Another Instrumental Method for Dealing with Endogenous Selection*

Xavier d'Haultfoeuille(*)
INSEE, Division Marchés et Stratégies d'Entreprise

Abstract

This paper considers a new method for dealing with endogenous selection. When selection depends directly on the dependent variable, the usual instrumental strategy based on the independence between the outcome and the instruments is likely to fail. On the other hand, it may be possible in this case to find an instrument which is independent of the selection variable, conditional on the outcome. This strategy 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, sharp bounds on parameters of interest can be obtained 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.

JEL classification numbers: C14, C21, C25.

*I would like to thank especially Jean-Claude Deville, for inspiring me this project, and Stéphane Bonhomme, for his fruitful suggestions. I also acknowledge three anonymous referees, 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.

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. [36]). 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. [33]), or the unconfoundedness assumption in the treatment effect literature (see for instance [26]). 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. [18] on tobit models, [1] or [19] on treatment effects). However, this assumption does not point identify the distribution of the outcome in general (see [36]). 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, [5] obtains identification at the infinity by imposing a linear structure. Lastly, using an appealing composite strategy, [32] 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.[1]

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 [6], [21], [42] and [43].[2] Apart from 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.

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

[2] 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. [6] and [43] propose sufficient conditions for identification in parametric models, and [21] studies identification when the support of the outcome is finite. We extend his result to a general situation here.

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, [3], [25] and [41]). Under this hypothesis and the conditional independence assumption, the joint distribution of the data is identified nonparametrically.[3] The key point is that it suffices, in this framework, to recover the probability of selection conditional on the outcome. This is similar to the unconfoundedness situation, where the problem reduces to identifying the propensity score. However, whereas the identification of the propensity score is trivial in the latter case, the conditional probability is harder to retrieve in the former. I show that this function satisfies an integral inverse problem, whose solution is unique under the completeness condition.

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 [11]). It is also related to the literature on auxiliary information, which has been developed either for efficiency reasons (see [20] or [27]) or, as here, to provide identification (see [20] and [40]). 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

[3] In particular, the marginal effect of the instrument on the outcome, or the effet of the selection variable on the outcome, are identified.

[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 [37] (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 [16] or [4]. 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 [20], [24], [40] or [45].

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 [29] for third graders in Chicago.

The rest of the paper is structured as follows. Section one is devoted to identification issues. Estimation methods are described in section two. The application to grade retention is presented in section three.[5]

[5]All proofs of the results are available in the CREST working paper.

1 Identification

1.1 The setting and main result

Let D, Y and Z denote respectively the selection dummy variable, the dependent variable, and the instruments. The first assumptions set the selection problem.

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

Assumption 2 *The distribution of 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 (Y, Z) are missing when $D = 0$. In this latter situation, auxiliary information on 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 Z may be overly strong, and we will see in Subsection 1.3 that it can be weakened to the knowledge of moments of Z, at the price of imposing parametric restrictions.

Assumptions 1 and 2 alone do not enable to point identify the distribution of (D, 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 assume here the following condition:[6]

Assumption 3 $D \perp\!\!\!\perp Z \,|\, Y$.

This assumption has also been made by [6], [43], [21] and [42] in the framework of non-response. It is also a particular case of assumption (41) of [35]. 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, Y, Z) follow the nonparametric system

$$\begin{cases} Y &= \varphi(Z, \varepsilon) \\ D &= \psi(Y, \eta). \end{cases} \tag{1.1}$$

[6] We could refine this assumption by supposing that $D \perp\!\!\!\perp Z \,|\, Y, X$ where X denote covariates whose distribution is identified. All the subsequent analysis would then hold conditional on X. We do not introduce such covariates until Subsection 1.4 for the ease of notations.

In this setting, we have the following result.

Proposition 1.1 *Suppose that system (1.1) holds with* $\eta \perp\!\!\!\perp (Z, \varepsilon)$. *Then assumption 3 holds.*

By letting $\psi(y, u) = \mathbb{1}\{u \leq P(D = 1|Y = y)\}$, we can suppose without loss of generality that η is independent of Y.[7] The exclusion restriction amounts to reinforce this into a conditional independence between η and (Y, Z).

As indicated previously, a dependence condition between Y and 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 $h(Y)$ 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 $g \in \mathcal{B}$,*

$$\Big(E(g(Y)|Z) = 0 \quad a.s.\Big) \Longrightarrow \Big(g(Y) = 0 \quad a.s.\Big). \tag{1.2}$$

Assumption 4 is weaker than the usual completeness condition, for which condition (1.2) must hold for any $g \in L^1_Y$, but stronger than bounded completeness, for which condition (1.2) must hold for bounded functions h only (see e.g. [39] 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 [41], [10]) and in nonclassical measurement error problems (see [7] and [25]),[8] while the bounded completeness condition has been used for instance by [3].

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 $\text{rank}(M) = s$, where M is the matrix of typical element $P(Y = y_i|Z = z_j)$ (see [41]). Hence, the support of Z must be at least as rich as that of Y ($t \geq s$) and the dependence between the two variables must be strong enough for s distinct conditional distributions $P(Y = .|Z = z_j)$ 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

[7] In this case, ψ is not necessarily structural.

