ can be any of the solutions. We do not impose it neither to lie in $]0, 1]$ nor to be unique, so that cases where the completeness condition 4 fails can also be handled.

Theorem 2.3 *Suppose that $P(D = 1) > 0$ and assumptions 1 and 2 hold for $\boldsymbol{Z}$ and $\boldsymbol{Z'}$. Then:*
a) Under assumption 6, $E\left[h(\boldsymbol{Y})|X\right] \leq E\left[E(h(\boldsymbol{Y})|\boldsymbol{Z'}, D = 1)|X\right]$ for all $h \in H^1_Y$. Moreover, this upper bound is sharp ;
b) Under assumptions 3', $E\left[Dh(\boldsymbol{Y})/Q(\boldsymbol{Y})|X\right] \leq E\left[h(\boldsymbol{Y})|X\right]$ for all function $h \in H^2_Y$. Moreover, this lower bound is sharp provided that at least one solution $Q(.)$ lies in $]0; 1]$.
c) For all function $h \in L^1_{\boldsymbol{Y}}$, these three expectations are equal when $D \perp\!\!\!\perp (\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{Z'})$ or when $\boldsymbol{Z} = \boldsymbol{Z'} = \boldsymbol{Y}$.

Part a) of theorem 2.3 is not specific to the methodology developed here, and is rather straightforward. Part b), on the other hand, shows that the moment condition used here leads to a sharp lower bound on this parameter. This lower bound does not depend on the choice of the solution $Q(.)$ of equation (2.4), so that no completeness condition is required. The bound also holds even if no solution $Q(.)$ lies in $]0;\ 1]$. In this case however, the bound may not be sharp because one could exploit the fact that the conditional independence assumption 3 is rejected by the data.

An important consequence of theorem 2.3 is that for all functions $h \in H_Y^1 \cap H_Y^2$, we can obtain a compact interval on $E(h(\mathbf{Y})|X)$. This is so even if $h(\mathbf{Y})$ is unbounded. In this sense, the result is similar to proposition 2, corollary 2 of Manski and Pepper (2000), under a different set of assumptions. In particular, we do not rely on the monotone treatment response condition, which is difficult to adapt to the context of selection models or nonresponse. Moreover, the monotone treatment response assumption can be strong in the context of treatment effects. In the Roy model with an unobserved sector developed in example 2, it asserts that almost surely, $Y_1 \geq Y_0$ (or $Y_0 \geq Y_1$), so that only one sector would be chosen at equilibrium, a rather unrealistic situation. Instead of this condition, assumption 3' supposes the existence of an instrument such that the probability of selection increases with this instrument. This assumption is rather weak and should be satisfied in many contexts, including treatment effects estimation or estimation of parameters with nonignorable missing data. In example 2, one could use standard instruments such as non-wage income or the number of children (actually, their opposite) for instance.

As part c) shows, the interval can be reduced to a point if D is fully missing at random. Hence, the length of the interval can be interpreted as a measure of the severity of the selection problem. Because the interval is also reduced to a point when $\mathbf{Z} = \mathbf{Z}' = \mathbf{Y}$, its length also reflects the quality of the chosen instruments. As the dependence between $(\mathbf{Z}, \mathbf{Z}')$ and $\mathbf{Y}$ increases, the knowledge of the distribution of the instruments enables to better predict parameters of the distribution of $\mathbf{Y}$. Besides, the upper (resp. lower) inequality turns into an equalities whenever $Y \perp\!\!\!\perp D|\mathbf{Z}'$ (resp. $\mathbf{Z} \perp\!\!\!\perp D|\mathbf{Y}$). Hence, Z and Z' must be chosen according to different logics. Z' intends to reduce selection on inobservables correlated with the outcome, whereas Z should be as independent of the selection (conditional on $\mathbf{Y}$) as possible.

The class of parameters for which a lower bound is available depends on H_Y^2 which, though identifiable, is rather abstract. In an informal way however, H_Y^2 will generally increase as the dependence between $\mathbf{Y}$ and $\mathbf{Z}$ becomes stronger. As a simple illustration, this set only includes constant function when $\mathbf{Y}$ and $\mathbf{Z}$ are independent (conditional on $D = 1$) but is equal to H_Y^1 when $\mathbf{Y} = \mathbf{Z}$. More formally, H_Y^2 is a subset of the range of the conditional expectation operator $f \mapsto (y \mapsto E(f(\mathbf{Z})|D = 1, \mathbf{Y} = y))$, which itself is linked to the null space of this operator. Indeed, when $(\mathbf{Y}, \mathbf{Z})$ has finite support, the dimension of the range will increase as the dimension of the null space decreases. Thus, at least in finite dimension, H_Y^2 will be maximal if the conditional expectation operator is injective, that is to say under a completeness condition on $\mathbf{Y}$ and $\mathbf{Z}$.[12] Hence, the quality of the instrument also matters for the range of applicability of the lower bound.

It is difficult to define precisely the set $H_Y^1 \cap H_Y^2$ of functions h such that an interval can be built on $E\left[h(\mathbf{Y})|X\right]$ without further restrictions. An interesting one is the following random coefficients linear model:

$$E(Z|\mathbf{Y}, D = 1) = \alpha(X) + \beta(X)Y, \quad \beta(.) > 0.$$

In this case indeed, $H_Y^1 \cap H_Y^2$ contains at least all functions $h(x, y) = \lambda(x)y$ with $\lambda(.) > 0$, so that $E[Y|X]$ can be bounded below and above. Besides, if Y and Z exhibit a positive dependence, this set will be equal to H_Y^2.

Lemma 2.1 *Suppose that for all (x, z), $y \mapsto F_{Z|X=x,Y=y,D=1}(z)$ is decreasing. Then $H_Y^2 \subset H_Y^1$.*

2.3 Parametric identification

Nonparametric identification stems from the uniqueness of a functional equation. However, one may be reluctant to use nonparametric estimators in practice, because of the curse of dimensionality for instance. Furthermore, assumption 2 may be too strong in some circumstances. Suppose for instance that instruments are observed only when $D = 1$ (as with unit nonresponse or attrition in a panel), but auxiliary information is available on these instruments. This auxiliary information may however not be sufficient to identify the full distribution of $\mathbf{Z}$. If $\mathbf{Z}$ is multivariate and its different components are observed through different sources which cannot be matched, only the marginal distributions will be identified. If the instruments are measured with a zero mean error in these auxiliary data, only $E(\mathbf{Z})$ can be recovered.

In such situations, assumption 2 fails but intuitively, information on $\mathbf{Z}$ can provide identification, at least in a parametric setting. Theorem 2.3 gives a rigorous treatment to this

[12] If (Y, Z) has infinite support and the conditional expectation operator is injective, one can show that the dimension of H_Y^2 is infinite.

idea. It generalizes the framework of Nevo (2002) to the case where $\mathbf{Y} \neq \mathbf{Z}$. It is also very similar to the theory of generalized calibration developed by Deville (2002) in a survey sampling framework to handle nonignorable nonresponse with instruments. Deville (2002), however, does not consider the issue of identification of $P(.)$.

Let us suppose here that $\mathbf{Y} \in \mathbb{R}^p$ and $\mathbf{Z} \in \mathbb{R}^q$. The identification result is based on the following assumptions.

Assumption 2' *$E(\boldsymbol{Z})$ is known. Moreover, $P(\boldsymbol{Y}) = F(\boldsymbol{Y}'\beta_0)$ where F is a known, differentiable and strictly increasing function from $\mathbb{R}$ to $]0,1[$, and $\boldsymbol{Y}$ is almost surely linearly independent conditional on $D = 1$.*

Assumption 4' *$rank(E(D\boldsymbol{Z}\boldsymbol{Y}'F'(\boldsymbol{Y}'\beta_0)/F^2(\boldsymbol{Y}'\beta_0))) = p$.*

Assumption 4" *$E(Z|\boldsymbol{Y}, D = 1) = \Gamma_1 X + \Gamma_2 Y$ where Γ_2 is full rank.*

Assumption 2' weakens 2 on data availability, at the price of imposing a parametric restriction on $P(.)$. Like assumption 4 in the nonparametric setting, assumption 4' is the rank condition. As usually, this condition implies that $q \geq p$. Lastly, assumption 4" is a particular case of assumption 4", which restricts the nonparametric regression of Z on $\mathbf{Y}$ to a linear form.

Theorem 2.4 *Suppose that assumptions 1, 2' and 3 are satisfied. then*
a) β_0 is locally identified if and only if assumption 4' holds.
b) if assumption 4" holds, β_0 is globally identified.

Local identification is obtained under a condition which is very similar to the rank condition in linear regressions with instruments. Theorem 2.4 also provides a sufficient and testable condition which ensures the global identification of β_0.

3 Estimation

We now turn to the parametric and nonparametric estimation of $P(.)$. The first assumption describes the sampling process. In the sequel, we let $Y^* = DY$.

Assumption 7 *We observe a sample $((D_1, X_1, Y_1^*, Z_1), ..., (D_n, X_n, Y_n^*, Z_n))$ of independent copies of (D, X, Y^*, Z).*

Assuming that the data are i.i.d. is standard in estimation, although this condition can be weakened without affecting consistency or rate of convergence. We also suppose, for the sake of simplicity, that Z is always observed in the data.

