ESTIMATION OF TREATMENT EFFECTS IN OBSERVATIONAL STUDIES BY RECOVERING THE ASSIGNMENT PROBABILITIES AND THE POPULATION MODEL

DANNY PFEFFERMANN, VICTORIA LANDSMAN

ABSTRACT

In observational studies the assignment of units to treatments is with unknown probabilities. Consequently, estimation and comparison of treatment effects based on the empirical distributions of the response under the various treatments can be biased since units exposed to one treatment could differ in important but unknown characteristics from units exposed to other treatments.

In this article we study the plausibility of analyzing observational data by deriving the parametric distribution of the observed response under a given treatment as a function of the distribution that would be obtained under a strongly ignorable assignment, and the assignment process, which is modeled as a function of the observed data (the response and covariate values). The use of this approach is founded by showing that the sample distribution of the observed responses is identifiable under some general conditions. The goodness of fit of this distribution can be tested by using standard test statistics since it refers to the observed data, but we also develop a new test. The proposed approach allows also testing the assumptions underlying the use of methods that employ instrumental variables, or methods that use propensity scores with a given set of covariates.

We assess the performance of the proposed approach and compare it to existing approaches using data collected in the year 2000 by OECD for the Programme for International Student Assessment (PISA). In the present application we compare students' scores in mathematics between public and private schools in Ireland and conclude, somewhat surprisingly, that the public schools perform better than the private schools. This finding is supported by one of the existing methods as well.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M07/10

Estimation of treatment effects in observational studies by recovering the assignment probabilities and the population model

Danny Pfeffermann,
Hebrew university of Jerusalem, *Israel*, *and University of Southampton, UK*
and Victoria Landsman
Hebrew University of Jerusalem, *Israel*

Summary. In observational studies the assignment of units to treatments is with unknown probabilities. Consequently, estimation and comparison of treatment effects based on the empirical distributions of the response under the various treatments can be biased since units exposed to one treatment could differ in important but unknown characteristics from units exposed to other treatments.

In this article we study the plausibility of analyzing observational data by deriving the parametric distribution of the *observed* response under a given treatment as a function of the distribution that would be obtained under a strongly ignorable assignment, and the assignment process, which is modeled as a function of the observed data (the response and covariate values). The use of this approach is founded by showing that the sample distribution of the observed responses is identifiable under some general conditions. The goodness of fit of this distribution can be tested by using standard test statistics since it refers to the observed data, but we also develop a new test. The proposed approach allows also testing the assumptions underlying the use of methods that employ instrumental variables, or methods that use propensity scores with a given set of covariates.

We assess the performance of the proposed approach and compare it to existing approaches using data collected in the year 2000 by OECD for the *Programme for International Student Assessment* (PISA). In the present application we compare students' scores in mathematics between public and private schools in Ireland and conclude, somewhat surprisingly, that the public schools perform better than the private schools. This finding is supported by one of the existing methods as well.

Keywords: Average treatment effect, Control functions, Instrumental variables, Propensity scores, Weighted estimators.

Contact details: Danny Pfeffermann, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, UK. Email: msdanny@huji.ac.il

1. Introduction

Observational studies are in common use for estimating and comparing the effects of different 'treatments' (medical treatments, teaching methods, new policies, etc.). In this kind of studies, the assignment of subjects to treatments often depends on latent assignment variables that are unknown to the investigator but could be related to the values of the response variable even when conditioning on known covariates. Consequently, a direct comparison of the response distributions (given the model covariates) or moments of these distributions between treatment groups might be misleading because units exposed to one treatment could differ in important but unknown characteristics from units exposed to other treatments.

Consider a finite population U composed of N elements, $\{1,...,N\}$. Suppose that every element $i \in U$ is potentially exposed to *m* treatments with responses $y_i^t, t = 1,...,m$. The random variable y_i^t represents the response that would be obtained if unit i had been exposed to treatment t. The target parameters of interest are population means like, $\mu^{p,t} = \frac{1}{N}\sum_{i=1}^{N} y_i^t$, or $\mu^t = \frac{1}{N}\sum_{i=1}^{N} E(y_i^t \mid x_i)$, where x_i defines a set of known covariates that affects the responses, and the expectation $E(y_i^t \mid x_i)$ is with respect to a 'superpopulation' model postulated for the responses. Very often, contrasts between the parameters $\mu^{p,t}$ or μ^t are of primary interest, such as the mean difference between two treatments, known as the *average treatment effect* (ATE). The assumption that every element in the population could possibly be exposed to every treatment, known as the "counterfactual approach", underlies many of the methods used in observational studies, starting with Neyman (1923/1990) and Fisher (1951). Rubin (1974, 1977), Rosenbaum (1984), Rosenbaum and Rubin (1983, 1984) and Smith and Sugden (1988) among others followed this formulation.

In practice, every element can be exposed to only one treatment if the net treatment effects are to be compared on 'equal grounds' (Holland, 1986). Also, it is rarely the case that all the population elements participate in the study. Let *S* define a sample of observational units of size n and denote by π_i the probability that element $i \in U$ is included in the sample. The probabilities π_i possibly depend on *sampling variables* Z_i, which may affect the treatment response but may not be

known to the analyst. In observational studies, unlike in survey sampling, the probabilities π_i are often unknown, as the selection to the sample could be by 'self-selection'. Every unit $j \in S$ is exposed to one of the m treatments with treatment assignment probabilities, $p_j^t = Pr[T(j) = t \mid j \in S]$; $\sum_{t=1}^{m} p_j^t = 1$, where T defines the assignment process. The probabilities p_j^t are assumed to depend on *treatment assignment variables* A_j, which again may affect the responses but are unknown in a typical observational study. The probability that unit $i \in U$ is included in the sample and assigned to treatment t is therefore,

$$P(i \in S, T(i) = t) = P(i \in S) \times P(T(i) = t \mid i \in S) = \pi_i \times p_i^t = q_i^t . \qquad (1.1)$$

After the assignments take place, the sample S is divided into sub-samples S^t of size n_t, $\sum_{t=1}^{m} n_t = n$, where $S^t = \{i \mid i \in S, T(i) = t\}, t = 1,...,m$. Epidemiologists sometimes refer to the bias induced by the sampling process as *selection bias*, and the bias resulting from the assignment process as *confounding bias*; see the discussion in Rothman (2002).

Sugden and Smith (1988) establish conditions on the sampling and assignment processes that allow ignoring them in the inference process. A simple special case is when all the sample selection probabilities π_i are equal and similarly for the assignment probabilities p_i^t, such that $q_i^t = q^t$ for every $i \in U$. The condition that every element in the population has the same probability of being exposed to a given treatment t constitutes a special case of a *strongly ignorable assignment*. An assignment process with sample inclusion probabilities π_i and treatment assignment probabilities p_i^t is strongly ignorable given x_i, if the sample model satisfies,

$$f_{S^t}(y_i^t \mid x_i) = f(y_i^t \mid x_i, i \in S^t) = f_p(y_i^t \mid x_i), \ \forall i \in U , \qquad (1.2)$$

where $f_p(y_i^t \mid x_i)$ defines the 'population' probability density function (*pdf*) of y_i^t under the formulation described above by which every element in the population is potentially exposed to each of the treatments $t = 1,...,m$. This definition of strong ignorability corresponds to the concept of 'noninformative sampling' in sample survey inference as defined in Pfeffermann *et al*. (1998). It is satisfied under the condition of independence between the assignment process and the response values, given the

covariates, as stated in Rosenbaum and Rubin (1983). The latter article assumes implicitly that the initial sample S is selected by simple random sampling.

The problem of observational studies is that although the measurements y_j^t are only taken after that the sample units are assigned to the various treatments, they may be related to the sampling variables Z_j and/or the treatment assignment variables A_j. If the effects of these variables on the responses are not accounted for by the covariates x_j included in the model, the ignorability condition (1.2) is no longer satisfied and the sample *pdf* $f_{S^t}(y_i^t \mid x_i)$ is different from the population *pdf* $f_p(y_i^t \mid x_i)$. As well known and illustrated in this article, ignoring the effect of the sample selection or the treatment assignment may result in highly biased estimators.

In this article we study the plausibility of approximating the *pdf* $f_{S^t}(y_i^t \mid x_i)$ of the observed responses under a given treatment by modeling the hypothetical population distribution under strong ignorability and the assignment rule. Fitting the resulting 'sample model' to the observed responses enables then to estimate the population model and hence estimate and compare the net treatment effects. The use of this approach is validated by showing that the sample model is identifiable under some general conditions on the population distribution under strong ignorability and the sampling/assignment rule. Furthermore, the goodness of fit of the sample model can be tested using simple test statistics. Estimating the population distribution and the assignment rule enables also to test the validity of applying propensity scores methods or instrumental variables in any given problem.

The paper is organized as follows. Section 2 contains a brief review of some of the classical methods in common use. Section 3 defines the sample model and discusses the estimation of the unknown model parameters. Section 4 defines sufficient conditions guaranteeing the identifiability of the sample model and Section 5 outlines test statistics for testing the goodness of fit of this model. Section 6 illustrates the application of the proposed approach and compares it to some other approaches proposed in the literature using data collected as part of the PISA program carried out by OECD. In this illustration we compare pupils' test scores in mathematics between public and private schools in Ireland. A simulation study that

uses the models fitted to this data set enables studying additional features of our approach. We conclude with a brief discussion in Section 7.

2. Methods in common use

In this section we review briefly some of the classical methods for observational studies in common use. This review is important for a better understanding of the approach outlined in subsequent sections and for the empirical comparison between the alternative methods in Section 6. We consider for convenience a two treatments case $(T = 0,1)$ and assume that the sample $S = S^0 \cup S^1$ of size n is selected with equal probabilities. The target parameter is defined to be the sample 'average treatment effect',

$$ATE = \frac{1}{n}\sum_{i=1}^{n}\left[E_p(y_i^1 \mid x_i) - E_p(y_i^0 \mid x_i)\right] = \frac{1}{n}\sum_{i=1}^{n} d_i\,, \tag{2.1}$$

where $E_p(\cdot)$ is the expectation under the population distribution. As mentioned earlier, most of the literature on observational studies does not distinguish between the initial sample before the assignment to treatments and the population from which the sample is taken. Note also that if the initial sample is selected with known probabilities, the ATE in the population can be estimated from the sample estimators $\{\hat{d}_i\}$ by application of classical sample surveys methods. See also below.