[8] Indeed, assumption 2.4 of [7] and assumption 2 of [25] are equivalent, under technical conditions, to a completeness condition.

the density of Y conditional on Z belongs to an exponential family (see [41]). We show here that assumption 4 is also satisfied under an additive decomposition, a large support assumption and technical restrictions on ε in system (1.1).

Proposition 1.2 *Consider system (1.1) with $Y \in \mathbb{R}$ and suppose that*

1. *(additive decomposition) $\varphi(z, \varepsilon) = \mu(\nu(z) + \varepsilon)$ and $Z \perp\!\!\!\perp \varepsilon$.*
2. *(large support) The measure of $\nu(Z)$ is continuous with respect to the Lebesgue measure and the support of $\nu(Z)$ is $\mathbb{R}$ almost surely.*
3. *(regularity conditions on ε) The distribution of ε admits a continuous density f_ε with respect to the Lebesgue measure. Moreover, $f_\varepsilon(0) > 0$ and there exists $\alpha > 2$ such that $t \mapsto t^\alpha f_\varepsilon(t)$ 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 Y is $\mathcal{B}-$complete for Z.

The additive decomposition and the large support condition are identical to the assumptions A1 and A2 made by [12] to study completeness and bounded completeness.[9] The regularity conditions on ε are satisfied for many distributions such as the normal, the student with degrees of freedom greater than one[10] or the stable distributions with characteristic exponent greater than one. Interestingly, these regularity conditions are hardly stronger than the one needed to achieve bounded completeness, namely, the zero freeness of the conditional characteristic function of ε (see [12], Theorem 2.1). Hence, in this framework at least, $\mathcal{B}$-completeness appears to be almost equivalent to bounded completeness.

Because identification is based on inverse probability weighted moment conditions, we also suppose that the conditional probability $P(Y) \equiv P(D = 1|Y)$ 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 Y, as in truncation

[9] The additive decomposition considered here 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 $\alpha_0 = -\infty < \alpha_1 < ... < \alpha_K = +\infty$. Simple tobit models correspond to $\mu(x) = \max(0, x)$. Duration models like the accelerated failure time model (for which $\mu(x) = \exp(x)$) or the proportional hazard model (for which μ is an unknown increasing function and $-\varepsilon$ is distributed according to a Gompertz distribution) also fit in this framework.

[10] See e.g. [38] for a proof that the conditions on the characteristic function of student distributions are indeed satisfied.

models for instance where $D = \mathbb{1}\{Y \geq s\}$, s denoting a fixed threshold. It also fails for random truncation models $D = \mathbb{1}\{Y \geq \eta\}$ if η is strictly greater than the infimum of Y. This would be the case, in example 2 below for instance, if the reservation wage η of individuals is always greater than the lowest potential wage Y.

Assumption 5 *$P(Y) > 0$ almost surely.*

Theorem 1.3 *Suppose that assumptions 1-5 hold. Then the distribution of (D, Y, Z) is identified.*

Basically, the result stems from the fact that under assumption 3 and 4, the equation in Q

$$E\left(\frac{D}{Q(Y)}\middle| Z\right) = 1 \tag{1.3}$$

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, Y, Z). We now present several potential applications of this framework.

Example 1: nonignorable nonresponse

In this case, an outcome Y is observed only if the individual answers to the survey or to a given question in the questionnaire ($D = 1$). The aim is to recover the full distribution of Y, given that nonresponse directly depends on Y. For instance, consider the variable $Y = 1$ if the individual has used drugs at least once during the month, 0 otherwise. Accepting to answer to the question "have you used drugs at least once during the last month?" is likely to depend on the answer Y itself. The method can be applied if an instrument affects Y but not directly D. In the drugs example, local drug prices affect the fact of using drugs but are unlikely to play directly on response on drug use. Note that in this example where Y is binary, the completeness condition is easy to check, since it is equivalent to a nonzero correlation between Y and the instrument.

Example 2: Roy model with an unobserved sector

In this example, Y (resp. η) denotes 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 η is never observed. Thus, in this case $D = \mathbb{1}\{Y \geq \eta\}$.[11] For instance, Y may represent the potential wage of an individual, which is observed only

[11] Following the previous discussion, assumption 5 will be satisfied if η can be lower than any value of Y, with a positive probability.

if the person enters the labor market, while η denotes his reservation wage. The aim is to recover the distribution of Y, or the effects of covariates X on Y. The usual exclusion restriction requires the existence a variable which affects η but not Y. On the other hand, the strategy above can be applied if there is an instrument Z which affects the potential wage but not directly the reservation wage, so that η is independent of Z conditional on Y (or conditional on (X, Y) if one adds covariates). A possible example of such an instrument is the local unemployment rate (see [17], for evidence that the local unemployment rate does not affect the reservation wage).[12]

Example 3: Sample from one response stratum

Suppose that a researcher seeks to study the effects of Y on a binary variable D, but Y is observed only for the stratum $D = 1$.[13] 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. In this case D is the dummy variable of being ill, while Y is the dummy of being vaccinated. 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.[14] Once more, the completeness condition is satisfied as soon as the correlation between Y and the instrument is not zero.

This example also covers truncated count data models. In this case, the aim is to recover the effect of Y on an integer valued variable N, given that Y is observed only when $N > 0$.[15] Consider for instance the estimation of the price elasticity of a good through the use of retail data.[16] If we observe the quantities sold N and the sales $N \times Y$, but not directly prices Y, then these prices can be deduced only when the quantities sold are positive. The

[12] No statistical test for completeness conditions has been developed yet in the case where Y is continuous. Thus, assumption 4 has to be maintained in this example. However, one can test implications of assumption 4 by checking for instance that $E(Y|Z)$ is not a constant function.