3.1 Parametric estimation

When $\mathbf{Y}$ has a finite support $\{y_1, ..., y_K\}$, the equation

$$E\left(\frac{D}{\sum_{k=1}^{K} P(y_k) 1\{\mathbf{Y} = y_k\}} - 1 \middle| \mathbf{Z} \right) = 0$$

provides identification of the parameters $(P(y_k))_{1 \leq k \leq K}$ if assumptions 3, 4 and 5 hold, by theorem 2.1. Hence, consistent and asymptotically normal estimators can be obtained by GMM in this case. Similarly, if $P(.)$ satisfies the restrictions of assumption 2', then

$$E\left[\left(\frac{D}{F(\mathbf{Y}'\beta_0)} - 1 \right) Z \right] = 0. \tag{3.1}$$

Moreover, the proof of theorem 2.4 (see equation (5.6)) ensures that under assumption 4", β_0 is identified globally by these conditions. Thus GMM can also be used in this framework.

3.2 Nonparametric estimation

When $\mathbf{Y}$ has continuous components and one is reluctant to rely on parametric restrictions on $P(.)$, the situation is more involved because a function, and not only parameters, must be estimated. This issue is similar to the one of nonparametric instrumental regression (see e.g. Newey and Powell, 2003, Darolles et al., 2002, Hall and Horowitz, Horowitz and Lee, 2007). For the sake of simplicity, we assume here that $X = \emptyset$ and $(Y, Z) \in [0,1]^2$. Moreover, since the paper is mainly focused on identification, we only prove consistency here. The analysis of the rate of convergence could be lead by adapting the arguments of Hall and Horowitz (2005).

Let us denote $f = 1/P$ and T be defined as

$$T(h)(z) = E(Dh(Y^*)|Z = z).$$

Then (2.4) may be written as

$$T(f) = 1.$$

Under assumptions 3 and 4 (with $\mathcal{A} = L^1_Y$), T is injective.[13] However, its inverse is not continuous, so that we are faced to an ill-posed problem.[14] To achieve consistency, we

[13] Indeed, by conditional independence, $T(h_1) = T(h_2)$ implies $E(P(Y)(h_1(Y) - h_2(Y))|Z) = 0$. By completeness and positivity of $P(.)$, this implies $h_1 = h_2$.

[14] One may argue that the constant function one is known, so that regularization is not needed here. Actually it is, because T is unknown and can be estimated only by a finite range estimator. This situation is similar to the one of Gagliardini and Scaillet (2006) in the framework of functional minimum distance.

adopt a Tikhonov regularization as Darolles et al. (2002), Hall and Horowitz (2005) and Horowitz and Lee (2007).

First, we consider a kernel estimator of T :

$$\widehat{T}(\phi)(z) = \frac{\sum_{i=1}^{n} D_i \phi(Y_i^*) K_{h_n}(z - Z_i)}{\sum_{i=1}^{n} K_{h_n}(z - Z_i)}$$

For any $1 < M < \infty$, let us define D_M as the subset of real measurable functions ϕ defined on $[0, 1]$ and such that $M \geq \phi(Y) \geq 1$ almost surely. For any square integrable function g defined on $[0, 1]$, let also $||g||^2 = \int_0^1 g(u)^2 du$. Our estimator of f satisfies

$$\widehat{f} \in \arg\min_{\phi \in D_M} \left\|\widehat{T}(\phi) - 1\right\|^2 + \alpha_n \left\|\phi\right\|^2.$$

Under the assumptions below, such a solution will always exist but may not be unique (see Bissantz et al., 2004). If not, $\widehat{f}$ is any of the solutions. The consistency result relies on the following assumptions. In the sequel, $\delta_n = h_n^2 + 1/nh_n$.

Assumption 8 *(a) $f \in D_M$. (b) The distribution of (Y, Z) is continuous with respect to the Lebesgue measure and the marginal densities f_Y and f_Z satisfy* $\sup_{y \in [0,1]} f_Y(y) < +\infty$ *and* $\inf_{z \in [0,1]} f_Z(z) > 0$.

Assumption 9 *For all $h > 0$ and $u \in$, $K_h(u) = K_1(u/h)$ where K_1 is positive, $\int K_1(u)du = 1$ and $\int uK_1(u)du = 0$.*

Assumption 10 *$\alpha_n \to 0$, $\delta_n \to 0$ and $\delta_n/\alpha_n \to 0$.*

Assumption 8-(a) strengthens assumption 5. Assumption 9 is weak and standard in nonparametric estimation. Assumption 10, which is identical to assumption 3 of Horowitz and Lee (2007), is also standard. It implies that the bandwidth h_n tends to zero at a slower rate than $1/n$, and that the regularization parameter α_n tends to zero at a slower rate than h_n^2.

Theorem 3.1 *Under assumptions 3-4 and 7-10,*

$$\lim_{n \to \infty} E\left(\left\|\widehat{f} - f\right\|^2\right) = 0$$

Theorem 3.1 implies that $\left\|\widehat{f} - f\right\|^2$ converges in probability to zero. With $\widehat{f}$ in hand, inverse probability weighting procedures can be used to estimate parameters on the whole population. Let $\widehat{f}^{-i}$ denotes the estimator of f obtained with the sample $(D_j, Y_j^*, Z_j)_{j\neq i}$. For any $g \in L^2_{Y,Z}$ and $\theta = E(g(Y, Z))$, define

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} D_i \widehat{f}^{-i}(Y_i^*)g(Y_i^*, Z_i).$$

Corollary 3.2 ensures that $\widehat{\theta}$ is consistent.

Corollary 3.2 *Suppose that assumptions 3, 4 and 7-10 hold. Then*

$$\lim_{n\to\infty} E\left(|\widehat{\theta} - \theta|\right) = 0$$

4 Application

4.1 Introduction

In this section, the strategy developed above is exploited to estimate bounds on the short term effects of grade retention among fifth grade students in France. Whereas most countries have almost completely given up grade retention as an educational policy,[15] the level of grade retention in France is still high. In 2002, for instance, a quarter of the students have repeated at least once in primary school (see Troncin, 2004). Yet, and despite the controversy on its effects in other countries,[16] there has been no serious attempts to measure its impact in the French educational system.[17]

There are two challenges in estimating the effects of grade retention on test score achievement in France. The first is the usual counterfactual problem. Indeed, unobservable characteristics such as motivation, maturity or parental involvement are likely to play both on

[15] A notable exception is United States. Indeed, several states have reintroduced this policy by tying promotion on a state or district assessment (see Jacob and Lefgren, 2004).

[16] Positive effects include the possibility for disadvantaged children to catch up (see e.g. Jacob and Lefgren, 2004) and the incentive for every student to increase their school efforts (see Jacob, 2005). On the other hand, most educational and sociological studies underline its harmful effects on the motivation of children (see e.g. Crahaye, 1996), drop outs (see Jimerson et al., 2002) and even academic performances (see e.g. the meta-analyses of Holmes, 1989, or Jimerson, 2001).However, usually, these studies rely on very few controls (see e.g. Lorence, 2006, for a discussion on the studies considered in the meta-analyses of Holmes and Jimerson), so that they probably underestimate the true effects of grade retention.

[17] Troncin (2005) measures the the effects of grade retention in the first grade of primary school using a propensity score matching approach, but he relies on data from one school only. Cosnefroy and Rocher (2004) study the effects in third grade on the same data as here, using a linear regression approach.

test score achievement and grade retention. Moreover, grade retention is not tied to any exogenous rule in France, contrary to Chicago or Chile schools (see respectively Jacob and Lefgren, 2004, and Manacorda, 2007).[18] To overcome this problem, credible bounds on the progression students would have made had they not repeat their grade are assumed. A second difficulty is that this progression depends on the ability at the end of the fifth grade, which is unobserved for repeating students. Standardized tests in France are taken indeed at the beginning of the third and sixth grades only, so that the ability of retained students at the end of their first year in fifth grade is unobserved.[19] Hence, we are faced to a severe missing data problem, since the main factor of grade retention is unobserved for retained students. The following subsection shows how the identification strategy developed above can be used to solve this issue.

The study is based on a panel of the French "Ministère de l'éducation Nationale" which follows 9641 children who enter the first grade of primary school in 1997. Among others, the panel reports the trajectories of children and their results in standardized tests at the beginning of the third and sixth grade.[20] Because the sixth grade test scores are reported in the database only for pupils who reached this grade in 2002 or in 2003, the initial sample comprises 7175 students who were in fifth grade in 2001 and in sixth grade either in 2002 or in 2003.[21]. 23.8 percent of this sample was excluded because of missing data either on the standardized test scores in third and sixth grade. The final sample consists in 5467 children. Table 1 displays the average scores on this sample. It especially underlines the important differences between retained and promoted pupils in terms of test achievement. The table also displays the progression of students retained in 6th grade during their first year in this grade. This progression is available because these students take the test twice, at the beginning of their first and second year in sixth grade. This feature of the sample will be useful in the following.