2.1. Methods for strongly ignorable treatment assignments

2.1.1. Regression methods

Suppose that the population model for the potential response y^t has the general form,

$$y^t = r^t(x) + u^t\,,\ E_p(u^t) = 0\,,\ t = 0,1\,, \tag{2.2}$$

where r^t is a deterministic function of x, and that the assignment process is ignorable such that (1.2) holds. Under this assumption, $E_{S^t}(y^t \mid x) = E_p(y^t \mid x) = r^t(x)$, $t = 0,1$, where $E_{S^t}(\cdot)$ is the expectation under the sample distribution $f_{S^t}(y_i^t \mid x_i) = f(y_i^t \mid x_i, i \in S^t)$ (see Section 3.1). Hence, one can estimate in this case the regressions $r^t(x), t = 0,1$ from the sample data in S^0 and S^1, and estimate,

$$A\hat{T}E = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{r}^1(x_i) - \hat{r}^0(x_i)\right]. \tag{2.3}$$

2.1.2. Imputation methods

Methods in this category impute the potential responses y_i^0 for $i \in S^0$ and y_j^1 for $j \in S^1$ by matching the covariates x. In practice, x is often of high dimension, in which case the one-dimensional 'propensity score' $e(x) = P(T = 1 | x)$ is used instead. Rosenbaum and Rubin (1983) show that under strongly ignorable treatment assignments, the potential responses y^1, y^0 are independent of T given $e(x)$, thus validating the use of the propensity scores for matching. In practice, the propensity scores are estimated by fitting logistic or probit models.

Mean imputation

Denote by $J_M^t(i)$ the M closest matches in S^t for unit $i \in S^{1-t}$ based on x or $\hat{e}(x)$. Then, for unit $i \in S^{1-t}$ $\hat{y}_i^{1-t} = y_i^{1-t}$ and $\hat{y}_i^t = \frac{1}{M}\sum_{j \in J_M^t(i)} y_j^t$, $t = 0,1$. Estimate,

$$A\hat{T}E = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^1 - \hat{y}_i^0). \tag{2.4}$$

See Abadie and Imbens (2006) for more details.

2.1.3. Propensity weighted contrasts

Propensity scores have been proposed also for constructing weighted estimators of the corresponding population means, similarly to the Horvitz and Thompson (1952) and Hajek (1971) estimators. Consider the estimator,

$$A\hat{T}E = [\sum_{i=1}^{n}\frac{T_i}{\hat{e}(x_i)}]^{-1}\sum_{i=1}^{n}\frac{T_i y_i}{\hat{e}(x_i)} - [\sum_{i=1}^{n}\frac{(1-T_i)}{1-\hat{e}(x_i)}]^{-1}\sum_{i=1}^{n}[\frac{1-T_i}{1-\hat{e}(x_i)}]y_i. \tag{2.6}$$

where $T_i = 1$ if $i \in S^1$ such that $y_i = y_i^1$, and $T_i = 0$ and $y_i = y_i^0$, otherwise (see also Rosenbaum, 1987.) For large samples this estimator is approximately unbiased for $(\bar{y}^1 - \bar{y}^0) = \sum_{i=1}^{n}(y_i^1 - y_i^0)/n$ under all possible assignments of a given sample (assuming correct specification of the propensity scores).

Robins *et al.* (1994) consider the estimator,

$$A\hat{T}E = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i y_i - [T_i - \hat{e}(x_i)]\hat{r}^1(x_i)}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)y_i + [T_i - \hat{e}(x_i)]\hat{r}^0(x_i)}{1-\hat{e}(x_i)}. \tag{2.7}$$

This estimator has the "double-robustness" property of being consistent even if only the model fitted for the propensity scores, or the regression models (2.2) in the two sub-samples are correctly specified. The estimator (2.7) resembles the GREG estimator (Sarndal, 1980), which is in common use in survey sampling. Recently, Qin and Zhang (2007) proposed a new estimator having a somewhat stronger robustness property than (2.7). See Imbens (2004) for review and discussion of semiparametric estimators under strongly ignorable treatment assignment.

2.2. Methods that use external variables to control the assignment

Two methods in common use in this group are the method of control functions and the method of instrumental variables.

2.2.1. Control functions

This method was originally proposed by Heckman (1978,1979). It assumes that the population model consists of two equations:

a) A structural equation modelling the potential responses; $y^t = r^t(x) + u^t$,

$E_p(u^t) = 0$, $t = 0,1$ (same as (2.2)).

b) A latent variable equation modelling the treatment assignment,

$$W = m(v) + u_v,\ E_p(u_v) = 0\ ;\ T = 1 \Leftrightarrow W \geq 0\,, \tag{2.8}$$

where W is a latent variable and m is a deterministic function of v, a set of observed covariates 'explaining' the choice of treatments. The problem is to model the sample expectations,

$$E_{S^t}(y^t \mid x,v) = r^t(x) + E_{S^t}(u^t \mid x,v),\ t = 0,1\,. \tag{2.9}$$

Assuming, $E(u^t \mid x,v,T) = E(u^t \mid T)$, $t = 0,1$, we have,

$$\begin{aligned} E(u^1 \mid x,v,T=1) &= E(u^1 \mid W \geq 0) = E(u^1 \mid u_v \geq -m(v)) = K_1(v) \\ E(u^0 \mid x,v,T=0) &= E(u^0 \mid W < 0) = E(u^0 \mid u_v < -m(v)) = K_0(v) \end{aligned}, \tag{2.10}$$

and hence,

$$E_{S^t}(y^t \mid x,v) = r^t(x) + K_t(v),\ t = 0,1\,. \tag{2.11}$$

The functions $K_t(v)$ are called 'control functions'. A common practice of fitting the model (2.11) is to assume that (u_v, u^1, u^0) is trivariate normal with expectation zero

and covariance matrix $\sum$. Heckman and Vytlacil (2006) review extensions of the method, including non-parametric estimation.

2.2.2. Instrumental variables

Suppose that the means in the population models (2.2) are linear, $r^t(x) = \mu^t + x'\beta^t$, such that $ATE = (\mu^1 - \mu^0) + \bar{x}(\beta^1 - \beta^0)$ where $\bar{x} = \sum_{i=1}^{n} x_i / n$. The (conventional) instrumental variables method assumes the sample model,

$$y = \mu^0 + \delta T + (1-T)x'\beta^0 + Tx'\beta^1 + u\,, \tag{2.12}$$

where $y = Ty^1 + (1-T)y^0$, u is the unobserved residual, which is correlated with the assignment variable T and $\delta = (\mu^1 - \mu^0)$.

The method assumes the availability of instrumental variables f satisfying,

a. $E(u \mid x, f) = 0$; *b.* $E_p(y^t \mid x, f) = E_p(y^t \mid x)$, $t = 0,1$; *c.* $\Pr(T = 1 \mid x, f) = g(x, f)$, a 'non-trivial' function of f; *d.* $Var(u \mid x, f)$ is constant. Let, $\tilde{x}' = (1, T, (1-T)x', Tx')$ be the vector of 'covariates', $z' = (1, g, (1-g)x', gx')$ the vector of 'instruments' and denote by $\theta = (\mu^0, \delta, \beta^0, \beta^1)$ the unknown parameters. Multiplying both sides of (2.12) by z' and taking expectations, implies using condition *a*,

$$E(z'\tilde{x})\theta = E(z'y)\,. \tag{2.13}$$

Estimation of θ in (2.13) is carried out in two steps:

1 – Estimate $\hat{g}(x, f) = \hat{E}(T \mid x, f) = \hat{P}(T = 1 \mid x, f)$ by fitting probit or logit regression;

2 - Estimate the vector parameter θ as $\hat{\theta}_{IV} = (\sum_{i=1}^{n} \hat{z}_i' \tilde{x}_i)^{-1} \sum_{i=1}^{n} \hat{z}_i' y_i$, where $\hat{z}'_i = (1, \hat{g}, (1-\hat{g})x', \hat{g}x')_i$. Wooldridge (2002, Ch.18.4) discusses different plausible conditions regarding the behavior of the error u in (2.12) and corresponding estimation procedures.

The method of instrumental variables has been extended for estimating other parameters of interest. Imbens and Angrist (1994) and Angrist *et al*. (1996) define a Local Average Treatment Effect (LATE) and show how to estimate it using instrumental variables. Local instrumental variables (LIV) is an alternative approach of implementing the method of control functions, see Heckman and Vytlacil (2006). Heckman and Navarro (2004) provide conditions under which the LATE is a special case of LIV.

2.3 Discussion

All the methods described above assume the existence of known variables that control the effect of the assignment process under certain conditions. Two major challenges with the use of these methods are therefore how to identify plausible 'control variables' and how to test that they satisfy the required conditions. Rosenbaum (2002) discusses methods of testing the sensitivity of the inference to different assumptions on confounding variables that affect the assignments.

In the remainder of this article we discuss an alternative approach for observational studies that does not require the use of control variables. Moreover, as illustrated in Section 6, the use of this approach allows testing the appropriateness of candidate instrumental variables and/or the use of propensity scores for inference.

3. An alternative approach for observational studies

Our proposed approach attempts to approximate the parametric sample distribution of the observed responses under a given treatment. The validity of this approach is studied theoretically in Sections 4 and 5, and empirically in Section 6.

3.1. The sample distribution

As described in the introduction, we assume that the sample S^t of units exposed to treatment t is generally obtained in two stages. First, a sample S of n observational units is obtained with inclusion probabilities π_i and then every unit $j \in S$ is assigned (or assigns itself) to one of the m treatments with probabilities, p_j^t , $\sum_{t=1}^{m} p_j^t = 1$. Alternatively, the assignment to treatments may take place in the population and then a sample is selected from each of the treatment groups. This scenario underlies the application in Section 6 where we compare students' proficiencies in public and private schools based on probability samples of students from the two types of schools. The analysis below applies to both cases but we assume for convenience that the sample selection takes place first. Denote by $q_j^t = \pi_j \times p_j^t$ the probability that unit $j \in U$ is included in the sample and assigned to treatment t , and by $f_p(y_j^t \mid x_j)$ the *population pdf* that would be obtained under a

strongly ignorable assignment process as defined by (1.2). The *sample pdf* of y_j^t for unit $j \in S^t$ is obtained by application of Bayes theorem as,

$$f_{S^t}(y_j^t \mid x_j) = f(y_j^t \mid x_j, j \in S^t) = [\Pr(j \in S^t \mid y_j^t, x_j) f_p(y_j^t \mid x_j)] / \Pr(j \in S^t \mid x_j), \qquad (3.1)$$

where $\Pr(j \in S^t \mid x_j) = \int \Pr(j \in S^t \mid y_j^t, x_j) f_p(y_j^t \mid x_j) dy_j^t$.