[13] In this case, Y is a covariate rather than an outcome. The notation Y is maintained however to ensure consistency with assumption 1.

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

[15] Hence, $D = \mathbb{1}\{N > 0\}$ here and recovering $P(N = k|Y)$ for all $k \in \mathbb{N}$ amounts to identify $P(D = 1|Y)$. Note that this example differs from the simple truncation model $D = \mathbb{1}\{Y \geq s\}$ described above. In particular, assumption 5 will hold as soon as $P(N = 0|Y) < 1$ almost surely.

[16] As discussed by [15], 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.

framework can be applied if there is an instrument whose distribution is identified and which affects prices but not directly the demand. Production cost shifters such as prices of the inputs may be good candidates for that.

1.2 Testability

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 (1.3) may have no solution. This is especially clear when (Y, Z) has a finite support. If indeed Y and Z take respectively m_1 and m_2 distinct values, with $m_2 > m_1$, (1.3) 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 (1.3) must be a positive probability, i.e. $Q(y) \in]0, 1]$ for all y.[17] 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 (1.3) 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, for instance, assumption 3 is rejected. 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 (1.3) does exist, one can expect that assumption 3 cannot be rejected, since intuitively, this equation makes use of all the available information. Theorem 1.4 formalizes this idea.

Theorem 1.4 *Suppose that assumption 1, 2 and 5 hold. Then assumption 3 can be rejected if and only if there exists no solution Q of equation (1.3) which belongs to $]0, 1]$.*

When Y is discrete and takes values in $\{y_1, ..., y_k\}$, a statistical test of assumption 3, under the maintained assumption 4, can be developed as follows. First, we can estime

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

$f = 1/P$ by GMM using (1.3). Then testing assumption 3 amounts to make a test of the multiple inequality constraints $f(y_j) \geq 1$ for $j = 1...k$ (see e.g. [14], section 21.4, for the implementation of such tests). The situation is more involved when Y is continuous. Under assumptions 1-5 and additional technical conditions, a consistent nonparametric estimator $\widehat{f}$ of f is developed in Subsection 2.2. This estimator is constraint to belong to $[1, M]$ with $M > 1$. It should be possible to build a consistent, unconstrained estimator $\widetilde{f}$ of f. Then, under the maintained assumptions 1, 2, 4 and 5, a test of assumption 3 could be based on the distance between $\widehat{f}$ and $\widetilde{f}$. Indeed, under assumption 3, $\widetilde{f}(y)$ should be greater than one for most values of y, so the distance between the two should be close to zero.[18]

1.3 Set identification without conditional independence

A second interesting feature of equation (1.3) is that it provides an informative bound on parameters of interest under monotonicity conditions, which are far weaker than the conditional independence condition of assumption 3. In the sequel, we let $\widetilde{Z}$ denote variables which may be different or not from Z and whose distribution is also identified. Besides, because monotonicity conditions are meaningful in ordered sets only, we restrict to the case where $(Y, Z) \in \mathbb{R}^2$. We replace assumption 3 by the following ones.

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

Assumption 6 *Almost surely,* $y \mapsto P(D = 1|Y = y, \widetilde{Z})$ *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 selection 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 [36] (p. 28), and actually implies it, as part a) of Theorem 1.5 shows.

[18] The critical region of such a test would depend on the asymptotic distribution of $(\widehat{f}, \widetilde{f})$, whose derivation is beyond the scope of the paper.

Theorem 1.5 below provides bounds on parameters of the form $E(h(Y))$ for $h \in H_Y^1$ or $h \in H_{YZ}^2$, where we let

$$\begin{aligned} H_T^1 &= \{h \in L_T^1 \text{ and } h \text{ is increasing}\} \ (T = Y \text{ or } Z), \\ H_{YZ}^2 &= \{h \in L_Y^1 / \exists \widetilde{h} \in H_Z^1 / h(Y) = E(\widetilde{h}(Z)|D = 1, Y)\}. \end{aligned}$$

The set H_Y^1 includes, among others, functions of the form $h(y) = \lambda y$ with $\lambda > 0$ and indicator functions $h_u(y) = \mathbb{1}\{y \geq u\}$, so that parameters of the form $E(h(Y)), h \in H_Y^1$, include the survival function of Y at each point. The set H_{YZ}^2 is more abstract. In an informal way, H_{YZ}^2 will increase as the dependence between Y and Z becomes stronger. As a simple illustration, this set only includes constant function when Y and Z are independent (conditional on $D = 1$) but is equal to H_Y^1 when $Y = Z$. More formally, H_{YZ}^2 is a subset of the range of the conditional expectation operator $g \mapsto (y \mapsto E(g(Z)|D = 1, Y = y))$, which itself is linked to the null space of this operator. Indeed, when (Y, 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_{YZ}^2 will be maximal if the conditional expectation operator is injective, that is to say under a completeness condition on Y and Z.[19]

It seems difficult to test formally that $h \in H_{YZ}^2$ for a given, increasing, function h. On the other hand, we can test the stronger condition:

$$E(Z|D = 1, Y) = \alpha + \beta h(Y), \quad \beta > 0 \tag{1.4}$$

Test of such functional forms are described for instance by [46] (Subsection 4.2).