[18] An Education Bill of the Minister of the Education in 2005 asserts that grade retention should be taken by teachers after discussion with parents, according to the ability of the student and his progression during the year.

[19] On the other hand, the ability of promoted students at the end of the fifth grade is observed by their sixth grade test.

[20] Tests corresponding to a given grade differ partly from year to year. The scores considered here are built using common items only. The three scores are also standardized on the final sample.

[21] Other situations correspond to missing data on the trajectories, grade-advanced pupils, pupils retained before the fifth grade and students in special classrooms.

	Retained in 5th grade	Retained in 6th grade	Promoted in both grades
3rd grade score	-1.48 (0.91)	-1.02 (0.90)	0.11 (0.94)
2002 6th grade score	-	-1.32 (0.81)	0.12 (0.93)
2003 6th grade score	-0.90 (0.87)	-0.64 (0.79)	-
Number of observations	120	365	4982

Table 1: Summary statistics.

4.2 Empirical strategy

We aim at identifying the average effects of retention in fifth grade on ability one year after. Let D_0 denote the dummy of promotion in sixth grade and let $Y_1(1)$ (resp. $Y_1(0)$) denote the achievement of a student who is promoted in sixth grade (resp. retained in fifth grade) in the sixth grade standardized test, at the end of the sixth grade (resp. at the end of his second fifth grade). The parameter of interest writes as

$$\Delta^{TT} = E(Y_1(0) - Y_1(1)|D_0 = 0) \tag{4.1}$$

As usually, $Y_1(0)$ is observed when $D_0 = 0$ but $Y_1(1)$ is not (see figure 1). Because there is no exogenous rule acting on grade retention decisions in France, it seems difficult to rely on an instrumental strategy to overcome this counterfactual issue. Rather, I suppose that the progressions of retained students had they been promoted in sixth grade can be bounded in the following way:

$$0 \leq E(Y_1(1) - Y|D_0 = 0, Y) \leq E(Y_1(1) - Y|D_0 = 1, D_1 = 0, Y). \tag{4.2}$$

The lower bound simply asserts that on average, retained students would not have regressed during one year, had they been promoted. The upper bound states that on average, their progression would have been smaller than the one of students with same initial test score and who are promoted in sixth grade and retained the year after. The idea behind this bound is that, on average, teachers do not make mistakes by retaining pupils who would have benefited more from the sixth grade than some of the promoted students. The two bounds somewhat represent two extreme situations. The lower bound corresponds to perfect decisions of retention, in that retained students would not have taken any advantage of being promoted. The upper bound corresponds to a fully randomized choice among students who would have equally benefited from being promoted.

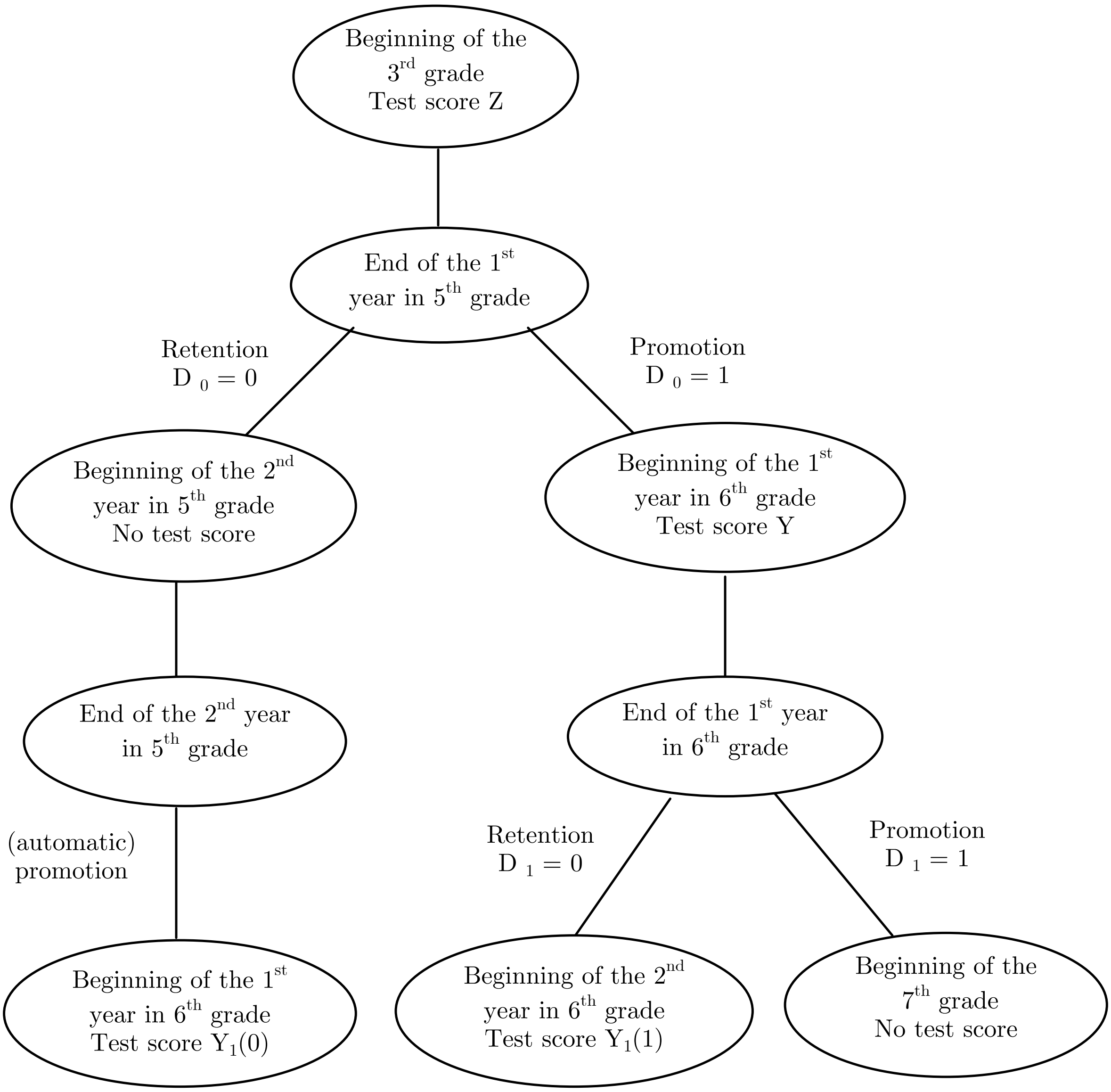
Beginning of the 3rd grade Test score Z
End of the 1st year in 5th grade
Retention D 0 = 0
Promotion D 0 = 1
Beginning of the 2nd year in 5th grade No test score
Beginning of the 1st year in 6th grade Test score Y
End of the 2nd year in 5th grade
End of the 1st year in 6th grade
(automatic) promotion
Retention D 1 = 0
Promotion D 1 = 1
Beginning of the 1st year in 6th grade Test score Y1(0)
Beginning of the 2nd year in 6th grade Test score Y1(1)
Beginning of the 7th grade No test score

Figure 1: Promotion, retention and available test scores.

Under condition (4.2), we get

$$E(Y_1(0)|D_0 = 0) - E\left[h(Y)|D_0 = 0\right] \leq \Delta^{TT} \leq E(Y_1(0)|D_0 = 0) - E(Y|D_0 = 0), \quad (4.3)$$

where $h(Y) = E(Y_1(1)|D_0 = 1, D_1 = 0, Y)$. Students retained in sixth grade take the standardized test twice. Thus, we observe both Y and $Y_1(1)$ for them, and $h(.)$ is identified. On the other hand, Y is unobserved for students retained in fifth grade, so that $E[h(Y)|D_0 = 0]$ and $E(Y|D_0 = 0)$ are not identified without further restrictions. We now rely on the results of section two to obtain bounds on Δ^{TT}, using the third grade standardized test score Z as an instrument.

First strategy: conditional independence

First, let us suppose that grade retention in fifth grade is independent of the third grade test score conditional on Y, i.e. a model of the form:

$$\begin{cases} Y &= g(Z, \varepsilon) \\ D &= h(Y, \eta) \end{cases}$$

where $\eta \perp\!\!\!\perp (Z, \varepsilon)$. The completeness condition is also supposed to hold. Informally, both will be satisfied if the third grade score affects the ability at the end of the fifth grade, measured by the sixth grade test score, but not directly grade retention. Under these assumptions, theorem 2.1 applies and we can identify $E(h(Y)|D_0 = 0)$ by

$$E\left[h(Y)|D_0 = 0\right] = \frac{1-p}{p} E\left[\frac{1-Q(Y)}{Q(Y)} h(Y)|D_0 = 1\right]$$

where $p = P(D_0 = 0)$ and $Q(Y)$ is the solution of $E(D/Q(Y) - 1|Z) = 0$. $E(Y|D_0 = 0)$ can be identified similarly. Then, using (4.2), we obtain the following lower and upper bounds on Δ^{TT}:

$$\underline{\Delta}_1^{TT} = E[Y_1(0)|D_0 = 0] - \frac{1-p}{p} E\left[\frac{1-Q(Y)}{Q(Y)} h(Y)|D_0 = 1\right] \quad (4.4)$$

$$\overline{\Delta}^{TT} = E[Y_1(0)|D_0 = 0] - \frac{1-p}{p} E\left[\frac{1-Q(Y)}{Q(Y)} Y|D_0 = 1\right], \quad (4.5)$$

Second strategy: monotonicity

Basically, the conditional independence condition holds if the test score at the beginning of the sixth grade is a perfect measure of ability at the end of fifth grade and if teachers only take into account the current ability when deciding whether to retain a student or not. If the second statement is rather plausible given that teachers usually do not observe children's ability before they enter their grade, the first statement seems too restrictive. Past scores probably bring additional information on the current ability and thus explain part of grade retention. On the other hand, it seems very plausible in this case that the dependence in both variable is monotonic, i.e., assumption 3' and 6 hold. To provide empirical evidence on this assumption, a logit model on D_1 among students who were promoted in sixth grade was estimated. For these students indeed, both Y and Z are known. The results, which are displayed table 2, confirm the monotonicity in both variables. As expected, we also observe a far smaller effects of the third grade test score.