Remark 1: It follows from (3.1) that the sample *pdf* is generally different from the *pdf* $f_p(y_i^t \mid x_i)$ under strong ignorability unless $\Pr(j \in S^t \mid y_j^t, x_j) = \Pr(j \in S^t \mid x_j)$ for all y_j^t, in which case the sampling and treatment assignment can be ignored in the inference process. See Rosenbaum (1987) for a similar condition.

Remark 2: The probabilities $\Pr(j \in S^t \mid x_j)$ are the *propensity scores,* introduced by Rosenbaum and Rubin (1983). See Section 2.1.2 above.

Remark 3: The probabilities $\Pr(j \in S^t \mid y_j^t, x_j)$ are generally not the same as the actual inclusion probabilities, $q_j^t = \Pr(j \in S^t)$, which as discussed in the Introduction, may depend on sampling variables Z_j and treatment assignment variables A_j that are possibly related to the responses y_j^t. Nonetheless, by regarding the probabilities q_j^t as realizations of random variables, the following relationship holds,

$$\Pr(j \in S^t \mid y_j^t, x_j) = \int \Pr(j \in S^t \mid y_j^t, x_j, q_j^t) f(q_j^t \mid y_j^t, x_j) dq_j^t = E(q_j^t \mid y_j^t, x_j). \qquad (3.2)$$

Substituting (3.2) in (3.1) gives an alternative representation for the sample *pdf* as,

$$f_{S^t}(y_j^t \mid x_j) = \frac{E(q_j^t \mid y_j^t, x_j) f_p(y_j^t \mid x_j)}{E(q_j^t \mid x_j)}. \qquad (3.3)$$

The use of (3.3) for inference instead of (3.1) has the advantage that it only requires specifying the form of the conditional expectations, $E(q_j^t \mid y_j^t, x_j)$.

The sample *pdf* defined by (3.1) or (3.3) was shown in recent years to provide a valuable modeling approach for inference from complex sample surveys; see the articles by Pfeffermann *et al*. (1998), Pfeffermann and Sverchkov (1999, 2003), Chambers *et al*. (2003), Sverchkov and Pfeffermann (2004) and Pfeffermann *et al*. (2006). These studies utilize the sample *pdf* for inference generalized linear and multi-level models, testing of distribution functions and prediction of finite population

totals. Pfeffermann and Sverchkov (1999, 2003) and Chambers *et al*. (2003) develop test statistics for testing the informativeness of the sampling process.

The obvious distinction between survey sampling and observational studies is that in survey sampling the sample inclusion probabilities are generally known for every element in the sample, which enables identifying and estimating the conditional expectations $E(\pi_i \mid y_i, x_i)$, and testing the informativeness of the sampling process. This is generally not the case in observational studies, requiring therefore to model the parametric forms of the probabilities $\Pr(j \in S^t \mid y_j^t, x_j)$ in (3.1) or the expectations $E(q_j^t \mid y_j^t, x_j)$ in (3.3). Fitting the logistic or probit function for these probabilities is a natural choice. As discussed below, modeling the sample *pdf* by use of (3.1) or (3.3) allows estimating the unknown parameters indexing the *pdf* $f_p(y_j^t \mid x_j)$ and the probabilities $\Pr(j \in S^t \mid y_j^t, x_j)$ or the expectations $E(q_j^t \mid y_j^t, x_j)$, and testing the goodness of fit of the estimated sample *pdf*.

3.2. Estimating the parameters of the sample distribution

So far we suppressed for convenience in the notation the parameters indexing the sample *pdf*. Consider the *pdf* (3.3). Testing the existence of possible treatment effects requires initially to allow for different parameters for different treatments. Adding the unknown parameters to the notation, the sample *pdf* under a given treatment t takes the form,

$$f_{S^t}(y_j^t \mid x_j; \alpha^t, \theta^t) = \frac{E(q_j^t \mid y_j^t, x_j; \alpha^t) f_p(y_j^t \mid x_j; \theta^t)}{E(q_j^t \mid x_j; \alpha^t, \theta^t)}. \tag{3.4}$$

Assuming that the inclusion in the sample and the assignment to the treatments are independent between units and that the responses y_j^t are likewise independent, the sample likelihood for treatment *t* takes the form,

$$L_{S^t}[\alpha^t, \theta^t; \{y_j^t, x_j\}] = \prod_{j=1}^{n_t} \frac{E(q_j^t \mid y_j^t, x_j; \alpha^t) f_p(y_j^t \mid x_j; \theta^t)}{E(q_j^t \mid x_j; \alpha^t, \theta^t)}. \tag{3.5}$$

Alternatively, the likelihood (3.5) can be replaced by the joint ('full') likelihood of the sample selection and the sample measurements, defined as,

$$L_S[\alpha^t,\theta^t;\{y_j^t,x_j;\, j\in S^t, x_i, i\notin S^t\}] = \\ = \prod_{j=1}^{n_t} E(q_j^t \mid y_j^t, x_j;\alpha^t) f_p(y_j^t \mid x_j;\theta^t) \prod_{i\notin S^t}[1-E(q_i^t \mid x_i;\alpha^t,\theta^t)] \quad . \tag{3.6}$$

The likelihood (3.6) has the advantage of comprising the model for the probabilities q_i^t for units outside the sample and thus using more information for estimating the model parameters, but finding the maximum is often more complicated. Notice that by dividing and multiplying by the product $\prod_{j=1}^{n_t} E(q_j^t \mid x_j;\alpha^t,\theta^t)$, the likelihood in (3.6) is seen to be the product of the sample likelihood (3.5) and the probability of observing the sample S^t, given the covariates x_k in and outside S^t. This likelihood is often applied in other areas, like when modeling data exposed to nonresponse, see, e.g., Greenlees *et al*. (1982), Gelman (2003, Ch.7), Pfeffermann and Sverchkov (2003) and Little (2004).

Maximization of either of the likelihoods (3.5) or (3.6) with respect to the unknown parameters yields the maximum likelihood estimators (*mle*) $\{\hat{\alpha}^t,\hat{\theta}^t, t=1,...,m\}$. Replacing the unknown model parameters by their *mle* yields the estimates,

$$\hat{f}_p(y_j^t \mid x_j) = f_p(y_j^t \mid x_j;\hat{\theta}^t);\;\; \hat{q}_j^t = \hat{E}(q_j^t \mid y_j^t, x_j) = E(q_j^t \mid y_j^t, x_j;\hat{\alpha}^t) \tag{3.7}$$

Remark 4: The separate likelihoods defined by (3.5) and (3.6) can be enhanced by modeling jointly the sample responses and assignment probabilities for all the sample units. This extension seems natural since every unit is assigned to one and only one of the treatments, implying $\sum_{t=1}^{m} E(q_j^t \mid x_j;\alpha^t,\theta^t) = 1$. Empirical evidence so far did not show any significant improvement by this joint modelling.

3.3. Calibration constraints

Suppose that the population size, *N*, is known and likewise some or all of the means, $\bar{X}_i = \sum_{k=1}^{N} x_{ki}/N$, or that they can be estimated unbiasedly (e.g., when the initial sample is selected with known probabilities π_i). Under the model, and for sufficiently large sample sizes, $\hat{N} = \sum_{j=1}^{n_t}(1/\hat{q}_j^t) \cong N$ and $\hat{\bar{X}}_i = \sum_{j=1}^{n_t}(x_{ji}/\hat{q}_j^t)/\hat{N} \cong \bar{X}_i$ for each t. Thus, the estimation process can be enhanced by maximizing the likelihoods (3.5) or (3.6) subject to the constraints,

$$\sum_{j=1}^{n_t}(1/q_j^t)=N \quad , \quad \sum_{j=1}^{n_t}(x_{ji}/q_j^t)/\sum_{j=1}^{n_t}(1/q_j^t)=\bar{X}_i,\ i=1,...,p^*, \tag{3.8}$$

with $p^* \le p = \dim(x)$. When the expectation under the population distribution is linear, i.e., $E_p(y_i^t \mid x_i) = x_i'\beta^t$, one can replace the p^* constraints in the right hand side of (3.8) by the constraint $\frac{1}{N}\sum_{j=1}^{n_t}(x_j'\beta^t/q_j^t)=\bar{X}'\beta^t$ where $\bar{X}'=(\bar{X}_1,...,\bar{X}_{p^*})$, thus reducing the number of constraints. Note that this constraint contains also β^t.

Changing the base sampling weights $w_i=(1/\pi_i)$ such that they satisfy constraints of the form (3.8) and thus utilize knowledge of population means of auxiliary variables that are related to the response variable of interest is very common in survey sampling estimation. See, Deville and Sarndal (1992).

3.4. Estimation of population parameters

3.4.1. *Estimation based on the population model under strong ignorability*

In this section we focus on the estimation of the means $\mu^t=\frac{1}{N}\sum_{i=1}^{N}E(y_i^t \mid x_i)$, $t=1,...,m$. If the covariates x_i are known for every unit $i \in U$, then by (3.7),

$$\hat{\mu}^t=\frac{1}{N}\sum_{i=1}^{N}\hat{E}_p(y_i^t \mid x_i,\theta^t)=\frac{1}{N}\sum_{i=1}^{N}E_p(y_i^t \mid x_i;\hat{\theta}^t). \tag{3.9}$$

Note that if $E_p(y_i^t \mid x_i;\theta^t)$ is linear, the computation of (3.9) only requires knowledge of the population means $\bar{X}_i$. The estimator $\hat{\mu}^t$ can be used also for predicting the mean $\mu^{p,t}=\sum_{i=1}^{N}y_i^t/N$. If the initial sample is selected with equal probabilities, μ^t can be estimated by the sample mean, $\hat{\mu}_s^t=\sum_{j\in S}\hat{E}_p(y_i^t \mid x_i;\theta^t)/n$.

Remark 5: The estimator (3.9) looks similar to the estimator used for the estimation of the ATE defined by (2.3). Note, however, the estimator (3.9) accounts for an informative treatment assignment process and it does not assume strong ignorability.