We suppose in the following that equation (1.3) admits a solution, and, as in the previous subsection, we let Q denote such a solution. More precisely, if the constant function $P(D = 1)$ is a solution, we let $Q(Y) = P(D = 1)$ 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 1.5 *Suppose that $P(D = 1) > 0$ and assumptions 1 and 2 hold for Z and $\widetilde{Z}$. Then:*
a) Under assumption 6, $E[h(Y)] \leq E\left[E(h(Y)|\widetilde{Z}, D = 1)\right]$ for all $h \in H_Y^1$. Moreover, this upper bound is sharp ;
b) Under assumptions 3', $E[Dh(Y)/Q(Y)] \leq E[h(Y)]$ for all function $h \in H_{YZ}^2$. Moreover, this lower bound is sharp provided that at least one solution Q lies in $]0; 1]$.

[19] If (Y, Z) has infinite support and the conditional expectation operator is injective, one can show that the dimension of H_{YZ}^2 is infinite.

c) For all function $h \in L^1_Y$, these three expectations are equal when $D \perp\!\!\!\perp (Y, Z, \widetilde{Z})$ or when $Z = \widetilde{Z} = Y$.

Part a) of Theorem 1.5 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 (1.3), 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 1.5 is that for all functions $h \in H^1_Y \cap H^2_{YZ}$, we can obtain a compact interval on $E(h(Y))$. This is so even if $h(Y)$ is unbounded. In this sense, the result is similar to proposition 2, corollary 2 of [37], 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 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 $Z = \widetilde{Z} = Y$, its length also reflects the quality of the chosen instruments. As the dependence between $(Z, \widetilde{Z})$ and Y increases, the knowledge of the distribution of the instruments enables to better predict parameters of the distribution of Y. Besides, the upper (resp. lower) inequality turns into an equalities whenever $Y \perp\!\!\!\perp D|\widetilde{Z}$ (resp. $Z \perp\!\!\!\perp D|Y$). Hence, Z and $\widetilde{Z}$ must be chosen according to different logics. $\widetilde{Z}$ intends to reduce selection on inobservables correlated with the outcome, whereas Z should be as independent of the selection (conditional on Y) as possible.

As noted before, H^2_{YZ} increases as the dependence between Y and Z becomes stronger. Hence, the quality of the instrument also matters for the range of applicability of the lower bound. If it seems difficult, without further restrictions, to describe the set $H^1_Y \cap H^2_{YZ}$ of functions h such that an interval can be built on $E[h(Y)]$, this set will contain at least all functions $h(y) = \lambda y$ with $\lambda > 0$ under the testable linear condition that $E(Z|D = 1, Y) = \alpha + \beta Y$ (with $\beta > 0$). In this case in particular, $E[Y]$ can be bounded below and above. Besides, if Y and Z exhibit a positive dependence, the following proposition states that the set $H^1_Y \cap H^2_{YZ}$ will be equal to H^2_{YZ}.

Proposition 1.6 *Suppose that for all z, $y \mapsto F_{Z|Y=y,D=1}(z)$ is decreasing. Then $H^2_{YZ} \subset H^1_Y$.*

1.4 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 Z. If 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 $E(Z)$ can be recovered.

In such situations, assumption 2 fails but intuitively, information on Z can provide identification, at least in a parametric setting. Theorem 1.5 gives a rigorous treatment to this idea. It generalizes the framework of [40] to the case where $Y \neq Z$. It is also very similar to the theory of generalized calibration developed by [11] in a survey sampling framework to handle nonignorable nonresponse with instruments. [11], however, does not consider the issue of identification of P.

As we consider a parametric framework here, we add explicitly covariates X. In the sequel, we suppose that $V = (X', Y')' \in \mathbb{R}^p$ and $W = (X', Z')' \in \mathbb{R}^q$. The identification result is based on the following assumptions.

Assumption 2' *$E(W)$ is known. Moreover, $P(D = 1|V) = F(V'\beta_0)$ where F is a known,*

differentiable and strictly increasing function from $\mathbb{R}$ *to* $]0,1[$*, and* V *is almost surely linearly independent conditional on* $D = 1$*.*

Assumption 3' $D \perp\!\!\!\perp Z|V$*.*

Assumption 4' $rank(E(DWV'F'(V'\beta_0)/F^2(V'\beta_0))) = p$*.*

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

Assumption 2' weakens assumption 2 on data availability, at the price of imposing a parametric restriction on P. The condition $P(D = 1|V) = F(V'\beta_0)$ with a known F is satisfied for instance if the selection equation is a logit or probit model. 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 Y to a linear form.

Theorem 1.7 *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 1.7 also provides a sufficient and testable condition which ensures the global identification of β_0.

2 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.

2.1 Parametric estimation

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

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

provides identification of the parameters $(P(y_k))_{1\leq k\leq K}$ if assumptions 3, 4 and 5 hold, by Theorem 1.3. Hence, consistent and asymptotically normal estimators can be obtained by GMM in this case. Similarly, if P satisfies the restrictions of assumption 2', then

$$E\left[\left(\frac{D}{F(V'\beta_0)} - 1\right) W\right] = 0. \tag{2.1}$$

Moreover, under assumption 4", β_0 is identified globally by these conditions. Thus GMM can also be used in this framework.