Variable	Estimate (std dev)
6th grade score	1.31 (0.08)
3rd grade score	0.23 (0.07)

Table 2: Logit estimation on the probability of retention in sixth grade.

Under assumptions 3' and 6, and provided that h is in $H^1_Y \cap H^2_Y$, we can apply theorem 2.3 to obtain the following bounds on $E(h(Y)|D_0 = 0)$:

$$\frac{1-p}{p} E\left[\frac{1-Q(Y)}{Q(Y)} h(Y)|D_0 = 1\right] \leq E\left[h(Y)|D_0 = 0\right] \leq E\left[E(h(Y)|Z, D_0 = 0)|D_0 = 1\right]$$

The same holds for $E(Y|D_0 = 0)$, provided that the identify function is in H^2_Y. Under these assumptions, we get the same upper bound on Δ^{TT} as under conditional independence, but another lower bound, which writes as

$$\underline{\Delta}_2^{TT} = E[Y_1(0)|D_0 = 0] - E\left[E(h(Y)|Z, D_0 = 1)|D_0 = 0\right] \tag{4.6}$$

Moreover, $\underline{\Delta}_2^{TT}$ and $\overline{\Delta}^{TT}$ are sharp by theorem 2.3.

4.3 Results

To compute the upper bound of Δ^{TT}, the following flexible parametric form on $y \mapsto Q(y)$ was used:

$$Q(y;\beta) = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{i=1}^k y 1\{y \geq \alpha_i\}\beta_i\right)}. \tag{4.7}$$

In the sequel, $k = 4$, $\alpha_1 = -\infty$ and $(\alpha_i)_{2\leq i\leq k}$ correspond to the estimated quantiles of order 8, 16 and 24 of Y.[22]. The parameter $\beta = (\beta_0, ..., \beta_6)$ is estimated through GMM, using as instrumental variables 1 and $(Z1\{Z \geq \gamma_i\})_{1\leq i\leq k}$, where $\gamma_1 = -\infty$ and the $(\gamma_i)_{2\leq i\leq k}$ are the estimated quantiles of order 8, 16 and 24 of Z. The estimated probability of promotion is displayed figure 2.[23]

The estimator of $\overline{\Delta}^{TT}$ is the empirical analog of (4.5):

$$\widehat{\overline{\Delta}^{TT}} = \frac{1}{n_0}\left[\sum_{i/D_i=0} Y_{1i}(0) - \sum_{i/D_i=1} \frac{Q(Y_i;\widehat{\beta})}{1-Q(Y_i;\widehat{\beta})} Y_i\right],$$

[22] Several specifications have been tried. Final results are unsensitive to the choice of k and $(\alpha_i)_{2\leq i\leq k}$.

[23] This plot corresponds to $\widehat{\beta}_0 = 3.07$, $\widehat{\beta}_1 = 0.75$, $\widehat{\beta}_2 = 4.13$,$\widehat{\beta}_3 = 34.3$, $\widehat{\beta}_4 = 0.42$.

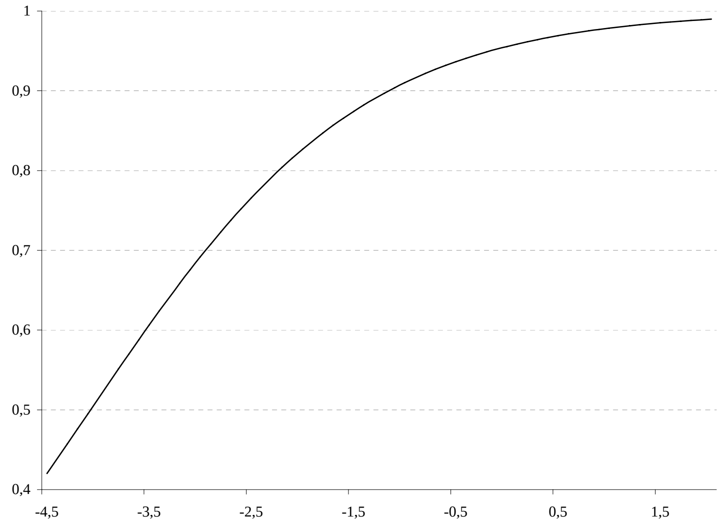

Figure 2: Probability of promotion according to the ability at the end of the fifth grade.

where n_0 denotes the number of pupils who repeat their fifth grade.

To compute the lower bounds of Δ^{TT}, $h(.)$ was first estimated using a kernel estimator.[24] The estimator of $\underline{\Delta}_1^{TT}$ is the empirical analog of (4.4):

$$\widehat{\underline{\Delta}_1^{TT}} = \frac{1}{n_0}\left[\sum_{i/D_i=0} Y_{1i}(0) - \sum_{i/D_i=1} \frac{Q(Y_i;\widehat{\beta})}{1-Q(Y_i;\widehat{\beta})}\widehat{h}(Y_i)\right],$$

The second lower bound is estimated by using a kernel estimator on $g(z) = E(h(Y)|Z = z, D_0 = 1)$ and reporting it in the empirical analog of (4.6):

$$\widehat{\underline{\Delta}_2^{TT}} = \frac{1}{n_0}\left[\sum_{i/D_i=0} Y_{1i}(0) - \sum_{i/D_i=1} \widehat{g}(Z_i)\right],$$

where $\widehat{g}(.)$ is the nonparametric estimator of $g(.)$.

The results are displayed in table 3. Under the assumption of a fully valid instrument, the interval only ranges positive values, so that grade retention leads to positive short terms effect even in the least favorable case.[25] The pattern is less clear if one weakens the instrumental exclusion restriction into a monotonicity condition. Under the extreme case where grade retention only depends on the third grade test score, this policy would be

[24] A gaussian kernel was chosen, and the bandwidth was estimated by cross validation.

[25] Indeed, the null hypothesis that the lower bound is negative is rejected at 5%.

Estimator	Value	95% Confidence interval
$\widehat{\overline{\Delta}^{TT}}$	1.17 (0.24)	[0.75,1.67]
$\widehat{\underline{\Delta}_1^{TT}}$	0.29 (0.16)	[0.02,0.65]
$\widehat{\underline{\Delta}_2^{TT}}$	-0.43 (0.06)	[-0.53,-0.30]

Standard errors were obtained through bootstrap with 1,000 replications.
Effects are measured in standard deviations terms.

Table 3: Bounds on Δ^{TT} under different assumptions.

harmful in terms of test achievement. This assumption does not seem very credible, though. As emphasized previously, the effects of Y on D_0 is probably much more important than the one of Z. Thus, even in the worst case, the true effect is more likely to be close to $\widehat{\underline{\Delta}_1^{TT}}$, that is to say around zero. In conclusion, and even if uncertainty is rather important,[26] the conclusion on short term effects of grade retention is rather positive. This result is in line with the results of Jacob and Lefgren (2004) for third graders in Chicago, but more optimistic than theirs on the sixth graders. This difference could reflect the opposition on grade retention decision rules in the two cases. Letting teachers and parents decide on the basis of their observation of the students during the whole year, and not on two tests only as in Chicago, may reduce measurement errors on the ability of children. On the other hand, such a discretionary process is likely to favour or penalize systematically some subpopulations of students, no matter of their ability, and thus decrease the efficiency of grade retention. The results suggest that the former effects overcome the latter.

5 Conclusion

This paper considers the issue of endogenous selection with instruments. The key assumption for identification, which contrasts with the usual ones in selection problems, is the independence between instruments and selection, conditional on the dependent variables. A general nonparametric identification result is obtained under a completeness condition. This framework can be applied to a broad class of selection models, including Roy models with an unobserved sector, nonignorable nonresponse or binary models with data taken

[26] This uncertainty is rather due to the endogenous selection on grade retention than on the true effect of the instrument on fifth grade retention. The former effect, which prevents us from recovering the counterfactual progression of retained students, accounts indeed for 55% of the width of the set.

from one response stratum. Set identification is also considered when the conditional independence condition fails. Under weaker conditions of monotonicity indeed, I show that there exists sharp and finite bounds on parameters of interest. This result is used to estimate bounds on the effects of grade retention in France.