3.4.2. Estimation based on estimated inclusion probabilities

The population parameters can be estimated also by use of Hajek (1971) estimators utilizing the estimated probabilities $\hat{q}_j^t$. The Hajek estimator is in common use in survey sampling applications. The resulting estimators have the form,

$$\hat{\mu}_H^{p,t} = \frac{\sum_{j\in S^t} y_j^t / \hat{q}_j^t}{\sum_{j\in S^t}(1/\hat{q}_j^t)} \quad ; \quad \hat{\mu}_H^t = \frac{\sum_{j\in S^t} \hat{E}_p(y_j^t \mid x_j;\theta^t)/\hat{q}_j^t}{\sum_{j\in S^t}(1/\hat{q}_j^t)}. \tag{3.10}$$

(Compare with (2.6)). Alternatively, one could use the 'doubly robustified' estimator,

$$\hat{\mu}_{DR}^t = \frac{\sum_{j\in S^t}[y_j^t - \hat{E}_p(y_j^t \mid x_j;\theta^t)]/\hat{q}_j^t}{\sum_{j\in S^t}(1/\hat{q}_j^t)} + \frac{1}{N}\sum_{i=1}^N \hat{E}_p(y_i^t \mid x_i,\theta^t). \tag{3.11}$$

(Compare with 2.7)). Large differences between the estimators in (3.9) and the estimators in (3.10) or (3.11) may indicate misspecification of either the population model under strong ignorability, or the treatment assignment probabilities. See Section 5 for a corresponding test statistic.

Remark 6: The estimators defined by (3.10) and (3.11) look similar to the estimators defined by (2.6) and (2.7), but as with the estimator (3.9), the estimators in (3.10) and (3.11) account for an informative treatment assignment process. This is reflected by the use of the probabilities $\hat{q}_j^t = \hat{Pr}(j \in S^t \mid y_j^t, x_j)$ instead of the propensity scores $\hat{e}_j^t = \hat{Pr}(j \in S^t \mid x_j)$.

4. Model identifiability

4.1. Identifiability problem

A major question underlying the use of the sample *pdf* (3.1) or (3.3) is model identifiability. By identifiability we mean the nonexistence of different pairs of population *pdf*s under strong ignorability and treatment assignment probabilities yielding the same sample *pdf*. Clearly, if different pairs exist, the model is not identifiable. At first thought it would seem that this is always the case since (3.1) for example is the sample *pdf* if the population *pdf* is $f_p(y_j^t \mid x_j)$ and the assignment probability is $\Pr(j \in S^t \mid y_j^t, x_j)$, but also if the population *pdf* is $f_{S^t}(y_j^t \mid x_j)$ and the units are assigned with equal probabilities. However, as shown below, under certain conditions, the sample *pdf* is generally identifiable.

In what follows we restrict to a single treatment t and assume that y_j^t is continuous. To simplify the notation in this section we denote by $q(y)$ the assignment probability to the sample S^t, denoted hereafter simply by S, and by

$f_p(y)$ the population *pdf* for treatment t under strong ignorability, assuming for convenience no covariates (see Remark 9 below). With this notation, the sample *pdf* for units in S is $f_s(y) = \dfrac{q(y) \cdot f_p(y)}{\int q(y) \cdot f_p(y) dy}$ and the identifiability of the sample model is defined as follows:

Model identifiability: The sample model $f_s(y)$ is identifiable if no different pairs $[f_p^{(1)}(y), q^{(1)}(y)]$, $[f_p^{(2)}(y), q^{(2)}(y)]$ exist that induce the same sample *pdf* $f_s(y)$.

4.2 Conditions for model identifiability

Suppose that there exist two treatment assignment probability rules (TAP) $q^{(1)}(y), q^{(2)}(y)$, and two *pdf*s $f_p^{(1)}(y), f_p^{(2)}(y)$ that are strictly positive on $J \subseteq \mathbf{R}$ yielding the same sample *pdf* $f_s(y)$, or equivalently,

$$\frac{q^{(1)}(y)}{q^{(2)}(y)} = K \frac{f_p^{(2)}(y)}{f_p^{(1)}(y)} \quad \forall y \in J \;\; ; \; K = \int q^{(1)}(y) \cdot f_p^{(1)}(y) dy \,/\, \int q^{(2)}(y) \cdot f_p^{(2)}(y) dy \,. \quad (4.1)$$

In what follows we assume that densities $f_p^{(1)}(y)$ and $f_p^{(2)}(y)$ that satisfy certain requirements are given, and define conditions under which no associated TAPs $q^{(1)}(y), q^{(2)}(y)$ exist that satisfy (4.1). This is done by studying the limit of each side of (4.1) as y tends to some limit point such as $+\infty, -\infty$ or 0, choosing the limit point in such a way that the left hand side of (4.1) converges to a finite positive number whereas the limit of the right hand side is either 0, ∞ or does not exist.

Remark 7: the use of this strategy enables to verify the identifiability of the sample *pdf* for many practical situations. Nonetheless, as shown later, there are other cases that need to be studied differently. Let $R_p(y) = f_p^{(2)}(y) / f_p^{(1)}(y)$.

Lemma 1 (similar to Lee and Berger, 2001): Assume that $J = [c, \infty)$ for some constant c. If the densities $f_p^{(1)}(y)$ and $f_p^{(2)}(y)$ are strictly positive on J and,

$$\lim_{y \to \infty} R_p(y) = 0 \text{, } \infty \text{ or does not exist,} \quad (4.2)$$

there are no $q^{(1)}(y), q^{(2)}(y)$ on J with finite positive limits at $y \to \infty$ satisfying (4.1).

Proof: Follows from (4.2) and taking the limit $y \to \infty$ on both sides of (4.1).

An example of (4.2) is two normal densities with different mean or variance. Another example is two Gamma densities with different location parameters. In both examples the limit of the ratio is either 0 or ∞. Examples of TAPs satisfying the requirement in the lemma are the Logistic and Probit functions with positive coefficients for the response values.

Lemma 2: Assume that $J=(-\infty,c]$ for some constant c. If $f_p^{(1)}(y)$ and $f_p^{(2)}(y)$ are strictly positive on J and,

$$\lim_{y\to-\infty} R_p(y)=0,\ \infty \text{ or does not exist,} \tag{4.3}$$

there are no $q^{(1)}(y), q^{(2)}(y)$ on J with finite positive limits at $y\to-\infty$ satisfying (4.1). The proof is similar to the proof of Lemma 1. Examples of (4.3) are two normal densities with different mean or variance or two double exponential (Laplace) densities with different location and scale parameters. In both examples the limit of the ratio is either 0 or ∞. Examples of TAPs satisfying the requirement in the lemma are the Logistic and Probit functions with negative coefficients for the response values.

Remark 8: When $q^{(j)}(y)=\dfrac{\exp(a_j+b_j y)}{1+\exp(a_j+b_j y)}$ and $f^{(j)}(y)=\mathrm{N}(\mu_j;\sigma_j^2)$, $j=1,2$, the sample model is identifiable by Lemma 1 If $b_j>0$, and by Lemma 2 if $b_j<0$.

Lemma 3: Assume that 0 is a limit point of J and that $f_p^{(1)}(y), f_p^{(2)}(y)$ are strictly positive in J and satisfy,

$$\lim_{y\to0^+\,(y\to0^-)} R_p(y)=0,\ \infty \text{ or does not exist.} \tag{4.4}$$

Then there are no $q^{(1)}(y), q^{(2)}(y)$ with finite positive limits at $y=0$ satisfying (4.1). The proof is again similar to the proof of Lemma 1. Examples of (4.4) are two Gamma *pdf*s with different location parameters or two Beta *pdf*s with different parameters. In both examples the limit of the ratio is either 0 or ∞, depending on the relative magnitude of the corresponding parameters. Logistic and Probit functions satisfy the requirement from the TAPs in the lemma.

Lemmas 1-3 cover many practical cases but as mentioned in Remark 7, there are other interesting and possibly practical cases that need to be studied separately. Below we consider cases where the TAPs are nonincreasing Logistic or Probit functions and the population densities are defined on the non-negative real line.

Case 1. *Logistic* assignment rules

Suppose that $q^{(j)}(y)=\dfrac{\exp(a_j+b_j y)}{1+\exp(a_j+b_j y)}$, $b_j<0,\ j=1,2$ and $J=[0,\infty)$. In this case the identity (4.1) can be expressed as,

$$\frac{1+\exp(a_2+b_2 y)}{1+\exp(a_1+b_1 y)}=K\frac{f_p^{(2)}(y)}{f_p^{(1)}(y)}\exp[(a_2-a_1)+(b_2-b_1)y],\ \forall y\in[0,\infty). \qquad (4.5)$$

The left hand side of (4.5) tends to 1 as $y\to\infty$. However, the limit of the right hand side depends on the forms of $f_p^{(1)}(y)$ and $f_p^{(2)}(y)$. In Appendix A we consider an example of two exponential densities.

Case 2. *Probit* assignment rules

Suppose that $J=[0,\infty)$ and $q^{(j)}(y)=\Phi(a_j+b_j y),\ b_j<0,\ j=1,2$, where $\Phi(\cdot)$ defines the normal cumulative *pdf*. The identity (4.2) is now,

$$\frac{\Phi(a_1+b_1 y)}{\Phi(a_2+b_2 y)}=K\cdot\frac{f_p^{(2)}(y)}{f_p^{(1)}(y)} \qquad \forall y\in[0,\infty). \qquad (4.6)$$

For y sufficiently large, the ratio $\dfrac{\Phi(a_1+b_1 y)}{\Phi(a_2+b_2 y)}$ can be bounded as (see Appendix B),

$$\frac{1-\varepsilon_1}{1+\varepsilon_2}\cdot\frac{a_2+b_2 y}{a_1+b_1 y}\cdot\frac{\varphi(a_1+b_1 y)}{\varphi(a_2+b_2 y)}<\frac{\Phi(a_1+b_1 y)}{\Phi(a_2+b_2 y)}<\frac{1+\varepsilon_1}{1-\varepsilon_2}\cdot\frac{a_2+b_2 y}{a_1+b_1 y}\cdot\frac{\varphi(a_1+b_1 y)}{\varphi(a_2+b_2 y)}, \qquad (4.7)$$

where $\varphi(\cdot)$ denotes the standard normal *pdf* and $\varepsilon_1,\varepsilon_2>0$ are arbitrarily small. Thus, by (4.7), and for y sufficiently large,

$$\frac{a_2+b_2 y}{a_1+b_1 y}\cdot\frac{1-\varepsilon_1}{1+\varepsilon_2}<K\frac{f_p^{(2)}(y)}{f_p^{(1)}(y)}\cdot\frac{\varphi(a_2+b_2 y)}{\varphi(a_1+b_1 y)}<\frac{a_2+b_2 y}{a_1+b_1 y}\cdot\frac{1+\varepsilon_1}{1-\varepsilon_2}. \qquad (4.8)$$

The left and right hand sides of (4.8) tend to (b_2/b_1) as $\varepsilon_1,\varepsilon_2\to 0; y\to\infty$. However, the limit of the middle part of (4.8) depends on the forms of the *pdf*s $f_p^{(1)}(y)$, $f_p^{(2)}(y)$. In Appendix C we consider an example of two exponential densities.