2.2 Nonparametric estimation

When 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. [10], [16], [23] or [41]). For the sake of simplicity, we assume that there is no covariate X and that $(Y, Z) \in [0, 1]^2$. 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 [16].

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

$$T\,\phi(z) = E(D\phi(Y^*)|Z = z).$$

Then (1.3) may be written as

$$T\,f = 1.$$

Under assumptions 3 and 4 (with $\mathcal{A} = L^1_Y$), T is injective.[20] However, its inverse is not continuous, so that we are faced to an ill-posed problem.[21] To achieve consistency, we adopt a Tikhonov regularization as [10], [16] and [23].

[20]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 $h_1 = h_2$.

[21]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 [13] in the framework of functional minimum distance.

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 ϕ defined on $[0,1]$, let also $||\phi||^2 = \int_0^1 \phi(u)^2 du$. Our estimator of f satisfies

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

where α_n is a regularization parameter which, basically, enables to rule out unstable solutions (see e.g. [4], for a discussion on regularization in ill-posed inverse problems). Under the assumptions below, such a solution will always exist but may not be unique (see [2]). 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 $u \in \mathbb{R}$, $K_h(u) = K_1(u/h)$ where K_1 is positive, $\int K_1(u)du = 1$ and $\int u K_1(u)du = 0$.*

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 non-parametric estimation. Assumption 10, which is identical to assumption 3 of [23], 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.[22]

Theorem 2.1 *Under assumptions 3-4 and 7-10,*

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

Theorem 2.1 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

[22] We suppose here that α_n is a deterministic sequence. See e.g. [13] for a data driven selection procedure.

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 2.2 ensures that $\widehat{\theta}$ is consistent.

Corollary 2.2 *Suppose that assumptions 3, 4 and 7-10 hold. Then*

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

3 Application

3.1 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,[23] the level of grade retention in France is still high. In 2002, for instance, a quarter of students have repeated at least once in primary school (see [44]). Yet, and despite the controversy on its effects in other countries,[24] there has been no serious attempts to measure its impact in the French educational system.[25]

The study is based on a panel of the French "Ministère de l'éducation Nationale" which follows 9641 children who entered 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

[23] A notable exception is United States. Indeed, several states have reintroduced this policy by tying promotion on a state or district assessment (see [29]).

[24] Positive effects include the possibility for disadvantaged children to catch up (see e.g. [29]) and the incentive for every student to increase their school efforts (see [28]). On the other hand, most educational and sociological studies underline its harmful effects on the motivation of children (see e.g. [9]), drop outs (see [31]) and even academic performances (see e.g. the meta-analyses of [22], 1989, or [30]). However, usually, these studies rely on very few controls (see e.g. [34], 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.

[25] [44] measures 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. [8] study the effects in third grade on the same data as here, using a linear regression approach.

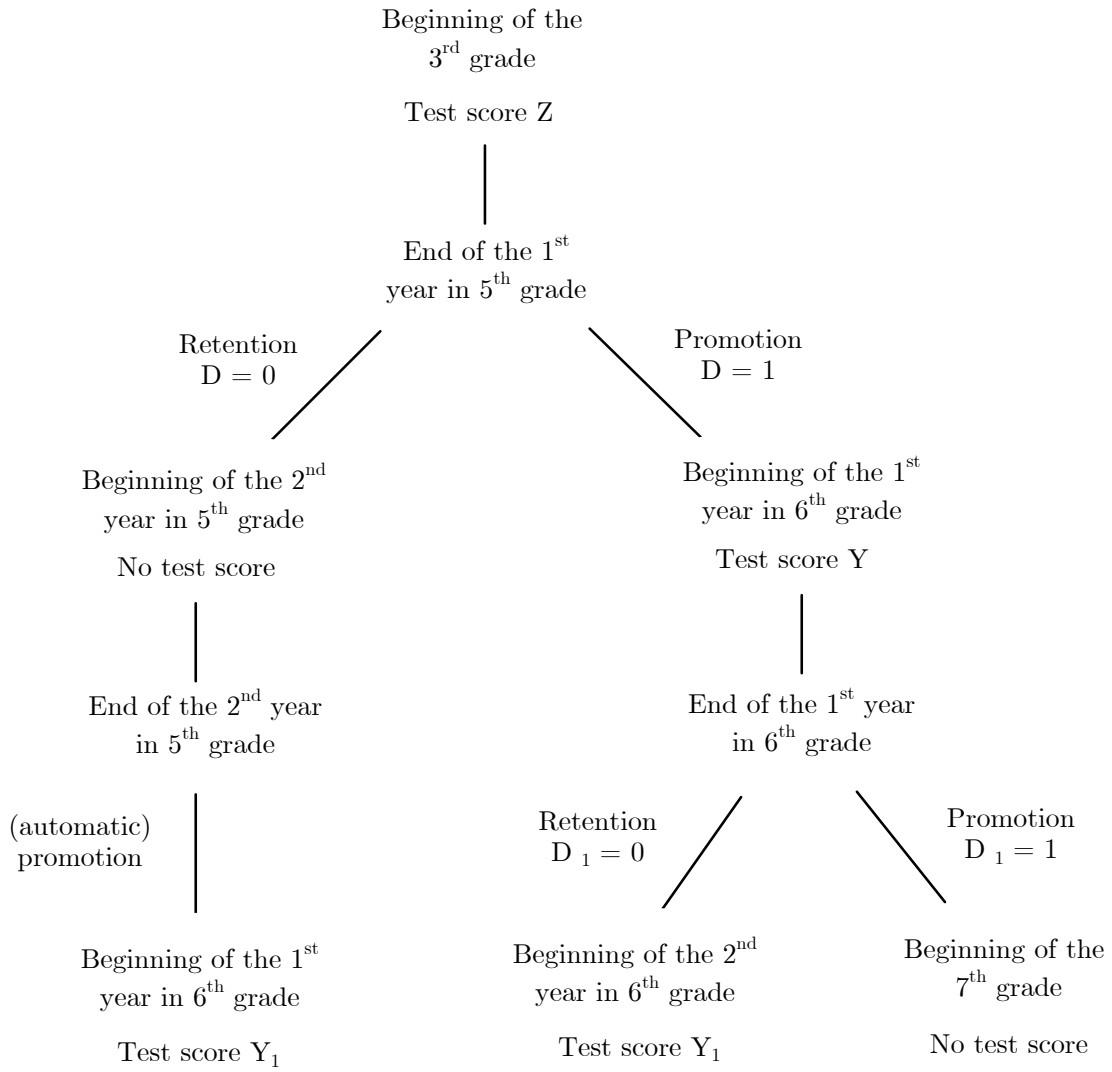

Figure 1: Promotion, retention and available test scores.

beginning of the third grade (variable Z) and sixth grade (variable Y for the 2002 test and Y_1 for the 2003 test).[26] 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.[27] 23.8 percent of this sample was excluded because of missing data on the standardized test scores in either third or sixth grade. The final sample consists in 5467 children. Among them, 2.2% were retained in 5th grade ($D = 0$), 6.7% in 6th grade ($D = 1$ and $D_1 = 0$)

[26]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.

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

while the others never repeated ($D = 1$ and $D_1 = 1$). Table 1 displays the average scores on this sample. The 2002 6th grade score is missing for children retained in 5th grade since they only entered this grade in 2003. Similarly, the 2003 6th grade score is not observed for children who never repeated, since they were in 7th grade in 2003. As expected, differences between retained and promoted pupils in terms of test achievement are large. On average, the fifth (resp. sixth) grade repeaters were already, in the 3rd grade test, more than 1.5 (resp. more than 1) standard deviations below the students who never repeated. 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 (see Figure 1). This feature of the sample will be useful in the following.

	Retained in 5th grade ($D = 0$)	Retained in 6th grade ($D = 1, D_1 = 0$)	Promoted in both grades ($D = 1, D_1 = 1$)
Number of observations	120	365	4982
3rd grade score Z	-1.48 (0.91)	-1.02 (0.90)	0.11 (0.94)
2002 6th grade score Y	-	-1.32 (0.81)	0.12 (0.93)
2003 6th grade score Y_1	-0.90 (0.87)	-0.64 (0.79)	-

Table 1: Summary statistics.

We focus here on the average effects of retention in fifth grade on test score achievement one year after. Let $Y_1(1)$ (resp. $Y_1(0)$) denote the 2003 sixth grade test score a student would have obtained if he had been promoted in sixth grade (resp. retained in fifth grade). The parameter of interest writes as

$$\Delta^{TT} = E(Y_1(0) - Y_1(1)|D = 0) \tag{3.1}$$

When $D = 0$, $Y_1(0)$ is observed by Y_1, but $Y_1(1)$ is unobserved. 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.[28] Rather, I suppose that the progressions of retained students had they been promoted in sixth grade can be bounded in the following way:

$$0 \leq E(Y_1(1) - Y|D = 0, Y) \leq E(Y_1(1) - Y|D = 1, D_1 = 0, Y). \tag{3.2}$$

[28]As an evidence of the discretionary nature of grade retention in France, 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.

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 were 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.

Under condition (3.2), we get

$$E(Y_1|D=0) - E\left[h(Y)|D=0\right] \leq \Delta^{TT} \leq E(Y_1|D=0) - E(Y|D=0), \tag{3.3}$$

where $h(Y) = E(Y_1(1)|D=1, D_1=0, Y)$. Students retained in sixth grade take the standardized test twice. Thus, we observe both Y and $Y_1(1)$ for them ($Y_1(1) = Y_1$ in this case), and h is identified. On the other hand, Y is unobserved for students retained in fifth grade, so that $E[h(Y)|D=0]$ and $E(Y|D=0)$ are not identified without further restrictions. Nonetheless, we can use the method developed previously to point or set identify them. Indeed, Y, the main factor of D, is unobserved when $D=0$. Besides, the third grade standardized test score Z is observed for both values of D and correlated with Y. We now consider the two cases corresponding respectively to the independence assumption $D \perp\!\!\!\perp Z|Y$ and the monotonicity conditions considered in Subsection 1.3.

3.2 Empirical strategies

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:

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

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 Y, but not directly grade retention. Under these assumptions, Theorem 1.3

applies and letting $p = P(D = 0)$, we can identify $E(h(Y)|D = 0)$ by

$$\begin{aligned} E\left[h(Y)|D=0\right] &= \frac{1}{p}\left(E[h(Y)] - (1-p)E[h(Y)|D=1]\right) \\ &= \frac{1}{p}\left((1-p)E\left[\frac{h(Y)}{P(Y)}|D=1\right] - (1-p)E[h(Y)|D=1]\right) \\ &= \frac{1-p}{p}E\left[\frac{1-P(Y)}{P(Y)}h(Y)|D=1\right]. \end{aligned}$$

$E(Y|D = 0)$ can be identified similarly. Then, using (3.2), we obtain the following lower and upper bounds on Δ^{TT}:

$$\underline{\Delta}_1^{TT} = E[Y_1|D=0] - \frac{1-p}{p}E\left[\frac{1-P(Y)}{P(Y)}h(Y)|D=1\right] \tag{3.4}$$

$$\overline{\Delta}^{TT} = E[Y_1|D=0] - \frac{1-p}{p}E\left[\frac{1-P(Y)}{P(Y)}Y|D=1\right], \tag{3.5}$$

To estimate these bounds, we first have to estimate h and P. h was estimated using a kernel estimator, with a gaussian kernel and a bandwidth estimated by cross validation (see Figure 2). P was estimated by the flexible parametric form

$$P(y;\beta) = \frac{1}{1+\exp\left(-\beta_0 - \sum_{i=1}^{k} y1\{y \geq \alpha_i\}\beta_i\right)}. \tag{3.6}$$

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.[29]. 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 estimator $P(.;\widehat{\beta})$ is displayed Figure 2.[30]

The estimator of $\underline{\Delta}_1^{TT}$ and $\overline{\Delta}^{TT}$ are then defined as being the empirical analog of (3.4) and (3.5):

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

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

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

Second strategy: monotonicity

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

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

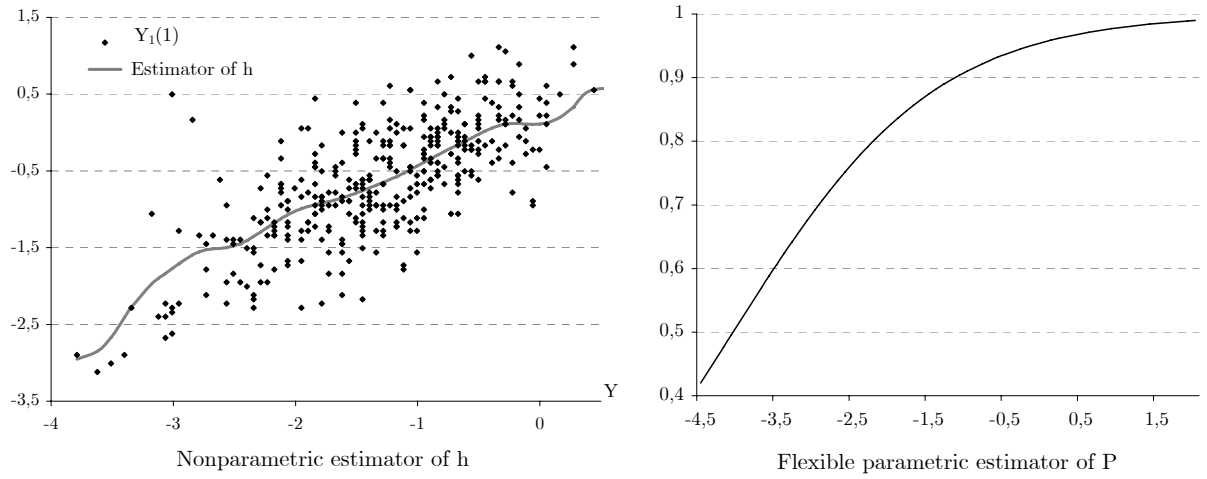

Figure 2: Estimation of h and P.

Basically, the conditional independence condition holds if Y 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. err.)
2002 6th grade score Y	1.31 (0.08)
3rd grade score Z	0.23 (0.07)

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

To apply Theorem 1.5 and obtain bounds on $E(h(Y)|D=0)$, we also need to check that $h \in H_Y^1 \cap H_{YZ}^2$. That h is increasing is apparent from Figure 2. To check that $h \in H_{YZ}^2$, we

implemented, as suggested in Subsection 1.3, a specification test of the form (1.4).[31] We obtain a positive and significant slope coefficient in (1.4) and do not reject, at the level of 1%, the linear specification. Hence, we do not reject the assumption that $h \in H^2_{YZ}$.

Under assumptions 3' and 6, and the condition $h \in H^1_Y \cap H^2_{YZ}$, we can apply Theorem 1.5 to obtain the following bounds on $E(h(Y)|D=0)$:

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

where Q denotes a solution of $E(D/Q(Y)-1|Z)=0$.[32]

To get bounds on $E(Y|D=0)$, we also check that the identity function belongs to H^2_{YZ}. This is true if $E(Z|D=1,Y)=\gamma+\lambda Y$ with $\lambda>0$. The specification test was not rejected at the level of 5%, so that we accept that the identity function belongs to $H^1_Y \cap H^2_{YZ}$. Under these assumptions, we get the same upper bound on Δ^{TT} as under conditional independence, but another lower bound, which writes as

$$\underline{\Delta}^{TT}_2 = E[Y_1|D=0] - E\left[E(h(Y)|Z,D=1)|D=0\right] \tag{3.7}$$

Moreover, $\underline{\Delta}^{TT}_2$ and $\overline{\Delta}^{TT}$ are sharp by Theorem 1.5.

To estimate $\underline{\Delta}^{TT}_2$, a kernel estimator $\widehat{g}$ of $g(z)=E(h(Y)|Z=z,D=1)$ was first estimated, and then plugged in the empirical analog of (3.7):

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

3.3 Results

The final 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.[33] 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 harmful in terms of test achievement. This assumption does not seem very credible,

[31] More precisely, we implemented the simple differencing test suggested by [46] (p. 701) with the kernel estimator $\widehat{h}$ instead of h.

[32] We do not use P here to emphasize the fact that the solution of this equation is not $P(D=1|Y)$ anymore. However, both P and Q are estimated with $P(.;\widehat{\beta})$.

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

though. As emphasized previously, the effects of Y on D 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.

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.

In conclusion, and even if uncertainty is rather important,[34] the conclusion on short term effects of grade retention is rather positive. This result is in line with the results of [29] 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.

4 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

[34] 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.

References

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

[2] N. Bissantz, T. Hohage, and A. Munk. Consistency and rates of convergence of nonlinear tikhonov regularization with random noise. *Inverse Problems*, 20:1773–1789, 2004.

[3] R. Blundell, X. Chen, and D. Kristensen. Nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75:1613–1669, 2007.

[4] M. Carrasco, J. P. Florens, and E. Renault. Linear inverse problems and structural econometrics: Estimation based on spectral decomposition and regularization. In J. J. Heckman and E. E. Leamer, editors, *Handbook of Econometrics*, volume 6. North Holland, 2006.

[5] G. Chamberlain. Asymptotic efficiency in semiparametric model with censoring. *Journal of Econometrics*, 32:189–218, 1986.

[6] C. Chen. Parametric models for response-biased sampling. *Journal of the Royal Statistical Society, Series B*, 63:775–789, 2001.

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

[8] O. Cosnefroy and T. Rocher. Le redoublement au cours de la scolarité obligatoire : nouvelles analyses, mêmes constats. *Education et Formation*, 70:73–82, 2004.

[9] M. Crahaye. *Peut-on lutter contre l'échec scolaire ?* De Boeck, 1996.

[10] S. Darolles, J. P. Florens, and E. Renault. Nonparametric instrumental regression. Working Paper, 2006.

[11] J. C. Deville. La correction de la non-réponse par calage généralisé. In *Actes des Journées de Méthodologie Statistique 2002*, pages 4–20. INSEE, 2002.

[12] X. d'Haultfœuille. On the completeness condition in nonparametric instrumental regression. *Econometric Theory, forthcoming*, 2008.

[13] P. Gagliardini and O. Scaillet. Tikhonov regularization for functional minimum distance estimators. Working Paper, 2006.

[14] C. Gouriéroux and A. Monfort. *Statistics and Econommetric Models*. Cambridge University Press, 1995.

[15] J. T. Grogger and R. T. Carson. Models for truncated counts. *Journal of Applied Econometrics*, 6:225–238, 1991.

[16] P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33:2904–2929, 2005.

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

[18] J. J. Heckman. Shadow prices, market wages, and labor supply. *Econometrica*, 42:679–694, 1974.

[19] J. J. Heckman and E. Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73:669–738, 2005.

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

[21] S. Hemvanich. The general missingness problems and estimation in discrete choice models. Working Paper, 2004.

[22] T. Holmes. Grade level retention effects : A meta-analysis of research studies. In L. A. Sheppard and M. L. Smith, editors, *Flunking Grades. Research and Policies on Retention*, pages 16–33. New York, The Falmer Press, 1989.

[23] J. L. Horowitz and S. Lee. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75:1191–1208, 2007.

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

[25] Y. Hu and S. Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76:195–216, 2008.

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

[27] G. W. Imbens and T. Lancaster. Combining micro and macro data in microeconometric models. *Review of Economic Studies*, 61:655–680, 1994.

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

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

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

[31] S. Jimerson, G. E. Anderson, and A. D. Whipple. 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, 2002.

[32] A. Lewbel. Endogenous selection or treatment model estimation. *Journal of Econometrics*, 141:777–806, 2007.

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

[34] J. Lorence. Retention and academic achievement research revisited from an united states perspective. *International Education Journal*, 7:731–777, 2006.

[35] C. F. Manski. The selection problem. In C. Sims, editor, *Advances in Econometrics, Sixth World Congress*. Cambridge University Press, 1994.

[36] C. F. Manski. *Partial Identification of Probability Distribution*. Springer, 2003.

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

[38] L. Mattner. Completeness of location families, translated moments, and uniqueness of charges. *Probability Theory and Related Fields*, 92:137–149, 1992.

[39] L. Mattner. Some incomplete but boundedly complete location families. *Annals of Statistics*, 21:2158–2162, 1993.

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

[41] W. Newey and J. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 2003.

[42] E. A. Ramalho and R. J. Smith. Discrete choice nonresponse. CEMMAP working paper, 2007.

[43] G. Tang, R. J. A. Little, and T. E. Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90:747–764, 2003.

[44] T. Troncin. 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, 2005.

[45] J. Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141:1281–1301, 2007.

[46] A. Yatchew. Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36:669–721, 1998.