The paper raises two challenging issues. First, we may wonder whether the ideas developed here could be adapted to generalized Roy model. In these models, selection depends on prediction on the dependent variable rather than on the dependent variable itself. Thus, the conditional independence condition breaks down but the structure of the model may provide information for point or at least set identification. Second, the sharp upper bounds are obtained on a set of parameters which is rather abstract. Further characterizations of this set appear desirable, for both theoretic and practical reasons.

Appendix

A1. Proofs

Theorem 2.1

By assumption 3 and the definition of $P(.)$,

$$\begin{aligned}
P(D=1|\mathbf{Z})E\left[\frac{1}{P(\mathbf{Y})}|D=1,\mathbf{Z}\right] &= E\left(\frac{D}{P(\mathbf{Y})}\middle|\mathbf{Z}\right) \\
&= E\left(\frac{E(D|\mathbf{Y},\mathbf{Z})}{P(\mathbf{Y})}\middle|\mathbf{Z}\right) \\
&= E\left(\frac{E(D|\mathbf{Y})}{P(\mathbf{Y})}\middle|\mathbf{Z}\right).
\end{aligned}$$

Hence,

$$E\left(\frac{D}{P(\mathbf{Y})}-1\middle|\mathbf{Z}\right)=0 \tag{5.1}$$

By assumption 2, $P(D=1|\mathbf{Z})$ can be identified from the data. Thus, for any $R(.)$, $E[D/R(\mathbf{Y})-1|\mathbf{Z}]$ can be computed from the data. Hence, any candidate for $P(.)$ must satisfy equality (5.1). Now let Q be such a candidate and let $g=P/Q-1$. g is bounded below by -1. Moreover, $Q(.)$ must satisfy $E[D/Q(\mathbf{Y})]=1$, which can also be written as $E[P(\mathbf{Y})/Q(\mathbf{Y})]=1$. This implies that

$$E\left[|g(\mathbf{Y}|\right]\leq E\left[\frac{P(\mathbf{Y})}{Q(\mathbf{Y})}\right]+1<\infty.$$

Hence, $g \in \mathcal{B}$. Moreover,

$$\begin{aligned} 0 &= E\left(\frac{D}{Q(\mathbf{Y})} - 1 \middle| \mathbf{Z}\right) \\ &= E\left(\frac{P(\mathbf{Y})}{Q(\mathbf{Y})} - 1 \middle| \mathbf{Z}\right) \\ &= E\left(g(\mathbf{Y})|\mathbf{Z}\right). \end{aligned} \tag{5.2}$$

This together with assumption 4 imply that $g(\mathbf{Y}) = 0$ a.s., so that $Q(\mathbf{Y}) = P(\mathbf{Y})$ a.s. Thus, $P(.)$ is identified.

To finish the proof, let $f_{D,\mathbf{Y},Z}(.,.,.)$ denote the density of $(D, \mathbf{Y}, Z)$ with respect to an appropriate measure. $f_{D,\mathbf{Y},Z}(1, \mathbf{y}, z)$ is identified by $f_{\mathbf{Y},Z|D=1}(\mathbf{y}, z)P(D = 1)$. Moreover, by assumption 3,

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

Similarly,

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

Thus,

$$f_{D,\mathbf{Y},Z}(0, \mathbf{y}, z) = \left[\frac{1 - P(\mathbf{y})}{P(\mathbf{y})}\right] f_{D,\mathbf{Y},Z}(1, \mathbf{y}, z).$$

Hence, the joint distribution of the data is identified □

Theorem 2.2

Part "if" of the theorem is trivial. To prove the "only if" implication, let us consider a solution $Q(.)$ which belongs to $]0, 1]$. Define also a function $g_{D,\mathbf{Y},Z}$ by

$$g_{D,\mathbf{Y},Z}(d, \mathbf{y}, z) = \left[\frac{1 - Q(\mathbf{y})}{Q(\mathbf{y})}\right]^{1-d} f_{\mathbf{Y},Z|D=1}(\mathbf{y}, z)P(D = 1).$$

$g_{D,\mathbf{Y},Z}$ is a density (with respect to a convenient measure λ), as it is nonnegative and integrates to one. Indeed,

$$\begin{aligned} &\int \left[g_{D,\mathbf{Y},Z}(0, \mathbf{y}, z) + g_{D,\mathbf{Y},Z}(1, \mathbf{y}, z)\right] d\lambda(\mathbf{y}, z) \\ = &\int \frac{f_{\mathbf{Y},Z|D=1}(\mathbf{y}, z)P(D = 1)}{Q(\mathbf{y})} d\lambda(\mathbf{y}, z) \\ = &\ E\left\{E\left[\frac{E(D|\mathbf{Y}, Z)}{Q(\mathbf{Y})}|\mathbf{Z}\right]\right\} \\ = &\ 1. \end{aligned}$$

Moreover,

$$g_{D,\mathbf{Y},Z}(1,\mathbf{y},z) = f_{\mathbf{Y},Z|D=1}(\mathbf{y},z)P(D=1) \tag{5.3}$$

and

$$\begin{aligned} g_{\mathbf{Z}}(\mathbf{z}) &= f_{\mathbf{Z}}(\mathbf{z})\int \frac{f_{\mathbf{Y},Z|D=1}(\mathbf{y},z))P(D=1)}{Q(\mathbf{y})f_{\mathbf{Z}}(\mathbf{z})}dy \\ &= f_{\mathbf{Z}}(\mathbf{z})E\left[\frac{E(D|\mathbf{Y},Z)}{Q(\mathbf{Y})}|\mathbf{Z}\right] \\ &= f_{\mathbf{Z}}(\mathbf{z}). \end{aligned}$$

This last equality, together with (5.3), ensures that $g_{D,\mathbf{Z}}(d,\mathbf{z}) = f_{D,\mathbf{Z}}(d,\mathbf{z})$. Thus, $g_{D,\mathbf{Y},Z}$ is coherent with the observed data. Lastly, because $g_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}|D=1}P(D=1)/Q(\mathbf{y})$, we get after straightforward manipulations:

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

In other words, the corresponding distribution of $(D,\mathbf{Y},Z)$ satisfies the independence condition of assumption 3. To conclude, if there exists a solution $Q(.)$ to equation (2.4) which lies in $]0,1]$, one can rationalize the observed data by a distribution which satisfies the independence condition □

Theorem 2.3

The result uses the following standard result, which is proved for the sake of completeness.

Lemma 5.1 *Let T denote a real random variable and $(h_1,h_2)\in(L^2_T)^2$ be increasing functions. Then cov$(h_1(T),h_2(T))\geq 0$.*

Proof: let (T_1,T_2) denote two independent copies of T. Then, because both h_1 and h_2 are increasing,

$$(h_1(T_1)-h_1(T_2))\times(h_2(T_1)-h_2(T_2))\geq 0.$$

Thus, taking expectation and using the fact that (T_1,T_2) are i.i.d, we get

$$2\left\{E\left[h_1(T)h_2(T)\right]-E\left[h_1(T)\right]E\left[h_2(T)\right]\right\}\geq 0.$$

The result follows □

a) By lemma 5.1 and assumption 6,

$$\text{cov}(h(\mathbf{Y}),P(D=1|Y,\mathbf{Z}')|\mathbf{Z}')\geq 0.$$

Thus,

$$E(h(\mathbf{Y})|\mathbf{Z}')P(D=1|\mathbf{Z}') \leq E\left(h(\mathbf{Y})D|\mathbf{Z}'\right).$$

This implies that

$$E(h(\mathbf{Y})|\mathbf{Z}') \leq E\left(h(\mathbf{Y})|D=1,\mathbf{Z}'\right).$$

Hence, by integration,

$$E[h(\mathbf{Y})|X] \leq E\left[E\left(h(\mathbf{Y})|D=1,\mathbf{Z}'\right)|X\right].$$

Moreover, this upper bound is sharp because the two terms are identical under the untestable assumption that $D \perp\!\!\!\perp Y|\mathbf{Z}'$.

b) Let $h \in H_Y^2$ and $\psi \in H_Z^1$ be such that $h(\mathbf{Y}) = E[\psi(\mathbf{Z})|D=1,\mathbf{Y}]$. We get

$$\begin{aligned}
E\left[\frac{Dh(\mathbf{Y})}{Q(\mathbf{Y})}|X\right] - E[h(\mathbf{Y})|X] &= E\left[\frac{DE(\psi(\mathbf{Z})|D=1,\mathbf{Y})}{Q(\mathbf{Y})}|X\right] - E[h(\mathbf{Y})|X] \\
&= E\left[\frac{D\psi(\mathbf{Z})}{Q(\mathbf{Y})}|X\right] - E[h(\mathbf{Y})|X] \\
&= E\left[\psi(\mathbf{Z})E\left(\frac{D}{Q(\mathbf{Y})}|\mathbf{Z}\right)|X\right] - E[h(\mathbf{Y})|X] \\
&= E\left[\psi(\mathbf{Z})|X\right] - E[h(\mathbf{Y})|X] \\
&= E\left[E\left(\psi(\mathbf{Z})|\mathbf{Y}\right) - E\left(\psi(\mathbf{Z})|D=1,\mathbf{Y}\right)|X\right].
\end{aligned}$$