Remark 9: So far we studied the identifiability of the sample model assuming that there are no covariates. In practice, both the probability assignment rule and the population *pdf* may depend on observable covariates x. For example, in the empirical analysis in Section 6 we use, $q(y;c,\delta,\gamma)=\dfrac{\exp(c+\delta y+x'\gamma)}{1+\exp(c+\delta y+x'\gamma)}$;

$f_p(y;\beta,\sigma^2) = \mathrm{N}(x'\beta;\sigma^2)$. Evidently, the identifiability arguments presented above apply to this case as well, provided that the covariate values are sufficiently spread to allow the identification of their coefficients. See Cox and Snell (1989, Section 3.4.3) for related discussion.

5. Model Assessment

Assessing the goodness of fit of an estimated model is an old problem underlying almost every statistical application. This is particularly imperative for models of the form (3.1) or (3.3) as both the distribution under strong ignorability and the assignment probabilities are generally unknown. On the other hand, once the identifiability of the sample *pdf* has been established, there is nothing unique in the present case and one faces the classical problem of having a random sample from an hypothesized *pdf* which has to be tested. Below we overview a few plausible test statistics that can be used for assessing the goodness of fit of the sample *pdf*.

5.1. Compare theoretical and empirical distributions

Once the model parameters $\{\alpha^t,\theta^t\}$ have been estimated, the cumulative sample distribution function (*cdf*) for sample unit $j \in S^t$ can be estimated as,

$$\hat{F}_j^t(y \mid x_j) = \int_{-\infty}^{y} f_{S^t}(y_j^t \mid x_j;\hat{\alpha}^t, \hat{\theta}^t)dy_j^t . \tag{5.1}$$

The 'expected' mean number of sample units with observations $y_j^t \le y$ under the hypothesized model is therefore, $\hat{F}_{S^t}(y;\hat{\alpha}^t, \hat{\theta}^t) = \sum_{j\in S^t} \hat{F}_j^t(y \mid x_j)/n_t$, which can be compared to the empirical proportion $\hat{F}_{EMP}^t(y) = \frac{1}{n_t}\sum_{j\in S^t} \mathrm{I}(y_j^t \le y)$, where $\mathrm{I}(y_j^t \le y)$ is the indicator function. The null hypothesis that the sample model fits the sample data can be tested by use of the Kolmogorov-Smirnov (KS) test statistic,

$$KS_t = \max_{y_i^t \in S^t} | \hat{F}_{EMP}^t(y_j^t) - \hat{F}_{S^t}(y_j^t;\hat{\alpha}^t, \hat{\theta}^t) |. \tag{5.2}$$

The KS test is known to be nonparametric, but this is only true if the parameters of the theoretical distribution are known. Otherwise, the distribution of the KS statistic depends in a complex way on the true values of the model parameters. Correct critical values can be obtained by use of parametric bootstrap. The procedure consists of generating many samples from the estimated hypothesized model, re-

estimating the unknown parameters from each bootstrap sample and the corresponding KS statistic, and then computing the critical values based on the bootstrap distribution of the KS statistic. See Babu and Rao (2004) for regularity conditions justifying the use of this procedure.

Another possibility of comparing the hypothesized distribution with the empirical sample distribution is by using the Moran (1951) test. Let $y_{(1)} < y_{(2)} < ... < y_{(n^t)}$ denote the ordered values of the response and let $p_{(i)} = \sum_{j \in S^t} \hat{F}_j^t(y_{(i)} \mid x_j)/n_t$, where $\hat{F}_j^t$ is defined by (5.1). Compute the differences (spacings), $D_i(\hat{\alpha}^t, \hat{\theta}^t) = p_{(i)} - p_{(i-1)}$, $i = 1...n_t$ with $p_0 = 0$, $p_{n_t} = 1$. The Moran test statistic is,

$$M(\hat{\alpha}^t, \hat{\theta}^t) = -\sum_{i=1}^{n^t} \log D_i(\hat{\alpha}^t, \hat{\theta}^t). \tag{5.3}$$

Cheng and Stephens (1989) show that under mild regularity conditions the statistic defined by (5.3) has asymptotically ($n_t \to \infty$) normal distribution with mean γ_m and variance σ_m^2, given, up to the order of m^{-1} by,

$$\gamma_m = m(\log m + \gamma) - \frac{1}{2} - \frac{1}{12m},\ \sigma_m^2 = m\left(\frac{\pi^2}{6} - 1\right) - \frac{1}{2} - \frac{1}{6m}, \tag{5.4}$$

where $m = n_t + 1$, and $\gamma \approx 0.5772$ is the Euler's constant. This property makes the test very attractive but its performance is known to be sensitive to the existence of 'close observations'. Cheng and Stephens (1989) propose modifications for the case of tied observations.

5.2. Compare estimates obtained from the estimated population distribution with estimates based on the estimated assignment probabilities

Section 3.4 considers two alternative methods of estimating the population parameters μ^t and $\mu^{p,t}$. The first method uses the estimated population *pdf* (Equation 3.9). The second method uses the estimated inclusion probabilities (Equations 3.10, 3.11). If the parametric forms of the population distribution under strong ignorability and the conditional expectations of the inclusion probabilities are correctly specified, we expect the two sets of estimators to be sufficiently close. Large differences would indicate that at least one of the models is misspecified. For

a given treatment *t*, we may test for example, $H_0 : \Delta^t = E(\hat{\mu}^t_{DR} - \hat{\mu}^t) = 0$ using the test statistic,

$$U^t = (\hat{\mu}^t_{DR} - \hat{\mu}^t) / S\hat{D}(\hat{\mu}^t_{DR} - \hat{\mu}^t) \tag{5.5}$$

Note that by (3.9) and (3.11), $\hat{\Delta}^t = \hat{\mu}^t_{DR} - \hat{\mu}^t = \dfrac{\sum_{j \in S^t} [y^t_j - \hat{E}_p(y^t_j \mid x_j ; \theta^t)] / \hat{q}^t_j}{\sum_{j \in S^t} (1/\hat{q}^t_j)}$

$= \sum_{j \in S^t} \hat{e}^t_j / \hat{q}^t_j / \sum_{j \in S^t} (1/\hat{q}^t_j)$. Under correct assignments, $\dfrac{E_{S^t}(e^t_i / q^t_i)}{E_{S^t}(1/q^t_i)} = E_p(e^t_i)$,

where $E_{S^t}(\cdot)$ is the expectation under the sample distribution (3.3) (Pfeffermann and Sverchkov, 1999), such that $\hat{\Delta}^t$ is asymptotically unbiased for the population mean of the residuals in treatment *t*, and $E_p(e^t_i) = 0$ if the population model for treatment *t* is specified correctly.

The asymptotic distribution of $\hat{\Delta}^t$ under correct model specification is obtained by noting that it is the solution of the estimating equations, $\sum_{j=1}^{n} T_j (e^t_j - \Delta^t) / q^t_j = 0$;

$\sum_{j=1}^{n} u(y^t_j, x_j, T_j, \alpha^t, \theta^t) = 0$, where $T_j = 1$ if $j \in S^t$ and zero otherwise, and $u(\cdot)$ is the score function with the likelihood defined by (3.6). Noting that $E[\sum_{j=1}^{n} T_j (e^t_j - \Delta^t) / q^t_j$,

$\sum_{j=1}^{n} u(y^t_j, x_j, T_j, \alpha^t, \theta^t)] = 0$ at the true parameter values $\alpha^t, \theta^t, \Delta^t$ under the joint distribution of (y^t_j, T_j), it follows from the theory of M-estimation (Stefansky and Boos, 2002) that

$$\sqrt{n}(\hat{\Delta}^t - \Delta) \xrightarrow{D} N(0, \Sigma). \tag{5.6}$$

In order to estimate Σ and apply the statistic U^t in (5.5), note that the score function can be written as $u(y^t_j, x_j, T_j, \alpha^t, \theta^t) = T_j g(y^t_j, x_j, \alpha^t, \theta^t) + (1 - T_j) s(x_j, \alpha^t, \theta^t)$, where $g(\cdot)$ and $s(\cdot)$ are the derivatives of the corresponding log-likelihood expressions. After some algebra and following the theory of M-estimation we obtain that,

$\Sigma = v_{11} - 2h'\mathrm{I}^{-1}v_{21} + h'\mathrm{I}^{-1}h$, with $v_{11} = \frac{1}{n}\sum_{j=1}^{n} E_p[(e_j^{\,t} - \Delta^t)^2 / q_j^{\,t}]$,

$v_{12} = \frac{1}{n}\sum_{j=1}^{n} E_p[(e_j^{\,t} - \Delta^t) g(y_j^{\,t}, x_j)]$, $\mathrm{I} = \mathrm{I}(\phi^t) = \frac{1}{n}\sum_{j=1}^{n} E\frac{\partial u_j}{\partial \phi^t}$; $\phi^t = (\alpha^t, \theta^t)$ is the Fisher Information matrix and $h' = -\frac{1}{n}\sum_{j=1}^{n} E(\partial[T_j \frac{e_j^{\,t} - \Delta^t}{q_j^{\,t}}] / \partial\phi^t)$. The estimator $\hat{\Sigma}$ is obtained by estimating $E_p(\cdot)$ by the corresponding sample value, substituting the unknown parameters by their sample estimates and estimating the population totals by inverse probability weighting. The estimator $\hat{\Sigma}$ is consistent for Σ under mild regularity conditions, Iverson and Randles (1989).