Now, because $\psi(X,.)$ and $z \mapsto P(D=1|\mathbf{Y},Z=z)$ are increasing with probability one, we have, similarly to a),

$$E(\psi(\mathbf{Z})|D=1,\mathbf{Y}) \geq E(\psi(\mathbf{Z})|\mathbf{Y}). \tag{5.4}$$

Thus,

$$E\left[\frac{Dh(\mathbf{Y})}{Q(\mathbf{Y})}|X\right] \leq E[h(\mathbf{Y})|X]. \tag{5.5}$$

Moreover, by the previous theorem, if there exists a solution $Q(.)$ to equation (2.4) which lies in $]0,1]$, one cannot reject that (5.4) and (5.5) are actually equalities. This implies that $E[Dh(\mathbf{Y})/Q(\mathbf{Y})|X]$ is a sharp lower bound of $E[h(\mathbf{Y})|X]$.

c) If $D \perp\!\!\!\perp (\mathbf{Y},\mathbf{Z}')$, by independence,

$$E\left[E(h(\mathbf{Y})|D=1,\mathbf{Z}')|X\right] = E\left[E(h(\mathbf{Y})|\mathbf{Z})|X\right] = E\left[h(\mathbf{Y})|X\right].$$

Moreover, because $P(D=1)$ is a solution to (2.4), $Q(\mathbf{Y}) = P(D=1)$, so that

$$E\left[\frac{Dh(\mathbf{Y})}{Q(\mathbf{Y})}|X\right] = E(h(\mathbf{Y})|X).$$

Now, if $\mathbf{Y} = \mathbf{Z}'$,

$$E\left[E(h(\mathbf{Y})|D=1,\mathbf{Z}')|X\right] = E\left[h(\mathbf{Y})|X\right].$$

Moreover, because $\mathbf{Y} = \mathbf{Z}$, equation (2.4) is equivalent to $Q(\mathbf{Y}) = P(D=1|\mathbf{Y})$. Hence,

$$E\left[\frac{Dh(\mathbf{Y})}{Q(\mathbf{Y})}|X\right] = E\left[\frac{E(D|\mathbf{Y})h(\mathbf{Y})}{Q(\mathbf{Y})}|X\right] = E\left[h(\mathbf{Y})|X\right] \;\square$$

Lemma 2.1

We have

$$E(\psi(\mathbf{Z})|D=1,\mathbf{Y}) = \int \psi(z)dF_{\mathbf{Z}|\mathbf{Y},D=1}(z).$$

Because $\psi(x,.)$ is increasing for all x, there exists a positive measure μ_x such that for all $z \leq z_1$,

$$\psi(x,z_1) - \psi(x,z) = \int_z^{z_1} d\mu_x(u).$$

Thus, for all $y = (x,y)$ and all $M \in \mathbb{R}$,

$$\begin{aligned} E(\psi(\mathbf{Z})|D=1,\mathbf{Y}=y) \;=\; & \int_M^{\infty}\int_M^z d\mu_x(u)dF_{\mathbf{Z}|\mathbf{Y}=y,D=1}(z) - \int_{-\infty}^M\int_z^M d\mu_x(u)dF_{\mathbf{Z}|\mathbf{Y}=y,D=1}(z) \\ & +2\psi(x,M). \end{aligned}$$

Hence, by Fubini's theorem on nonnegative functions,

$$E(\psi(\mathbf{Z})|D=1,\mathbf{Y}=y) = \int_M^{\infty}(1-F_{\mathbf{Z}|\mathbf{Y}=y,D=1}(u))d\mu_x(u) - \int_{-\infty}^M F_{\mathbf{Z}|\mathbf{Y}=y,D=1}(u)d\mu_x(u) + 2\psi(x,M).$$

Consequently, we get, for all x and $y \leq y_1$,

$$\begin{aligned} & E(\psi(\mathbf{Z})|D=1,X=x_0,Y=y_1) - E(\psi(\mathbf{Z})|D=1,X=x_0,Y=y) \\ = & \int [F_{\mathbf{Z}|D=1,X=x_0,Y=y}(u) - F_{\mathbf{Z}|D=1,X=x_0,Y=y_1}(u)]d\mu_x(u). \end{aligned}$$

By assumption, the right hand side is nonnegative. The result follows.

Theorem 2.4

a) β_0 satisfies

$$E\left(\frac{D\mathbf{Z}}{F(\mathbf{Y}'\beta_0)}\right) = E\left(\frac{\mathbf{Z}}{F(\mathbf{Y}'\beta_0)}E(D|\mathbf{Z},\mathbf{Y})\right) = E\left(\frac{\mathbf{Z}}{F(\mathbf{Y}'\beta_0)}E(D|\mathbf{Y})\right) = E(\mathbf{Z}),$$

where the second equality stems from assumption 3. Local identification only requires that the differential of $g : \beta \rightarrow E(D\mathbf{Z}/F(\mathbf{Y}'\beta))$ is full rank at $\beta = \beta_0$. This differential is $-E(D\mathbf{Z}\mathbf{Y}'F'(\mathbf{Y}'\beta_0)/F^2(\mathbf{Y}'\beta_0))$, so the result follows from assumption 4'.

b) Suppose that there exists β such that

$$E\left(\frac{D\mathbf{Z}}{F(\mathbf{Y}'\beta)}\right) = E(\mathbf{Z}) = E\left(\frac{D\mathbf{Z}}{F(\mathbf{Y}'\beta_0)}\right). \tag{5.6}$$

Then

$$E\left(\left(\frac{1}{F(\mathbf{Y}'\beta_0)} - \frac{1}{F(\mathbf{Y}'\beta)}\right)\mathbf{Z}(\beta_0 - \beta)\middle| D = 1\right) = 0.$$

Thus

$$E\left(\left(\frac{1}{F(\mathbf{Y}'\beta_0)} - \frac{1}{F(\mathbf{Y}'\beta)}\right)E(\mathbf{Z}|\mathbf{Y}, D = 1)(\beta_0 - \beta)\middle| D = 1\right) = 0.$$

Now, by assumption 4",

$$E(\mathbf{Z}|\mathbf{Y}, D = 1) = \begin{pmatrix} I_r & 0 \\ \Gamma_1 & \Gamma_2 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} \equiv \Gamma\mathbf{Y},$$

where I_r is the identity matrix of size r. Moreover, because Γ_2 is full rank, Γ is also full rank. Hence

$$E\left(\left(\frac{1}{F(\mathbf{Y}'\beta_0)} - \frac{1}{F(\mathbf{Y}'\beta)}\right)(\mathbf{Y}'\beta_0 - \mathbf{Y}'\beta)\middle| D = 1\right) = 0.$$

Because F is strictly increasing, for any $x \neq y$, $(x - y)(1/F(x) - 1/F(y)) < 0$, so that

$$\mathbf{Y}'\beta_0 = \mathbf{Y}'\beta \quad P^{D=1} \text{ a.s.}$$

Because $\mathbf{Y}$ is linearly independent almost surely, $\beta = \beta_0$ □

Theorem 3.1.

As Horowitz and Lee (2007), we adapt the proof of theorem 2 of Bissantz et al. (2004). By definition of $\widehat{f}$,

$$\max\left(\left\|\widehat{T}(\widehat{f}) - 1\right\|^2, \alpha_n\left\|\widehat{f}\right\|^2\right) \leq \left\|\widehat{T}(\widehat{f}) - 1\right\|^2 + \alpha_n\left\|\widehat{f}\right\|^2 \leq \left\|\widehat{T}(f) - 1\right\|^2 + \alpha_n\left\|f\right\|^2 \tag{5.7}$$

Because $E(\left\|\widehat{T}(f) - 1\right\|^2) = O(\delta_n)$ (see e.g. Györfi et al., 2002) and $\delta_n/\alpha_n \to 0$, we get

$$\limsup E(\left\|\widehat{f}\right\|^2) \leq \left\|f\right\|.$$

Inequalities (5.7) and $\delta_n/\alpha_n \to 0$ also implies that $E(\left\|\widehat{T}(\widehat{f}) - 1\right\|^2) \to 0$. Besides, D_M is weakly closed as a closed and convex set (see Bissantz et al., 2004). Moreover, for all $\phi \in D_M$, by Jensen's inequality,

$$T(\phi)^2 \leq E(\phi(Y)^2 \,|Z).$$

Hence,

$$\begin{aligned}\left\|T(\phi)\right\|^2 &\leq \int \left[\int \phi(y)^2 f_{Y|Z}(y|z)dy\right] dz \\ &\leq \int \phi(y)^2 \left[\int \frac{f_{Z|Y}(z|y)f_Y(y)}{f_Z(z)} dz\right] dy \\ &\leq \frac{\sup_{y\in[0,1]} f_Y(y)}{\inf_{z\in[0,1]} f_Z(z)} \int \phi(y)^2 dy,\end{aligned}$$

where the second inequality follows from Fubini's theorem and Bayes theorem. Hence, by assumption 8-(b), there exists $A < +\infty$ such that

$$\left\|T(\phi)\right\|^2 \leq A\left\|\phi\right\|^2.$$

This inequality and the linearity of T proves that it is continuous. Hence, T is weakly continuous. This and the fact that D_M is weakly closed ensures that T is weakly sequentially closed (see Bissantz et al., 2004). Consequently, we can apply the end of the proof of theorem 2 of Bissantz et al. (2004), and the result follows □

Corollary 3.2

By the triangular inequality,

$$|\widehat{\theta} - \theta| \leq \frac{1}{n}\sum_{i=1}^{n} D_i |g(Y_i^*, Z_i)||\widehat{f}^{-i}(Y_i^*) - f(Y_i^*)| + \left|\frac{1}{n}\sum_{i=1}^{n} D_i g(Y_i^*, Z_i) f(Y_i^*) - \theta\right| \qquad (5.8)$$

By assumption 8, $|D_i f(Y_i^*)| \leq M$. Hence, $E[|D_i g(Y_i^*, Z_i) f(Y_i^*)|^2] < \infty$ and by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} D_i g(Y_i^*, Z_i) f(Y_i^*) \xrightarrow{L^2} E(Dg(Y^*, Z)f(Y^*)).$$

Moreover,

$$\begin{aligned}E(Dg(Y^*, Z)f(Y^*)) &= E(Dg(Y, Z)f(Y)) \\ &= E(E(D|Y, Z)g(Y, Z)f(Y)) \\ &= \theta.\end{aligned}$$

Thus the second term of the r.h.s. of (5.8) tends to zero in quadratic mean.

Now, because the $(\widehat{f}^{-i}(Y_i^*))_i$ are identically distributed, the first term T_1 of the r.h.s. of (5.8) satisfies

$$\begin{aligned}E\left(|T_1|\right) &= E\left(D_1|g(Y_1^*, Z_1)||\widehat{f}^{-1}(Y_1^*) - f(Y_1^*)|\right) \\ &\leq \sqrt{E\left(|g(Y_1^*, Z_1)|^2\right) E\left(|\widehat{f}^{-1}(Y_1^*) - f(Y_1^*)|^2\right)},\end{aligned}$$

by the Cauchy-Schwartz inequality. Now, by independence between Y_1^* and $\widehat{f}^{-1}$,

$$E(|\widehat{f}^{-1}(Y_1^*) - f(Y_1^*)|^2) \leq \sup_{y\in[0,1]} f_Y(y) E\left(||\widehat{f}^{-1} - f||^2\right).$$

Thus, the left hand side tends to zero by theorem 3.1. As a consequence, $E(|T_1|)$ also tends to zero. This yields the announced result □

A2. Discussion on identification in system (2.1)

In this appendix, we provide sufficient conditions for assumptions 3 and 4 to hold in system (2.1). Lemma 5.2 shows that the conditional independence condition holds whenever η is independent of (Z, ε). Lemma 5.3 presents sufficient conditions for the completeness condition.

Lemma 5.2 *Suppose that $\eta \perp\!\!\!\perp (Z, \varepsilon)|X$. Then assumption 3 holds.*

Proof: let $H_{x,y} = \{u/h(x, y, u) = 1\}$ and $G(x, y, z) = \{u \in \mathbb{R}/g(x, z, u) = y\}$. We get, for all (x, y, z),

$$\begin{aligned} P(D = 1|X = x, Y = y, Z = z) &= P(\eta \in H_{x,y}|X = x, Y = y, Z = z) \\ &= P(\eta \in H_{x,y}|X = x, \varepsilon \in G(x, y, z), Z = z) \\ &= P(\eta \in H_{x,y}|X = x) \\ &= P(\eta \in H_{x,y}|X = x, Y = y) \\ &= P(D = 1|X = x, Y = y). \end{aligned}$$

Thus, assumption 3 holds □

Lemma 5.3 *Suppose that*

1. *(additive decomposition) $g(x, z, \varepsilon) = \mu(\nu(x, z) + \varepsilon)$ and $Z \perp\!\!\!\perp \varepsilon|X$.*

2. *(large support) For P^X-almost all x, the measure of $\nu(x, Z)$ is continuous with respect to the Lebesgue measure and the support of $\nu(x, Z)$ conditional on $X = x$ is $\mathbb{R}$ almost surely.*

3. *(regularity conditions) The distribution of ε conditional on X admits almost surely a continuous density $f_{\varepsilon|X}$ with respect to the Lebesgue measure. Moreover, $f_{\varepsilon|X}(0) > 0$ and there exists $\alpha > 2$ such that $t \mapsto t^\alpha f_{\varepsilon|X}(t)$ is bounded. Lastly, the conditional characteristic function of ε does not vanish and is infinitely often differentiable in $\mathbb{R}\backslash A$ for some finite set A.*

Then $\boldsymbol{Y}$ *is* $\mathcal{B}-$*complete for* $\boldsymbol{Z}$.

The additive decomposition and the large support condition are identical to the assumptions A1 and A2 made by d'Haultfœuille (2008) to study completeness.[27] In this framework, the regularity conditions imposed here are not sufficient in general to achieve completeness (see d'Haultfœuille, 2008, proposition 2.2). Actually, they are hardly stronger than the one needed to achieve bounded completeness, namely, the zero freeness of the conditional characteristic function of ε (see d'Haultfœuille, 2008, theorem 2.1). Hence, in this framework at least, $\mathcal{B}$-completeness appears to be almost equivalent to bounded completeness.

Proof: in the following, the dependence in X is omitted for the sake of simplicity. Hence, all statements must be understood conditional on this variable. We proceed in three steps.

1. First, we prove that there exists positive c_1, c_2 and $0 < \alpha' < \alpha - 2$ such that

$$c_1 \leq (f_\varepsilon \star g)(x) \times (1 + |x|)^{\alpha'+1} \leq c_2, \tag{5.9}$$

where g denote the density of an α'-stable distribution of characteristic function $\exp(-|t|^{\alpha'})$ and $\star$ denotes the convolution product.

To prove (5.9), note that g satisfies, for well chosen $c < C$ (see e.g. Mattner 1992, p. 146),

$$c \leq g(x) \times (1 + |x|)^{\alpha'+1} \leq C \tag{5.10}$$

Let $I = [a, b] \subset [-1, 1]$ denote an interval such that $\inf_{x\in I} f_\varepsilon(x) = m > 0$ (such an interval exists by the regularity conditions). For all x and $t \in I$,

$$\begin{aligned} 1 + |x - t| &\leq 1 + \max(|x - a|, |x - b|) \\ &\leq 1 + |x| + \max(|a|, |b|) \\ &\leq 2(1 + |x|). \end{aligned}$$

[27] Note that the additive decomposition considered here still encompasses many nonlinear models, beyond the nonparametric additive models for which $\mu(x) = x$. Usual ordered choice models correspond to $\mu(x) = \sum_{k=1}^{K} k\mathbb{1}_{]\alpha_{k-1};\alpha_k]}(x)$ (where $\mathbb{1}_A(x) = 1$ if $x \in A$, 0 otherwise) for some given thresholds $\alpha_0 = -\infty < \alpha_1 < ... < \alpha_K = +\infty$. Count data models can also be handled by taking $\mu(x) = [\exp(x)]$ (where $[a]$ denotes the integer part of a). Simple tobit models correspond to $\mu(x) = \max(0, x)$. Lastly, duration models like the accelerated failure time model or the proportional hazard model also fit in this framework. The first corresponds to $\mu(x) = \exp(x)$, while in the second, μ is an unknown increasing function and $-\varepsilon$ is distributed according to a Gompertz distribution.

Thus,

$$\begin{aligned}
(f_\varepsilon \star g)(x) &\geq \int_I f_\varepsilon(t) g(x-t)dt \\
&\geq mc \int_I \frac{dt}{(1+|x-t|)^{\alpha'+1}} \\
&\geq \frac{mc(b-a)}{2^{\alpha'+1}(1+x)^{\alpha'+1}}.
\end{aligned}$$

This shows the first inequality of (5.9). To prove the second one, remark that by the regularity conditions, there exists M such that

$$(1+|t|)^\alpha f_\varepsilon(t) \leq M \tag{5.11}$$

Moreover, for all $x \geq 0$ and $t < x/2$, we get $1+|x-t| \geq (1+x)/2$. Thus, using both (5.10) and (5.11), we get

$$\begin{aligned}
\int_{-\infty}^{x/2} f_\varepsilon(t) g(x-t)dt &\leq \frac{2^{\alpha'+1}MC}{(1+x)^{\alpha'+1}} \int_{-\infty}^{x/2} \frac{dt}{(1+|t|)^\alpha} \\
&\leq \frac{2^{\alpha'+1}MC}{(1+x)^{\alpha'+1}} 2 \int_{-\infty}^{0} \frac{dt}{(1-t)^\alpha} \\
&\leq \frac{2^{\alpha'+2}MC}{(\alpha-1)(1+|x|)^{\alpha'+1}}. \qquad (5.12)
\end{aligned}$$

Moreover, because $g(x-t) \leq C$ and $\alpha - 1 > \alpha' + 1$,

$$\begin{aligned}
\int_{x/2}^{+\infty} f_\varepsilon(t) g(x-t)dt &\leq MC \int_{x/2}^{+\infty} \frac{dt}{(1+t)^\alpha} \\
&\leq \frac{2^{\alpha-1}MC}{(1+x)^{\alpha-1}} \\
&\leq \frac{2^{\alpha-1}MC}{(1+x)^{\alpha'+1}}.
\end{aligned}$$

This, together with (5.12), shows that for all $x \geq 0$, there exists a constant C' such that $(f_\varepsilon \star g)(x) \times (1+|x|)^{\alpha'+1} \leq C'$. The same reasoning can be applied to any $x < 0$, and the second inequality of (5.9) follows.