5.3. Assess the coherence of estimated propensity scores for different treatments

Since every sample unit is assigned to one and only one of the treatments, under correct model specification $\sum_{t=1}^{m} \Pr(j \in S^t \mid x_j) = 1$ for every unit j, where $\Pr(j \in S^t \mid x_j) = E(q_j^t \mid x_j) = \int E(q_j^t \mid y_j^t, x_j) f_p(y_j^t \mid x_j) dy_j^t$ is the propensity score (denominator of 3.1 or 3.3). Thus, one can test the sample models by testing the null hypothesis, $H_0 : \sum_{t=1}^{m} \Pr(j \in S^t \mid x_j) = 1$ for all j. A plausible test statistic is therefore,

$$M_s = \max_{j \in S} |1 - \sum_{t=1}^{m} \hat{\Pr}(j \in S^t \mid x_j)|, \qquad (5.7)$$

where $\hat{\Pr}(j \in S^t \mid x_j) = \hat{E}(q_j^t \mid x_j)$. Note in this respect that the sample models are fitted independently for each treatment (see Section 3.2).

The distribution of the test statistic (5.7) under the null hypothesis has yet to be established (and possibly approximated by use of parametric bootstrap), and its use is restricted therefore at this stage to descriptive analysis.

6. Application of the new approach to the PISA survey

6.1. Data used for present application

We study the performance of the proposed approach and compare it to the other methods described in Section 2, using data collected in *Ireland* in the year 2000 by OECD for the *Programme for International Student Assessment* (PISA). The purpose

of this program is to study the proficiency of pupils aged 15 in mathematics, science and reading in 34 countries.

6.2. Sampling design

The sampling design underlying the PISA study is in most countries a stratified two-stage sampling design. The strata are defined by size, type of school and gender composition. Within each stratum, the first stage of sampling is a probability proportional to size (PPS) sample of schools with the size defined by the 'anticipated' number of 15 years old pupils enrolled in the school. A minimum of 150 schools has been selected in each country (or all the schools if there are less than 150 schools in the country). The second stage consists of an equal probability sample of 35 pupils from the corresponding age group in each of the sampled schools (or all the pupils in schools with less than 35 pupils aged 15).

By this sampling design, pupils included in the sample in a given country are not equally representative of the pupils aged 15 in the country and each pupil is assigned therefore a sampling weight. The weight is the reciprocal of the product of the school inclusion probability and the pupil's inclusion probability within his school, adjusted for non-participation of schools and nonresponse of pupils. We performed some of the analyses described below incorporating the weights but found that it had no effect on the values of the estimates, implying that the sample selection is noninformative for the models we use. For more information on the PISA sampling design and weighting see PISA 2000 Technical Report, Chapters 4 and 6.

In the present application we compare proficiency scores in mathematics between public schools and private schools in Ireland. This is a good example of an observational study because pupils attending the two types of schools are different in their family background and other important characteristics. The whole dataset has been analyzed previously by Vandenberghe and Robin (2004) using existing methods. The data from Ireland is of particular interest because different existing methods provide ATE estimates with opposite sings (see Section 6.5 and Vandenberghe and Robin, 2004). The sample data refers to 1256 students in private schools ($t = 1$) and 702 students in public schools ($t = 0$).

6.3. Computation of response values

The response value in the PISA study (proficiency in mathematics in the present application) is not observed directly even for sampled pupils and is treated as a

missing value. PISA uses two approaches for imputing the missing proficiencies: a maximum likelihood approach and a multiple imputation approach. Let the binary variable d_{ij} take the value 1 if pupil j answers correctly question i of the PISA examination and 0 otherwise. The probability $\Pr(d_{ij}=1)$ is the logistic probability, $\Pr(d_{ij}=1 \mid a_i, b_i, \psi_j) = [1+\exp(-a_i(\psi_j - b_i)]^{-1}$. The parameter a_i measures how question i distinguishes between persons of different proficiency; the parameter b_i represents the 'difficulty' of question i and ψ_j is the unobserved proficiency score. The imputed score for student j is the MLE $\hat{\psi}_j$. Note that the logistic models have no covariates, implying conditional independence of the answers on background characteristics, given the score ψ_j.

The second approach draws at random multiple values from the conditional distribution of ψ_j given the indicators $\underline{d}_j = (d_{1j},...,d_{mj})$, ($m$ is the number of questions), and covariates x_j representing individual background characteristics like age and gender. The conditional *pdf* of ψ_j given $\underline{d}_j$ and x_j is expressed as,

$$f(\psi_j \mid d_j, x_j) \propto \prod_{i=1}^{m} [\Pr(d_{ij}=1)]^{d_{ij}} [\Pr(d_{ij}=0)]^{(1-d_{ij})} f(\psi_j \mid x_j, \lambda, \sigma), \qquad (6.1)$$

where $\Pr(d_{ij}=1 \mid a_i, b_i, \psi_j)$ is modeled as above and $f(\psi_j \mid x_j, \lambda, \sigma^2)$ is the normal distribution with mean $x_j'\lambda$ and variance σ^2. Note that the responses to the various questions are assumed to be independent given the parameters $\eta_{ij} = (a_i, b_i, \psi_j)$. Five imputed values of ψ_j are drawn for every student j in the sample.

In the present application we use the second approach and following Vandenberghe and Robin (2004) we standardized the values by dividing them by their empirical standard deviation. The use of this approach enables estimating the variances of the ATE estimates using multiple imputation theory (Rubin, 1987). Denote by $\hat{ATE}_d$ the ATE estimate from imputed data set d, $d = 1,...,5$. Following the theory of multiple imputation,

$$\hat{ATE} = \sum_{d=1}^{5} \hat{ATE}_d / 5 \;;\; \hat{Var}(\hat{ATE}) = (1+1/5)B + V, \qquad (6.2)$$

where $B = \sum_{d=1}^{5} (A\hat{T}E_d - A\hat{T}E)^2 / 4$ is the 'between' imputation variance and $V = \sum_{d=1}^{5} \hat{V}_d / 5$ is the 'within' imputation variance, $\hat{V}_d = V\hat{a}r(A\hat{T}E_d)$. For the ATE estimator $(\hat{\mu}^1 - \hat{\mu}^0)$ with $\hat{\mu}^t$ defined by (3.9) we computed $\hat{V}_d$ using the estimated inverse information matrix. For the ATE estimator $(\hat{\mu}^1_{DR} - \hat{\mu}^0_{DR})$ with $\hat{\mu}^t_{DR}$ defined by (3.11), we estimated $\hat{V}_d$ similarly to the estimation of Σ in Section 5.2. Note that $\hat{\mu}^1_{DR}$ and $\hat{\mu}^0_{DR}$ are independent since they refer to different treatments.

6.4. Model for PISA data

In the analysis that follows we model the sample *pdf* (3.1) by assuming a normal distribution for the potential population responses and the logistic model for the assignment probabilities. Thus, using the notation of Section 3,

$$f_p(y_j^t \mid x_j) = N(x_j'\beta^t, \sigma_t^2)\,; \Pr(j \in S^t \mid x_j, y_j^t) = \frac{\exp(c^t + \delta^t y_j^t + x_j'\gamma^t)}{1 + \exp(c^t + \delta^t y_j^t + x_j'\gamma^t)},\ t = 0,1. \quad (6.3)$$

In (6.3) $t = 0$ defines public schools and $t = 1$ private schools. As shown in Section 4, the sample *pdf* is identifiable for $\delta^t \neq 0$ (see Remark 8).

6.4.1 Explanatory variables

Six explanatory variables (covariates) were found to be significant in at least one of the models fitted to the PISA data. Gender (1 for girls 0 for boys), father's education (F.E= 1 for high education, 0 otherwise), family socio-economic index (S.E.I), index of home educational resources (H.E.R), average socio-economic index of the student's schoolmates (S.E.S, proposed by Vandenberghe and Robin, 2004 to account for potential peer effects), and school location (S.loc= 1 if school located in an urban area, 0 otherwise). The continuous variables have been standardized.

Remark 10: Vandenberghe and Robin (2004) considered additional variables, but these were not found to be significant in our analysis.

Remark 11: The variable school location was used by Vandenberghe and Robin (2004) as an instrumental variable. The authors fit the model (2.13) but impose $\beta^1 = \beta^0 = \beta$. They show that it has a significant effect on the probability of attending private schools in all the countries (thus satisfying Condition *c* in Section 2.2.2). However, the approaches considered in the literature for observational studies do

not permit testing directly the other requirement from an instrumental variable that the school location is exogenous to the student's proficiency given the model covariates (Condition *a*). The authors claim that this requirement is plausible using similar arguments to Hoxby (2000). As mentioned in Section 2.3, the use of our approach enables testing this requirement (see below).

6.4.2 Computational details

We computed the maximum likelihood estimates of the unknown parameters by maximizing the full likelihood (3.6) with respect to $\theta^t = (\beta^t, \sigma_t)$; $\alpha^t = (c^t, \delta^t, \gamma^t)$. For this, we used the maximization routine *nlm* in R (Development Core Team (2004)). The choice of the initial values plays a crucial role in the convergence of the maximization algorithm. However, empirical investigations show that for a fixed value of the coefficient δ^t, the maximization is not sensitive to the choice of the initial values for the other parameters. We applied therefore the following algorithm which performs well in our application.

1. Define a grid of plausible values for δ^t around zero. Maximize the likelihood for each value δ^t with respect to the other parameters using as initial values for β^t and σ_t the values obtained by fitting a linear regression model to the sample data and zeroes for c^t and γ^t. The parameters maximizing the likelihood over all the grid values of δ^t are taken as the initial values.

2. Maximize the likelihood with respect to all the parameters (including δ^t) with initial values obtained in Step 1.

6.5 Results

Tables A1-A4 show the estimates and standard errors (Std. Error) obtained for the private and public schools. Note that $\hat{\delta}^1 > 0$, $\hat{\delta}^0 < 0$, but $\hat{\delta}^1$ is close to zero and not significant. On the other hand, $\hat{\delta}^0$ is far from zero and highly significant, indicating that for given values of the covariates, the probability to attend a public school decreases very rapidly as the score increases. This finding suggests that pupils attending public schools have a priori lower scores, and not because of a poor quality of public schools. Note also that the instrument, school location, is

nonsignificant in Tables A2 and A4 but highly significant in Tables A1 and A3. We discuss this outcome in Section 6.6.

Table A5 shows the estimates of the population means by type of school and the estimates of the ATE as obtained under our approach. We show the two estimates considered in Section 3.3: the estimate (3.9) that is based on the estimated population model and the doubly-robustified estimate (3.11). The two ATE estimates are similar, negative and very significant, indicating the very interesting and somewhat surprising result that the mean proficiency in public schools after accounting for the school selection process is actually higher in public schools than in private schools. Table A6 shows the ATE estimates obtained by some of the existing methods reviewed in Section 2, using Stata (StataCorp, 2004) and R packages (R Development Core Team, 2004). For the propensity score matching method we used a one-to-one matching algorithm with replacement (see Section 2.1.2). For the control functions method we used the two-step Heckman's (1979) method, assuming that (u_v, u^0, u^1) is trivariate Normal (see Section 2.2.1). Notice that unlike the ATE estimates in Table A5, the crude difference between the unadjusted sample means in the two types of schools is positive, suggesting that the mean proficiency is higher in private schools than in private schools. This outcome illustrates the problem of observational studies very pronouncedly. All the methods except for the method of instrumental variables yield very small ATE estimates.