2. Now let us show that for any $h \in \mathcal{B}$ such that $E\left[h(Y)|Z\right] = 0$ a.s., we get, almost everywhere (a.e. for short),

$$(h \circ \mu) \star \phi = 0, \tag{5.13}$$

where $\phi = f_{-\varepsilon} \star g$.

By definition of $\mathcal{B}$, there exists K such that $h(Y) \geq K$ almost surely. Let $\widetilde{h}(u) = h(\mu(u)) - K$. Using the additive decomposition, we get

$$\begin{array}{rcl} \mathbb{E}[h(Y) - K|Z] & = & \mathbb{E}[\widetilde{h}(\nu(Z) + \varepsilon)|Z] \\ & = & \displaystyle\int \widetilde{h}(\nu(Z) + u) f_\varepsilon(u) du \\ & = & \displaystyle\int \widetilde{h}(u) f_{-\varepsilon}(\nu(Z) - u) dt \end{array}$$

This implies, by the large support assumption, that

$$\mathbb{E}[h(Y)|Z] = 0 \text{ a.s.} \Leftrightarrow \int \widetilde{h}(u) f_{-\varepsilon}(t - u) dt = -K \text{ for a.e. } t. \tag{5.14}$$

In other words, $\widetilde{h} \star f_{-\varepsilon} = -K$. Let α' and g be defined as previously. We get, a.e.,

$$\left(\widetilde{h} \star f_{-\varepsilon}\right) \star g = -K.$$

Because $\widetilde{h}, f_{-\varepsilon}$ and g are nonnegative functions, we can apply Fubini's theorem, so that $\widetilde{h} \star (f_{-\varepsilon} \star g) = -K$ a.e. Equation (5.13) follows.

3. Finally, let us prove that the location family generated by ϕ is complete. This proves the result because then, $h \circ \mu = 0$ a.e. and thus $h(Y) = 0$ almost surely. For this purpose, we check the conditions of theorem 1.1 of Mattner (1992). First, ϕ satisfies condition (i) of this theorem by (5.9) and proposition 1.2 of Mattner (1992). Second, the characteristic function Ψ_ϕ corresponding to the density ϕ writes as

$$\Psi_\phi(t) = \Psi_\varepsilon(-t) \times \exp(-|t|^{\alpha'}) \tag{5.15}$$

where Ψ_ε denotes the characteristic function of ε. Thus, by the regularity conditions, Ψ_ϕ is infinitely differentiable on $\mathbb{R}\backslash(A \cup \{0\})$ and condition (ii) of Mattner's theorem holds. Lastly, by (5.15) and the regularity conditions once more, Ψ_ϕ does not vanish anywhere. Thus theorem 1.1 in Mattner (1992) can be applied, and the proof is finished □

References

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

Baker, S. G. and Laird, N. M. (1988), 'Regression analysis for categorical variables with outcome subject to nonignorable nonresponse', *Journal of the American Statistical Association* **83**, 62–69.

Bissantz, N., Hohage, T. and Munk, A. (2004), 'Consistency and rates of convergence of nonlinear tikhonov regularization with random noise', *Inverse Problems* **20**, 1773–1789.

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

Carrasco, M., Florens, J. P. and Renault, E. (2006), Linear inverse problems and structural econometrics: Estimation based on spectral decomposition and regularization, *in* J. J. Heckman and E. E. Leamer, eds, 'Handbook of Econometrics', Vol. 6, North Holland.

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

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

Chen, X. and Hu, Y. (2006), Identification and inference of nonlinear models using two samples with arbitrary measurement errors. Cowles foundation discussion paper no. 1590.

Cosnefroy, O. and Rocher, T. (2004), 'Le redoublement au cours de la scolarité obligatoire : nouvelles analyses, mêmes constats', *Education et Formation* **70**, 73–82.

Crahaye, M. (1996), *Peut-on lutter contre l'échec scolaire ?*, De Boeck.

Darolles, S., Florens, J. P. and Renault, E. (2002), Nonparametric instrumental regression. Working paper 05-2002 CRDE.

Deville, J. C. (2002), La correction de la non-réponse par calage généralisé, *in* 'Actes des Journées de Méthodologie Statistique 2002', INSEE, pp. 4–20.

d'Haultfœuille, X. (2008), On the completeness condition in nonparametric instrumental regression. CREST, working paper.

Florens, J. P., Heckman, J. J., Meghir, C. and Vytlacil, E. (2003), Instrumental variables, local instrumental variables and control functions. IDEI Working Paper 249.

Gagliardini, P. and Scaillet, O. (2006), Tikhonov regularization for functional minimum distance estimators. Working Paper.

Grogger, J. T. and Carson, R. T. (1991), 'Models for truncated counts', *Journal of Applied Econometrics* **6**, 225–238.

Györfi, L., Kohler, M., Kryzak, A. and Walk, H. (2002), *A Distribution-Free Theory of Nonparametric Regression*, New York: Springer.

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

Haurin, D. R. and Sridhar, K. S. (2003), 'The impact of local unemployment rates on reservation wages and the duration of search for a job', *Applied Economics* **35**, 1469–1475.

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

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

Hellerstein, J. K. and Imbens, G. W. (1999), 'Imposing moment restrictions from auxiliary data by weighting', *The Review of Economics and Statistics* **81**, 1–14.

Hemvanich, S. (2004), The general missingness problems and estimation in discrete choice models. Working Paper.

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

Holmes, T. (1989), Grade level retention effects : A meta-analysis of research studies, *in* L. A. Sheppard and M. L. Smith, eds, 'Flunking Grades. Research and Policies on Retention', New York, The Falmer Press, pp. 16–33.

Horowitz, J. L. and Lee, S. (2007), 'Nonparametric instrumental variables estimation of a quantile regression model', *Econometrica* **75**, 1191–1208.

Horvitz, D. G. and Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**, 663–685.

Hu, Y. and Schennach, S. (2008), 'Instrumental variable treatment of nonclassical measurement error models', *Econometrica* **76**, 195–216.

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

Imbens, G. W. and Lancaster, T. (1994), 'Combining micro and macro data in microeconometric models', *Review of Economic Studies* **61**, 655–680.

Jacob, B. A. (2005), 'Accountability, incentives and behavior: Evidence from school reform in chicago', *Journal of Public Economics* **89**, 761–796.

Jacob, B. A. and Lefgren, L. (2004), 'Remedial education and student achievement: A regression-discontinuity analysis', *Review of Economics and Statistics* **86**, 226–244.

Jimerson, S. (2001), 'Meta-analysis of grade retention research: Implications for practice in the 21st century', *School Psychology Review* **30**, 420–437.

Jimerson, S., Anderson, G. E. and Whipple, A. D. (2002), 'Winning the battle and losing the war: Examining the relationship between grade retention and dropping out of high school', *Psychology in the Schools* **39**, 441–457.

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

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

Lorence, J. (2006), 'Retention and academic achievement research revisited from an united states perspective', *International Education Journal* **7**, 731–777.

Manacorda, M. (2007), The cost of grade retention. Working paper.

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

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

Manski, C. F. and Pepper, J. V. (2000), 'Monotone instrumental variables: With an application to the returns to schooling', *Econometrica* **68**, 997–1010.

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

Mattner, L. (1993), 'Some incomplete but boundedly complete location families', *Annals of Statistics* **21**, 2158–2162.

Nevo, A. (2002), 'Using weights to adjust for sample selection when auxiliary information is available', *Journal of Business and Economics Statistics* **21**, 43–52.

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

Park, T. and Brown, M. B. (1994), 'Models for categorical data with nonignorable nonresponse', *Journal of the American Statistical Association* **89**, 44–52.

Ramalho, E. A. and Smith, R. J. (2007), Discrete choice nonresponse. CEMMAP working paper.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999), 'Adjusting for nonignorable drop-out using semiparametric nonresponse models', *Journal of the American Statistical Association* **94**, 1096–1120.

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

Troncin, T. (2005), Le redoublement : radiographie d'une décision à la recherche de sa légitimité. PhD Thesis, available at http://tel.archives-ouvertes.fr/docs/00/14/05/31/PDF/05076.pdf.

Wooldridge, J. (2007), 'Inverse probability weighted estimation for general missing data problems', *Journal of Econometrics* **141**, 1281–1301.