Table A7 shows the p-values of the goodness of fit test statistics discussed in Section 5. The first 3 statistics are nonsignificant with p-values higher than 12%, thus supporting the use of the selected models. As mentioned in Section 5.3, the theoretical critical values of the M_s statistic are unknown but notice its very low value. Computing the critical values by parametric bootstrap yields a p-value of 0.30.

6.6. *Testing of assumptions of existing methods*

We mentioned before that the use of the proposed approach enables testing some of the assumptions underlying the existing methods. Note first that the coefficient of y is not significant in the logistic model for the private schools, thus seemingly supporting the use of methods that use the propensity scores. However, the coefficient of y is highly significant in the logistic model for public schools, indicating that the covariates used in this study do not fully explain the choice of

public schools and hence that the use of methods that use the propensity scores with these covariates is not valid.

Next consider the instrument, 'school location'. We notice in Tables A2 and A4 that the coefficient of the instrument is not significant in the two population models, implying that Condition *b* underlying the use of instrumental variables is satisfied (Section 2.2.2). Similarly, the instrument is highly significant in the two logistic models (Tables A1 and A3) as assumed under Condition *c*. However, in the public schools the assignment probabilities depend heavily on y, despite of including in the model the covariates and the instrument, indicating that Condition *a* is not satisfied and hence that the school location is not a proper instrument. Note, however, that the use of the method of instrumental variables with this instrument yields the closest ATE estimate to the estimate obtained under the new approach.

6.7 *Simulation study*

The simulation study is divided into two parts. In the first part we generated independently 400 data sets from the model fitted to the data from Ireland when the response values are the averages of the five imputed values (see section 6.3). The sample sizes for the two types of schools were the same as in the original samples. This part of the simulation study is therefore an application of parametric bootstrap and it was carried out in order to study the performance of the proposed approach and as another validation of the empirical results reported in Section 6.5. The simulations allowed us also to compute the critical values of the KS test statistic (Section 5.1) and of the M_S test statistic (Section 5.3; as noted there, the validity of the use of parametric bootstrap for calculating the distribution of the M_S statistic has yet to be studied). In order to save in space we don't show the empirical means and standard deviations of the model parameter estimates obtained for the 400 runs but the means are generally very close to the true parameters and the standard deviations are close to the standard errors computed for the original sample.

Table B1 shows the empirical means of the estimates of the population means and the ATE over the 400 simulations, and the corresponding empirical standard deviations (Std) of the means. The row titled "original sample" shows the estimates obtained for the original samples from Ireland. Notice that these estimates are mildly different from the estimates in Table A5 because the response values are now the averages of the five imputed values. As is evident, the empirical means are close to

the original estimates, even though the differences are 'significant'. Interesting, the empirical means are almost identical to the values shown in Table A5, which are also means over the estimates obtained for the five separate sets of imputed values. Table B2 shows the empirical means of the ATE estimates obtained under the existing methods. As can be seen, the empirical means always have the same sign as the original estimates shown in Table A6, and except in two cases the mean estimates and the original estimates are close. The empirical standard deviations are close to the standard errors shown in Table A6.

All in all, the results obtained for this part of the simulation study show that indeed the model parameters can be estimated almost unbiasedly with acceptable standard error estimates, despite the rather complicated structure of the sample model. Obtaining similar ATE estimates under the existing and the new method as for the original samples can be used as another indication of the goodness of fit of the models fitted to the data from Ireland and the corresponding ATE estimates.

In the second part of the simulation study we generated independently 400 other data sets from the same model as above, except that the residual error terms in the two populations models (public and private schools) were generated from a *t*-distribution with 4 degrees of freedom instead of the normal distribution. For each set we fit the model that assumes normal error terms, like in the first part. This part of the simulation study was carried out mostly in order to study the performance of the goodness of fit test statistics under a misspecified model. The $t_{(4)}$ *pdf* is not very far from the N(0,1) *pdf* and yet, we find that in this case some of the parameter estimates are highly biased, interestingly, more so in the two logistic models, despite the fact that these models have not been changed. Table C1 shows the empirical means of the estimates of the population means and the ATE as obtained under the misspecified model. As can be seen, the empirical means are biased in this case but the biases are not extreme.

Table C2 shows the percentage of samples for which the goodness of fit tests rejected the misspecified model at the 5% nominal level. These percentages indicate the power of the various tests. The KS test basically rejects the misspecified model in all the samples from private schools and in 60% of the samples from public schools. The statistic U^t has somewhat better power than KS in public schools but very low power in private schools. The Moran test has low

power in both types of schools. Thus, KS shows overall the best performance, and with the mild model misspecification considered, a power of 60% as obtained for the public schools is not unexpected.

7. Discussion

In this article we propose a new approach for observational studies that recovers the treatment assignment model and the population model, before the assignment, from the sample data. On first thought, this seems impossible but we show in Section 4 that the sample model holding for the observed data, which incorporates the population model and the assignment probabilities, is identifiable under mild conditions. Furthermore, the goodness of fit of the sample model can be tested by standard test statistics because the sample model refers to the sample data. We develop also in Section 5 a new test that compares the estimate of the population mean obtained under the recovered population model, with an estimate of the population mean that uses the estimated assignment probabilities.

The advantage of the proposed approach over existing methods that use the propensity scores or instrumental variables for estimating the treatment effects is that it does not require knowledge of the covariates or instruments that explain the assignment to treatments. Moreover, as illustrated in Section 6, the use of the new method actually enables to test the appropriateness of the use of these methods.

We applied the new approach for comparing the proficiency scores in mathematics of children aged 15 between public and private schools in Ireland. Our analysis shows that although the average score of pupils in the sample from private schools is significantly higher than the average score of pupils from public schools, the picture is reversed once the effect of the school selection is accounted for properly. A similar conclusion is reached by application of the method of instrumental variables, but the difference between the two types of schools is more profound under the new method.

TABLES

A. PISA data in Ireland

The model was fitted for each set of imputed responses separately. The results in Tables A1-A6 are obtained using the theory of multiple imputation described in Section 6.2. Table A7 refers to a single data set with the responses defined by the mean of the five imputed responses (after standardization).

Private schools

Table A1. Assignment (logistic) model for private schools

Coefficient	C^1	δ^1	Gen.	F.edu	S.E.I	H.E.R	S.E.S	S.loc
Estimate	-1.45	0.23	0.70	0.04	- 0.08	3.28	0.14	1.13
Std. Error	1.55	0.25	0.13	0.12	0.08	0.20	0.07	0.13

Table A2. Population (normal) model for private schools

Parameter	σ_1	Const	Gen.	F.edu	S.E.I	H.E.R	S.E.S	S.loc
Estimate	0.88	6.24	-0.23	0.17	0.16	0.34	0.19	- 0.09
Std. Error	0.02	0.11	0.06	0.06	0.03	0.11	0.03	0.06

Public schools

Table A3. Assignment (logistic) model for public schools

Coefficient	C^0	δ^0	Gen.	F.edu	S.E.I	H.E.R	S.E.S	S.loc
Estimate	15.08	-2.18	-0.78	0.17	0.35	-2.85	0.29	-1.57
Std. Error	2.80	0.55	0.20	0.23	0.14	0.49	0.14	0.26

Table A4. Population (normal) model for public schools

Parameter	σ_0	Const	Gen.	F.edu	S.E.I	H.E.R	S.E.S	S.loc
Estimate	1.20	6.98	0.18	0.10	0.16	1.48	0.31	0.23
Std. Error	0.10	0.18	0.10	0.09	0.05	0.31	0.04	0.15

Table A5. Estimation of population means and ATE

	Private School		Public School		ATE	
	$\hat{\mu}^1 = \overline{x}'\hat{\beta}^1$	$\hat{\mu}^1_{DR}$	$\hat{\mu}^0 = \overline{x}'\hat{\beta}^0$	$\hat{\mu}^0_{DR}$	$\hat{\Delta} = \hat{\mu}^1 - \hat{\mu}^0$	$\hat{\Delta}_{DR}$
Estimate	6.16	6.16	7.22	7.05	-1.05	-0.89
Std. Error	0.08	0.08	0.25	0.20	0.25	0.23

Table A6. Estimation of ATE by existing methods

Method	$\overline{y}^1 - \overline{y}^0$	Reg.	Propens. Matching	Hajek	Doubly Rob.	Instrum. Variables	Control Function
Estimate	0.354	0.130	0.214	0.160	0.170	- 0.726	- 0.166
Std. Error	0.045	0.052	0.103	0.051	0.052	0.247	0.130

Table A7. Goodness of fit test statistics and p-values (in parentheses)

Statistics	KS	Moran	U^t	M_S
Private schools	0.0218 (0.125)	- 0.4643 (0.64)	-0.48 (0.64)	0.04812
Public schools	0.0399 (0.166)	- 0.1647 (0.87)	-1.24 (0.22)	

B. Simulations from model fitted to data from Ireland (400 simulated data sets)

Table B1. Estimation of population means and ATE

	Private School		**Public School**		**ATE**	
	$\hat{\mu}^1 = \overline{x}'\hat{\beta}^1$	$\hat{\mu}^1_{DR}$	$\hat{\mu}^0 = \overline{x}'\hat{\beta}$	$\hat{\mu}^0_{DR}$	$\hat{\Delta} = \hat{\mu}^1 - \hat{\mu}^0$	$\hat{\Delta}_{DR}$
Original sample	**6.17**	**6.17**	**7.27**	**7.11**	**-1.10**	**-0.94**
Emp. mean	6.17	6.17	7.20	7.07	- 1.02	- 0.90
Std of mean	0.005	0.005	0.006	0.009	0.007	0.01

Table B2. Estimation of ATE by existing methods

Method	$\overline{y}^1 - \overline{y}^0$	**Reg.**	**Propens. Matching**	**Hajek**	**Doubly Rob.**	**Instrum. Variables**	**Control Function**
Emp. mean	0.361	0.134	0.126	0.158	0.167	-0.676	-0.365
Emp. Std	0.044	0.051	0.100	0.048	0.048	0.244	0.231

C. Simulations from misspecified model (400 simulated data sets)

Table C1. Estimation of population means and ATE

	Private School		**Public School**		**ATE**	
	$\hat{\mu}^1 = \overline{x}'\hat{\beta}^1$	$\hat{\mu}^1_{DR}$	$\hat{\mu}^0 = \overline{x}'\hat{\beta}^0$	$\hat{\mu}^0_{DR}$	$\hat{\Delta} = \hat{\mu}^1 - \hat{\mu}^0$	$\hat{\Delta}_{DR}$
Original Sample	**6.17**	**6.17**	**7.27**	**7.11**	**-1.10**	**-0.94**
Emp. mean	6.06	5.98	7.58	7.19	-1.52	-1.21
Std of mean	0.01	0.015	0.015	0.03	0.018	0.031

Table C2. Percentage of rejection of misspecified model at 5% nominal level

Statistics	KS	Moran	M_S	U^t
Private School	99.7%	49%	53.8%	16%
Public School	59.7%	19.5%		66%

Appendix A: Identifiability of the sample *pdf* when the population *pdf* is exponential and the assignment rule is logistic

Suppose that $f_p^{(j)}(y) = \theta_j \exp(-\theta_j y),\ q^{(j)}(y) = \dfrac{\exp(a_j + b_j y)}{1+\exp(a_j + b_j y)},\ b_j < 0, j = 1,2$;

$J = [0,\infty)$. The right hand side of (4.5) is therefore,

$G(y) = K\dfrac{\theta_2}{\theta_1}\exp[(a_2 - a_1) + (b_2 - b_1 + \theta_1 - \theta_2)y]$. If $b_2 - b_1 \neq \theta_2 - \theta_1$, letting $y \to \infty$ on both sides of (4.5) yields a contradiction. If $b_2 - b_1 = \theta_2 - \theta_1$, (4.5) takes the form,

$$\frac{1+\exp(a_2 + b_2 y)}{1+\exp(a_1 + b_1 y)} = K\frac{\theta_2}{\theta_1}\exp(a_2 - a_1)\ ,\forall y \in J\ .$$

Differentiating both sides with respect to y shows that it can only hold if $a_1 = a_2, b_1 = b_2$ and $\theta_1 = \theta_2$, establishing the identifiability of the sample *pdf*.

Appendix B: Bounds on the ratio of two probit assignment probabilities

In order to bound the ratio $\dfrac{\Phi(a_1 + b_1 y)}{\Phi(a_2 + b_2 y)}$ $(b_1, b_2 < 0)$, we use the following results:

1. $\lim\limits_{x\to-\infty}\dfrac{-x\Phi(x)}{\varphi(x)} = 1$ (Feller, 1968, pp. 175), 2. If $a \geq b > 0$, $c > d > 0$, then $ac > bd$.

It follows from Result 1 that for $\varepsilon > 0$ and sufficiently small negative x,

$$\frac{1-\varepsilon}{-x}\cdot\varphi(x) < \Phi(x) < \frac{1+\varepsilon}{-x}\cdot\varphi(x)\ .$$

The bounds in (4.7) follow after some algebra using Result 2.

Appendix C: Identifiability of the sample pdf when the population *pdf* is exponential and the assignment rule is probit.

Let $f_p^{(j)}(y) = \theta_j \exp(-\theta_j y),\ q^{(j)}(y) = \Phi(a_j + b_j y),\ b_j < 0,\ j = 1,2;\ J = [0,\infty)$. The middle part of (4.8) is therefore

$$G(y) = K\frac{\theta_2}{\theta_1}\exp[(a_1^2 - a_2^2)/2 + (\theta_1 + a_1 b_1 - \theta_2 - a_2 b_2)y + ((b_1^2 - b_2^2)/2)y^2].$$

If $\theta_1 + a_1 b_1 - \theta_2 - a_2 b_2 \neq 0$ or $b_1^2 - b_2^2 \neq 0$, taking the limit of (4.8) when $y \to \infty$ yields a contradiction. If $\theta_1 + a_1 b_1 - \theta_2 - a_2 b_2 = 0$ and $b_1^2 - b_2^2 = 0$ (equivalent to $b_1 = b_2 = b$), then for y sufficiently large (4.8) takes the form,

$$\frac{a_2 + by}{a_1 + by}\cdot\frac{1-\varepsilon_1}{1+\varepsilon_2} < K\frac{\theta_2}{\theta_1}\exp[(a_1^2 - a_2^2)/2] < \frac{a_2 + by}{a_1 + by}\cdot\frac{1+\varepsilon_1}{1-\varepsilon_2}$$

Letting $y \to \infty$ we get $K\frac{\theta_2}{\theta_1}\exp[(a_1^2 - a_2^2)/2] = 1$. Also, by (4.6) at $y = 0$,

$\frac{\Phi(a_1)}{\Phi(a_2)} = K\frac{\theta_2}{\theta_1}$. Thus, $\frac{\Phi(a_1)}{\Phi(a_2)} = \frac{\exp(-a_1^2/2)}{\exp(-a_2^2/2)} \Leftrightarrow \frac{\Phi(a_1)}{\Phi(a_2)} = \frac{\varphi(a_1)}{\varphi(a_2)} \Leftrightarrow \frac{\varphi(a_1)}{\Phi(a_1)} = \frac{\varphi(a_2)}{\Phi(a_2)}$

$\Leftrightarrow a_1 = a_2$ since $\lambda(x) = \frac{\varphi(x)}{\Phi(x)}$ is a one-to-one function. It follows that $a_1 = a_2,\ b_1 = b_2$ and $\theta_1 = \theta_2$, establishing the identifiability of the sample *pdf* in this case.

References

Abadie, A. and Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, **74** (1), 235-267.

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**, 444-472.

Babu, G.J. and Rao, C.R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhya, Series A*, **66** (1), 63-74.

Chambers, R.L. Dorfman, A. and Sverchkov, M. (2003). Nonparametric regression with complex survey data. In, *Analysis of Survey Data*. Ed. C Skinner and R. Chambers. New York: Wiley

Cheng, R.C.H., and Stephens, M.A. (1989). A goodness-of-fit test using Moran's statistic with estimated parameters. *Biometrika*, **76**, 385-392.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

Deville, J.C. and Sarndal, C.E. (1992) Calibration estimator in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Fisher, R.A. (1951) *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Feller W. (1968) *An Introduction to Probability Theory and Its Applications.* Volume 1, 3rd Edition. Wiley: New York.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003) *Bayesian Data Analysis*. CRC Press.

Greenlees, J.S., Reece W.S. and Zieschang, K.D. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, **77**, 251-261.

Hajek, J. (1971). Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One". *The Foundations of Survey Sampling*, Godambe, V.P. and Sprott, D.A., eds., 236, Holt, Rinehart and Winston.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, **46**: 931-959.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**(1), 153-161.

Heckman, J. and Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics*, **86**(1), 30-57.

Heckman, J. and Vytlacil, E. (2006) Econometric evaluation of social programs. *Handbook of Econometrics*, **6**, J. Heckman and E. Leamer, eds., Amsterdam: North Holland.

Holland, P.W. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945-960.

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.

Hoxby, C.M. (2000) Does competition among public schools benefit students and taxpayers? *The American Economic Review* **90** (5), 1209-1238.

Imbens, G. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, **86**, 4–30.

Imbens, G.W. and Angrist, J.D. (1994) Identification and estimation of local average treatment effects. *Econometrica,* **62**, 467-475.

Iverson, H.K. and Randles, R.H. (1989) The effects on convergence of substituting parameter estimates into U-statistics and other families of statistics. *Probability and Related Fields*, **81**, 453-471.

Lee, J. And Berger, J.O. (2001) Semiparametric Bayesian analysis of selection models. *Journal of the American Statistical Association*, **96**, 1397-1409.

Little, R. J. (2004) To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, **99**, 546-556.

Moran, P. (1951) The random division of an interval – Part II. *JRSS*, *Series B*, **13**, 147-150.

Neyman, J. (1990, [1923]) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, **5**, 465-472.

PISA 2000. *Technical Report.* (2002) Edited by Ray Adams and M.Wu. OECD. Paris.

Pfeffermann, D., Krieger, A.M. and Rinnot, Y. (1998) Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, **8**, 1087-1114.

Pfeffermann, D., Moura, F. A. S. and Nascimento-Silva, P. L. (2006) Multilevel modeling under informative sampling. *Biometrika*, **93**, 943-959.

Pfeffermann, D. and Sverchkov, M. (1999) Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series* B, 166-186.

Pfeffermann, D. and Sverchkov, M. (2003) Fitting generalized linear models under informative sampling. *Analysis of Survey Data*, Chapter 11, C. Skinner and R. Chambers, eds., New York: Wiley.

Qin, J. and Zhang, B. (2007) Empirical-likelihood-based inference in missing response problems and its application in observational studies. Journal Royal Statistical Society B, **69** (1), 101-122.

R Development Core Team (2004) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Robins, J.M., Rotnizky, A. and Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-886.

Rosenbaum, P. R. (1984) Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, **79**, 565-574.

Rosenbaum, P.R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association*, **82**, 387-394.

Rosenbaum, P.R. (2002) *Observational Studies*. Springer-Verlag. 2nd edition.

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for treatment effects. *Biometrika*, **70**, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984) Reducing bias in observational studies using the subclassification on the propensity score. *Journal of the American Statistical Association*, **79**, 516-524.

Rothman K. J (2002) *Epidemiology. An introduction*. Oxford University Press.

Rubin, D.B (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701.

Rubin, D.B. (1977) Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**, 1-26.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York.

S?rndal, C.E. (1980) On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, **67** (3), 639-650.

Smith, T.M.F. and Sugden, R.A. (1988) Sampling and assignment mechanisms in experiments, surveys and observational studies. *International Statistical Review*, **56**, 165-180.

StataCorp. 2004. *Stata Statistical Software: Release 7.* College Station, TX: StataCorp LP.

Stefanski, L.A. and Boos, D.D. (2002) The calculus of M-estimation. *The American Statistician*, **56**, 29-38.

Sverchkov, M. and Pfeffermann, D. (2004) Prediction of finite population totals based on the sample distribution. *Survey Methodology*, **30**, 79-92.

Vandenberghe, V. and Robin, S. (2004) Evaluating the effectiveness of private education across countries: a comparison of methods. *Labour Economics*, **11** (4), 487-506.

Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA.