Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2007/09

Finding alternative sources of identification in generalized selection models using principal stratification

Fabrizia Mealli, Barbara Pacini

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Finding alternative sources of identification in generalized selection models using principal stratification[1]

Fabrizia Mealli, mealli@ds.unifi.it
Dipartimento di Statistica, Università di Firenze

Barbara Pacini, barbara.pacini@unibo.it
Dipartimento di Scienze Statistiche, Università di Bologna

Abstract

In this paper we consider two approaches for dealing with "endogenous selection" problems when estimating causal effects, namely selection models (SM) and principal stratification (PS). Our main goal is to highlight similarities and differences of the two approaches, by first investigating the different nature of their parametric hypotheses. In order to support our reasoning, we show their different performances by simulations under both approaches. We argue that principal stratification is able to suggest alternative identification strategies not always easily translated into assumptions of a selection model.

1 Introduction

When the goal of inference is that of estimating causal effects, we usually have to face problems related to how data are observed. In observational studies the most relevant of such problems is the fact that assignment to treatment is not under the control of the investigator; in addition some studies, both observational and experimental, may be affected by different sorts of post-treatment selection of observations due to, e.g., non response, truncation or censoring "due to death". Moreover, one may be interested in decomposing the total effect of a treatment on an outcome into a direct effect and an indirect one mediated by another intermediate variable. All such complications require to somehow control for them, but the use of the standard statistical conditioning is improper (Rubin, 1978; Heckman, 1974; Rosembaum, 1984; Rubin, 2004).

A relatively recent approach to deal with post-treatment complications is Principal Stratification (PS), as first defined by Frangakis and Rubin (2002) within the framework of the Rubin Causal model (Rubin, 1974; Holland, 1986) and applied mainly in experimental studies (Barnard et. al., 2003; Zhang et. al., 2006; Mattei and Mealli, 2007). In Frangakis and Rubin (2002), PS was introduced in order to give a formal definition of surrogate endpoints; it was then used to define direct and indirect effects (Mealli and

[1]This paper was presented at the Causal Inference conference held in Uppsala, June 2007. The authors are grateful for comments by the conference participants. Financial support for this research was provided by Miur Cofin 2005 grant.

Rubin, 2003). As Rubin (2004) points out, the PS framework can be viewed as having its seeds in the Instrumental Variables (IV) method of estimation of causal effects. This is described within the context of the potential outcomes in Angrist, Imbens and Rubin (1996) from the frequentist perspective, and in Imbens and Rubin (1997) from the Bayesian perspective, although it has roots in work by economists such as Tinbergen (1930) and Haavelmo (1944). Indeed, the approach to adjust for noncompliance applied in Angrist, Imbens and Rubin (1996) and in Imbens and Rubin (1997) is a special application of the PS framework, where the compliers are a principal stratum with respect to the post-treatment compliance behavior.

Despite the use of PS to solve particular issues, the framework appears to be a very general one, that can be applied in various contexts. Furthermore, the framework may lead to both parametric and semi(non)parametric inference, depending on the set of assumptions that can be reasonably maintained, as well as whether point or partial identification is to be achieved.

In the econometric literature post-treatment complications are usually described as problems of endogenous selection and include treatment assignment in observational studies, self-selection, non response, censoring or truncation "due to death". They are usually represented by means of selection models (SM, Heckman, 1974; Gronau, 1974). Since the seminal work of Heckman, various extentions of the model have been proposed, which include semi and nonparametric versions (Pagan and Ullah, 1997; Vella, 1998). While parametric selection models can be rather restrictive in terms of distributional assumptions, semi and nonparametric extensions usually require additional exclusion restrictions to maintain point-identification of the parameters of interest.

In the paper we consider both SM and PS to estimate causal effects in the presence of post-treatment complications under different sets of assumptions. Indeed, causal inference requires some assumptions about the population, the sampling process and the behavior of the subjects under study. The *credibility* of (causal) inference decreases with the strength of the assumptions maintained (Manski, 2003). Note that hypotheses are not all on the same ground, and they may have different nature, as well as a different degree of agreement.

The main goal of the paper is to highlight similarities and differences of the two approaches (SM and PS), first investigating the different nature of their parametric hypotheses. We also argue that Principal Stratification is able to suggest alternative identification strategies not always easily translated into assumptions of a selection model. In order to support our reasoning, we show, by simulations under both approaches, their different performances.

The paper is organized as follows: in section 2 principal stratification is presented together with its main characteristics; section 3 is devoted to recall the basic features of selection models; a first comparison in terms of aims and developments of the two approaches is conducted in section 4. In section

5 and 6 some simulation results are shown, which refer to an empirical setting with nonignorable post-treatment nonresponse. Simulation are aimed at studying the different performances of the two approaches together with their robustness. Specific identification strategies are presented in section 7 in a rather complicated setting, with more then one post-treatment complications. Section 8 summarizes our findings and concludes.

2 Principal Stratification and its Role for Causal Inference in Experimental and Observational Studies

Principal stratification has been first introduced by Frangakis and Rubin (2002), in order to address post-treatment complications in an experimental setting. Here, we show that the framework can be easily extended to an observational setting under specific hypotheses on the assignment mechanism. We first introduce "potential outcomes" (see Rubin, 1979) for one post-treatment variable and a binary treatment. If unit i in the study ($i = 1, \ldots, N$) is to be assigned to treatment t ($t = 1$ for treatment and $t = 0$ for no treatment), we denote with $Q_i(t)$ a post-treatment potential variable, which is, without loss of generality, assumed to be an indicator equal to 1 if a specific post-treatment event happens and 0 otherwise. For example, Q may represent a response indicator for a specific item in a questionnaire, or a survival indicator: in these examples, $Q_i = 0$ precludes the observation of $Y_i(t)$, which is the potential outcome variable of main interest, i.e., with respect to which we are interested in investigating the causal effect of T, defined on a single unit as a comparison between $Y_i(1)$ and $Y_i(0)$. Alternatively, we may be interested to decompose the effect of T on Y in a direct effect and an indirect one mediated by a variable Q (in this case Y would be observed for every i-th unit).

In an observational setting, various hypotheses can be posed on the assignment mechanism. In what follows, we will assume that treatment assignment is unconfounded given a vector X of pretreatment variables:

$$\text{Assumption } 1: \quad T \perp Q(0), Q(1), Y(0), Y(1)|X$$

In other words, we assume that within cells defined by the values of pretreatment variables X, the treatment is randomly assigned or, at least, is assigned independently of the post-treatment variables considered relevant for the study. If we indicate with $Tobs_i$ the observed treatment assignment, the observed data are

$$\left(Tobs_i, Q(Tobs_i), Y(Tobs_i), X_i\right) \qquad i = 1, \ldots, N,$$

Consider now the potential outcomes $Q_i(0)$ and $Q_i(1)$. Within each cell defined by specific values of the pretreatment variables, the units under study

can be stratified into four groups, named Principal Strata, according to the joint value of the potential variables $(Q_i(0), Q_i(1))$; the strata are the following:

$$11 = \{i : Q_i(1) = Q_i(0) = 1\} \quad \text{with proportion } \pi_{11|X_i}$$

$$10 = \{i : Q_i(1) = 1, Q_i(0) = 0\} \quad \text{with proportion } \pi_{10|X_i}$$

$$01 = \{i : Q_i(1) = 0, Q_i(0) = 1\} \quad \text{with proportion } \pi_{01|X_i}$$

$$00 = \{i : Q_i(1) = Q_i(0) = 0\} \quad \text{with proportion } \pi_{00|X_i} = 1-\pi_{11|X_i}-\pi_{10|X_i}-\pi_{01|X_i}$$

Let G_i represent the principal stratum indicator for subject i. The principal stratum indicator G_i is not affected by treatment assignment $Tobs_i$, so it only reflects characteristics of subject i, and can be regarded as a covariate, which is only partially observed in the sample (Angrist et al., 1996); by unconfoundedness, however, it is guaranteed to have the same distribution in both treatment arms, within cells defined by pre-treatment variables. We usually need to adjust for the principal strata, which synthesize important unobservable characteristics of the subjects in the study.

Usually, information on causal effects is contained in a particular principal stratum: in the present exemplified context direct information on the causal effect can be found in the 11 stratum (e.g., respondents or survivors under treatment and control or units with the same levels of the intermediate variable under treatment and control), because only for units belonging to this stratum one can consistently compare $Y(1)$ and $Y(0)$. The effect to be estimated is usually an average effect within stratum: $E(Y(1) - Y(0)|11)$.

If Q represents non response or death, in fact, in all strata but 11 only one potential outcome or none can be observed, whereas if Q is an intermediate variable evidence on the direct effect cannot be disentangled from the indirect effect trough Q, except for stratum 11 for which the effect on Q is absent (Q is always equal to one).

The purpose of inference is to estimate (by likelihood or Bayesian procedures[2]) the probabilities of strata belonging ($\pi_{11|X_i}$, $\pi_{10|X_i}$,$\pi_{01|X_i}$) and the distribution of the potential outcomes within each stratum, under different identifying distributional and behavioral assumptions.

Note that the same framework can be easily extended to cases with non binary treatment, non binary post-treatment variable and more than one post-treatment variables (an example of such extensions is presented in section 7).

[2]In what follows, we will focus on likelihood based inference, in order to better highlight identification problem and also to conform to the standard literature on selection models.

3 Solving endogenous selection problems by means of sample selection models

Sample selection models have been introduced in order to consistently estimate regression models' parameters with non random samples (Heckman, 1974; Gronau, 1974); they have been also widely used for the estimation of treatment effects with endogenous selection. Assuming, as in the previous section, that T in unconfounded (Assumption 1) given a vector of X variables, but the outcome variable is observed only for a nonrandom sample; then the standard specification of a selection model is the following:

$$Y = \alpha + \beta X + \gamma T + u_1 \tag{1}$$

$$Q^* = \omega + aX + bT + u_2$$

where $u_1 \sim N(0, \sigma_1)$, $u_2 \sim N(0, \sigma_2)$ and $corr(u_1, u_2) = \rho$. Y is observed only for $Q^* > 0$, so that Q^* is the latent variable underlying the observed selection indicator Q. As far as causal inference is concerned, this specification implicitly assumes that treatment effect is constant and measured by γ; moreover monotonicity with respect to Q holds by construction: $b > 0$ implies $Q|X, T = 1 \geq Q|X, T = 0$ (i.e., response under treatment cannot be less than response under control), the opposite inequality would hold for $b < 0$. More complex specifications are possible, allowing for heterogeneous effects and non monotonicity: in this case, separate models for $T = 0$ and $T = 1$ are usually formulated, as we will see in section 6.

Under the standard hypotheses given above, model (1) can be consistently estimated by maximum likelihood, although caution is required when no exclusion restriction on the Xs is imposed. The model can also be estimated by a two-step procedure (Heckman, 1974); note however that in this case identification is achieved only trough a nonlinear transformation of the Xs in the Mills ratio, so that at least a continuous regressor is needed (Olsen, 1980).

Despite identification, Little (1985) and Copas and Li (1997) observe that inference is not robust and very sensitive to parametric assumptions. Relaxing normality and further extensions of the basic model will be reviewed in the next section, according to a chronological order as they appear in the literature since the seminal paper by Heckman.

4 Basic aims and developments of the two approaches

Sofar, the essential features of the SM and PS approaches have been presented, which provide a first understanding of their different "spirit". In fact, selection models aim at estimating parameters of a model (first equation of

model (1)), model that should be valid for the whole population in the absence of complications (i.e., if the data come from a random sampling). Because observations come from a nonrandom sampling procedure, it becomes necessary to include a selection equation as in model (1), in order to "correct" the estimation of the causal effect. The intrinsic nature of this approach is parametric, although semi and nonparametric versions have been developed.

On the contrary, the PS approach focuses on information contained in specific subgroups of units, aiming at producing valid inference conditional on such subgroups, without a priori extending results to the whole population. The intrinsic nature of this approach in nonparametric, although the framework may lead to both parametric and semi(non)parametric inference, depending on the set of assumptions that can be reasonably maintained, as well as whether point or partial identification is to be achieved.

Principal stratification	**Sample selection models**
Basic characteristics	
Data on endogenously selected sample	Data on endogenously selected sample
Non parametric hypotheses	Parametric hypotheses
Finding meaningful subgroups to refine inference	Extending results to the whole population
Causal inference on subgroups	Causal inference: Average treatment effect
Possible extensions	
Reducing the number of PS by assumptions	Removing parametric hypotheses
Monotonicity (Angrist et a., 1996; Mealli et al., 2004) (Frangakis et al., 2004)	Non normality assumption on error terms (Lee, 1982,1983; Gallant and Nychka, 1987; Pagan and Vella, 1989; Honore *et al.*, 1997)
Stochastic dominance (Rubin & Zhang, 2002) (Zhang et al., 2006)	Nonparametric correction term: a) maintaining single index restriction (Newey, 1990; Cosslet, 1991; Lee, 1994)
Distributional hypotheses (Zhang et al., 2005)	b) avoiding single index restriction
Restrictions on covariate effects (Jo, 2003; Mattei & Mealli, 2006)	(Choi, 1990; Ahn & Powell, 1993; Ichimura, 1993; Li and Wooldridge, 2002)

Table 1: Basic characteristics and possible extensions of PS and SM

The basic assumptions embedded in a standard selection model have already been summarized and refer to the parametric distributional hypothesis for the error terms and their additivity with respect to the explanatory variables, constant treatment effect and monotonicity. Possible departures from a fully parametrized model concern primarily the relaxation of distributional and functional assumptions. While preserving the parametric nature of the model, the normality assumption can be avoided by specifying arbitrary para-

metric marginal distributions for the error terms, but still coming back to joint normality by appropriate transformation of the marginals (Lee, 1982, 1983; Gallant and Nychka, 1987). This strategy can be followed within a maximum likelihood estimation framework.

Alternatively, within a two-step estimation setting, only the distribution of u_2 must be specified, together with the linearity assumption of the relationship between u_1 and u_2 (Olsen, 1980; Wooldridge, 1994). Also, because joint normality implies this linear relationship, normality can be relaxed by including terms capturing deviations from linearity (Lee, 1984; Pagan and Vella, 1989). Honore *et al.* (1997) have proposed an alternative approach assuming that errors are simmetrically distributed conditional on the regressors.

Avoiding parametric assumptions on the error distribution, while maintaining additivity and imposing an exclusion restriction on X, the conditional expectation $E(u_1|X, T, Q = 1)$, which is relevant for the correction of the outcome equation in model (1):

$$E[Y|X, T, Q = 1] = \alpha + \beta X + \gamma T + E(u_1|X, T, Q = 1) \tag{2}$$

can be expressed in the following general way:

$$E(u_1|X, T, Q = 1) = g(v(X, T; \theta)). \tag{3}$$

Different identification and estimation strategies have been proposed in the literature depending on the specification of v and g. When v is supposed linear w.r.t X and T (single index restriction) θ can be estimated without distributional assumptions on u_2, and in the second step, g can be nonparametrically estimated using flexible functional forms (such as series expansion or step functions). This estimation strategy is proposed for example in Cosslet (1991). Similarly, Robinson (1988), Newey (1990) and Lee(1994) suggest, for the second step, a different nonparametric estimation procedure based on differencing out the selectivity bias, without a direct estimation of $g((X, T)'\theta)$.

Moreover, avoiding linearity in v w.r.t X and T, i.e. removing the single index restriction, Ahn and Powell (1993) propose a non parametric kernel method to estimate the *propensity score* $Pr(Q = 1|X, T)$; on the same ground, Choi (1990) suggests the use of series expansions to approximate the unknown function v. Li and Wooldridge (2002), Lee (1994) and Ichimura (1993) suggest alternative multistep estimation strategies. Note that additivity of error terms is always maintained. All these approaches in estimating model (2) can potentially be framed into the GAM and GAPLM (Generalized Additive Partial Linear Models) literature (Hastie and Tibshirani, 1990; Green and Silverman, 1994; Haerdle *et al.*, 2006).

It must be noted that all the semiparametric approaches involving the estimation of (3) need identifying conditions that always require at least an exclusion restriction; in addition the intercept term is not identifiable because it cannot be disentangled from $g(v(X, T; \theta))$. This is a relevant problem

when, in order to remove monotonicity and constant treatment effect, separate model for $T = 0$ and $T = 1$ are specified and thus the treatment effects also involves the comparison of intercept terms (cf. section 6). Heckman (1990) and Schafgans and Zinde-Walsh (2002) suggest possible solutions to retrieve the intercept term based on an identification "at infinity" condition. For a recent empirical comparison of parametric and semiparametric selection models see Christofides *et al.* (2003).

As shown above, recent developments of selection models are aimed at removing specific hypotheses on error terms: the definition of error terms requires a more or less specified model both for the outcome and the selection equation(s).

This is instead avoided in the PS approach, where the latent strata are generated by the primitive potential outcomes. Usually, identification strategies exploit the comparison between observed groups and latent groups (strata). This comparison can sometimes imply only bounds for treatment effects (Zhang and Rubin, 2003; Imbens and Rubin, 1997); point identification can instead be reached by means of assumptions that usually relate to specific behavioral hypotheses about the strata. Some of such assumptions aim at reducing the number of strata: following our introductory example, imposing monotonicity (response under treatment cannot be less than response under control) allows, for example, to exclude either stratum 10 or stratum 01. Other hypotheses impose certain features of the distribution of outcomes within or among strata: these include various forms of exclusion restrictions (Mealli *et al.*, 2004), various versions of stochastic dominance that assume, for example, that the distribution of the outcome in one or more strata stochastically dominate that of other strata (Zhang and Rubin, 2003; Zhang *et al.*, 2006), and various forms of ignorability and non ignorability model for the selection (Frangakis and Rubin, 1999).

In case point identification cannot be achieved solely on the base of such nonparametric (structural) hypotheses, or simply for efficiency reasons, some distributional hypotheses can be introduced. Note however, as will be also clearer with our simulation exercises, that such distributional assumptions do not usually involve the joint specification of a model for the outcome and the selection process, but rather refer to distributions for the outcome variables conditional on the strata. These distributions have explicit implications on the probability law of variables within observed groups, in terms of mixture distributions, so that the theory on mixture models can be exploited for both identification and specification testing.

Analogously, identification and efficiency improvements can be achieved by exploiting covariates: plausible behavioral hypotheses *within or among groups* defined by the values of the covariates can be translated into restrictions on coefficients *within or among strata*. One can, for example, exclude some interaction terms (Jo, 2002), or impose the same coefficients across strata for some covariates (Frangakis, 2006). Some practical examples of this approach will be presented in the following sections.

Using PS, whatever the assumptions made, the result of inference is always a causal effect within one or more strata. An issue that often arises regarding the PS approach is that we cannot univocally identify the group the causal effect refers to, so we cannot univocally estimate the individual causal effects. This issue also characterizes the Instrumental Variable literature where, under certain assumptions, only the effect on specific subpopulations can be identified (Angrist et al., 1996). Note, however, that the fact that proper causal effects can only be defined and estimated for latent subgroups of units is a limitation created by the selection mechanism, rather than a drawback of the framework of principal stratification.

4.1 Equivalence under particular conditions

Although the two approaches have been developed separately, Vytlacil (2002) has shown an equivalence in the specific linear IV setting. He proves that assumptions on the principal strata directly translate into assumptions on a selection latent index model and vice-versa: under the LATE (Latent Average Treatment Effect, Imbens and Angrist, 1994) independence assumption and the LATE monotonicity assumption the two models are shown to be observationally equivalent. Note, however, that interpretation of inference is usually not the same, in the sense that the estimated causal effect explicitly refers to a specific subgroup of individuals within the PS framework, whereas it refers to the whole population within the selection model setting. Furthermore, under alternative hypothesis on the strata, the equivalence cannot be stated, so that the results are not general and not easily generalizable.

In what follows we first focus on a hypothetical simple empirical setting in order to show, by simulations, differences in models assumptions and differences in performance of the estimator under the two approaches. More general complications will be introduced and analyzed in subsequent sections.

5 Empirical setting and simulation exercise

We first consider the following simplified setting in the field of evaluation of financial aids to firms. Let T be a binary treatment which represents public financial assistance to firms ($T = 1$ for treatment and $T = 0$ for no treatment). We assume that T is unconfounded given a vector of pre-treatment covariate X. We assume a single continuous covariate X which can be considered a univariate summary of the pre-treatment variables (e.g., propensity score); the intermediate post-treatment variable Q represents the response to a post-treatment questionnaire on firms' performances, the outcome variable of interest being the turnover Y. So we are facing the post-treatment complication of a nonignorable missing mechanism of the outcome variable. This artificial setting is consistent with evidence from the real world, where typ-

ically missingness on turnover variables can rarely be assumed missing at random[3].

Note that, within this context and in the absence of exclusion restrictions, it would not be possible to nonparametrically point identify treatment effects and most of the semiparametric versions of the selection models would require instruments. For these reasons, we decide to stay within a parametric setting, exploring the different nature of the parametric assumptions of the two approaches.

Given these premises, we first simulate under a parametric sample selection model and estimate the causal effect using both approaches; we then simulate under the principal strata model, again estimating the causal effect using both methods.

5.1 Simulating under SM

We assume the following data generating processes[4]:

$$\log(Y) = \alpha + \beta X + \gamma T + u_1 \tag{4}$$

$$Q^* = \omega + aX + bT + u_2$$

where $u_1 \sim N(0, \sigma_1 = 1)$, $u_2 \sim N(0, \sigma_2 = 1)$ and $corr(u_1, u_2) = \rho = \{-0.2, -0.5\}$. Y is observed only for $Q^* > 0$, and $\alpha = 12.5$, $\beta = 1.5$, $\gamma = 1$,$\omega = -0.15$, $a = 0.82$ and $b = 0.76$. T is assumed to follow a logit model:

$$Pr(T = 1|X) = \frac{exp(c + dX)}{1 + exp(c + dX)}$$

where $c = 0.5$ and $d = 1.3$.

We first draw a random sample of size $n = 1000$ of X from a standard normal distribution; we then simulate 1000 samples of size n for Y, T, and Q^*.

Concerning the estimand of interest in the main equation, the causal effect in log-scale is measured by the parameter γ and is thus assumed constant; moreover, in the selection equation, monotonicity of T with respect to Q is implied by the model specification, i.e., $Q^*(T = 1) \geq Q^*(T = 0)$. Trying to bridge the specified selection model with the underlying latent strata and using notation introduced in section 2, we can derive the following correspondence:

[3]For example, Mattei and Mauro (2007), from a survey of Tuscan artisan enterprises, found evidence of nonignorability of nonresponse.

[4]Only some simulation results are reported in the paper, which are representative of the main conclusions drawn from the MonteCarlo study using different scenarios for the parameters' values. All scenarios have been suggested by empirical evidence on real case studies concerning Italian firms (where Y is expressed in euros). As far as the correlation parameter is concerned, a grid of negative values have been considered, consistenly with the evidence of negative correlation between non response propensity and level of turnover.

- $\{i : u_{2i} > -\omega - x_i a\} \equiv \{i : Q_i(1) = Q_i(0) = 1\}$, now denoted as *RR* stratum;
- $\{i : -\omega - b - x_i a < u_{2i} < -\omega - x_i a\} \equiv \{i : Q_i(1) = 1, Q_i(0) = 0\}$, denoted as *RN* stratum;
- $\{i : u_{2i} < -\omega - b - x_i a\} \equiv \{i : Q_i(1) = Q_i(0) = 0\}$, denoted as *NN* stratum.

As previously noted, monotonicity is implicitly assumed and implies that the *NR* stratum ($\{i : Q_i(1) = 0, Q_i(0) = 1\}$) is empty.

Model 4 is estimated by full maximum likelihood, under the stated parametric assumptions.

We now show how the same empirical setting can be formalized within the PS framework, define the causal effect within this framework, and, using the same simulated datasets, estimate the causal effect by maximizing the likelihood implied by the Principal Stratification.

There are potentially three latent strata within each cell defined by pre-treatment covariate X:

$$RR = \{i : Q_i(1) = Q_i(0) = 1\} \quad \text{with proportion } \pi_{RR}$$

$$RN = \{i : Q_i(1) = 1, Q_i(0) = 0\} \quad \text{with proportion } \pi_{RN}$$

$$NN = \{i : Q_i(1) = Q_i(0) = 0\} \quad \text{with proportion } \pi_{NN}$$

The causal effect of interest is the effect within the *RR* stratum, because only for individuals belonging to the stratum we have observations on both $Y(1)$ and $Y(0)$. In order to form the likelihood, it is necessary to state the correspondence between observed groups defined by T and Q and latent strata, as shown in the following table:

Observed subgroups $O(T, Q)$	Turnover Y	Latent strata
$O(1,1) = \{i : T_i = 1, Q_i = 1\}$	OBS	RR or RN
$O(1,0) = \{i : T_i = 1, Q_i = 0\}$	·	NN
$O(0,1) = \{i : T_i = 0, Q_i = 1\}$	OBS	RR
$O(0,0) = \{i : T_i = 0, Q_i = 0\}$	·	RN or NN

Note that two of the four observed groups results from a mixture of two principal strata; however, unlike standard mixture models, some units have zero probability of belonging to some strata and this may facilitate disentangling mixture.

We specify the likelihood function, considering the contribution of each observed subgroup separately, making parametric assumptions that are consistent with the DGP. Specifically, we assume lognormality of Y conditional on the principal strata:

$$
\begin{aligned}
f(log(Y_{T=1})|RR, X) &\sim N(\alpha_{T=1,RR} + \beta_{T=1,RR}X, \sigma_{T=1,RR}) \\
f(log(Y_{T=0})|RR, X) &\sim N(\alpha_{T=0,RR} + \beta_{T=0,RR}X, \sigma_{T=0,RR}) \\
f(log(Y_{T=1})|RN, X) &\sim N(\alpha_{T=1,RN} + \beta_{T=1,RN}X, \sigma_{T=1,RN})
\end{aligned}
\tag{5}
$$

Because the DGP implies a constant treatment effect, we impose the following restrictions: $\beta_{T=1,RR} = \beta_{T=0,RR} = \beta_{T=1,RN} = \beta$. We further assume that $\sigma_{T=1,RR} = \sigma_{T=0,RR} = \sigma_{T=1,RN} = \sigma$. The distribution of the principal strata is modelled as a multinomial logit, where:

$$
\begin{aligned}
\pi_{RR|X} &= \frac{\exp(\delta_{RR} + \gamma_{RR}X)}{1 + \exp(\delta_{RR} + \gamma_{RR}X) + \exp(\delta_{RN} + \gamma_{RN}X)} \\
\pi_{RN|X} &= \frac{\exp(\delta_{RN} + \gamma_{RN}X)}{1 + \exp(\delta_{RR} + \gamma_{RR}X) + \exp(\delta_{RN} + \gamma_{RN}X)} \\
\pi_{NN|X} &= 1 - \pi_{RR|X} - \pi_{RN|X}
\end{aligned}
\tag{6}
$$

Denoting $\theta = \{\alpha_{T=1,RR}, \alpha_{T=0,RR}, \alpha_{T=1,RN}, \beta, \sigma, \delta_{RR}, \gamma_{RR}, \delta_{RN}, \gamma_{RN}\}$, we can now write the likelihood function as follows:

$$
L(\theta|X, T, Q, log(Y)) \propto \prod_{i \in O(1,1)} [\pi_{RR_i} N_i(\alpha_{T=1,RR} + \beta X_i, \sigma) + \pi_{RN_i} N_i(\alpha_{T=1,RN} + \beta X_i, \sigma)]
$$

$$
\times \prod_{i \in O(1,0)} \pi_{NN_i} \times \prod_{i \in O(0,1)} [\pi_{RR_i} N_i(\alpha_{T=0,RR} + \beta X_i, \sigma)] \times \prod_{i \in O(0,0)} [\pi_{RN_i} + \pi_{NN_i}]
$$

Note that, while identifiability in the selection model is driven by the joint normality assumption for the distribution of error terms, within the PS likelihood identification is achieved thank to results of finite mixture distribution theory (see e.g., McLachlan and Peel, 2000). Indeed, likelihood results in a finite mixture distribution and identification is straightforward, except when $\pi_{RR} = \pi_{RN}$.

In Table 2 we report the results of our MonteCarlo analysis, where $Diff = \alpha_{T=1,RR} - \alpha_{T=0,RR}$ is the difference between the intercepts, and is thus an estimate of the causal effect within the RR stratum, to be compared with the estimate of γ from the selection model approach.

It is evident from the results that, despite the fact that the PS framework does not reflect exactly the DGP, the estimation of the causal effect through PS performs as well as the one for γ (see also Fig. 5.1 and Fig.5.1), especially for relatively small values of ρ. In fact, for small correlation values data do not carry a lot of information on the parameters (the likelihood of selection models is rather flat), so the performances of the two approaches are more or less equivalent.

$\rho = -0.2$				$\rho = -0.5$			
Parameter	Mean	Q1	Q3	Parameter	Mean	Q1	Q3
γ	1.03	0.89	1.14	γ	1.03	0.91	1.11
ρ	-0.13	-0.37	0.07	ρ	-0.44	-0.62	0.34
$\alpha_{T=1,RR}$	13.33	13.17	13.48	$\alpha_{T=1,RR}$	13.12	12.98	13.24
$\alpha_{T=0,RR}$	12.31	12.26	12.37	$\alpha_{T=0,RR}$	12.03	11.98	12.09
$Diff$	1.02	0.85	1.18	$Diff$	1.09	0.94	1.24

Table 2: Estimation results under selection model as data generating process: MonteCarlo average, 1^{th} and 3^{rd} quartile over 1000 replications

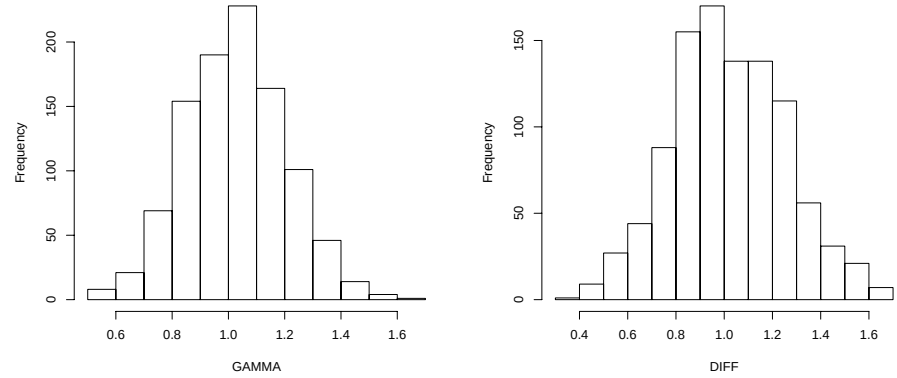

Figure 1: Histogram for γ and $Diff$ MonteCarlo distributions - 1000 replications under selection model, $\rho = -0.2$

5.2 Simulating under PS

Previous section shows that the performances of the two approaches is similar when the data are generated under the SM assumptions. We now investigate their performances when data are simulated under the PS approach. We still assume a logit model for treatment T, with $c = 0.5$ and $d = 1.3$. We assume a multinomial logit for the distribution of principal strata, as in (6), where

$$\delta_{RR} = 0.5, \quad \gamma_{RR} = 2$$

$$\delta_{RN} = 0.1, \quad \gamma_{RN} = 1$$

The multinomial logit parameters were chosen in such a way, so that the proportions of the strata are similar to those implied by the previously simulated selection model. Furthermore, we specify a lognormal model for turnover,

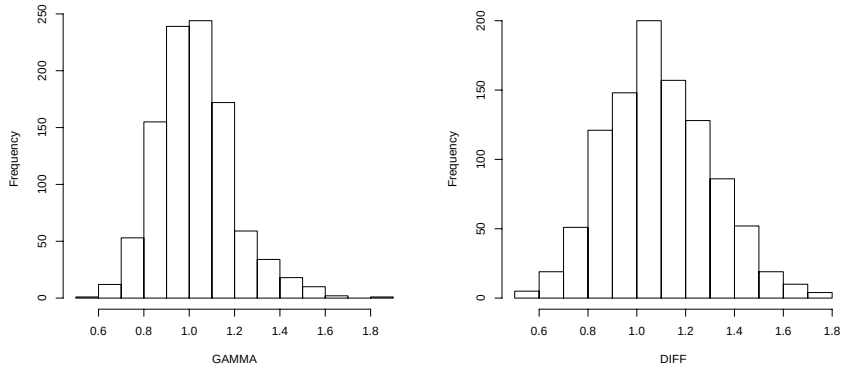

Figure 2: Histogram for γ and $Diff$ MonteCarlo distributions - 1000 replications under selection model, $\rho = -0.5$

as in (5), with the following parameters:

$$\alpha_{T=1,RR} = 12, \quad \beta_{T=1,RR} = 2, \quad sd_{T=1,RR} = 1$$

$$\alpha_{T=0,RR} = 11, \quad \beta_{T=0,RR} = 2, \quad sd_{T=0,RR} = 1$$

$$\alpha_{T=1,RN} = 13, \quad \beta_{T=1,RN} = 2, \quad sd_{T=1,RN} = 1$$

Results, reported in Table 3, show the bad performance of the maximum likelihood estimator of the selection model: the MonteCarlo average estimate of γ has a large positive bias, highlighting that the bivariate normality is not appropriate to model the selection process. This is confirmed by the estimate of ρ, which is not significantly different from zero in 946 samples over 1000.

Parameter	Mean	Q1	Q3
γ	1.33	1.20	1.45
ρ^*	-0.11	-0.33	0.09
$\alpha_{T=1,RR}$	12.03	11.93	12.10
$\alpha_{T=0,RR}$	11.00	10.94	11.06
$Diff$	1.03	0.91	1.13

*946 cases with ρ not significantly different from zero

Table 3: Estimation results under principal stratification as data generating process: MonteCarlo average, 1^{th} and 3^{rd} quartile over 1000 replications

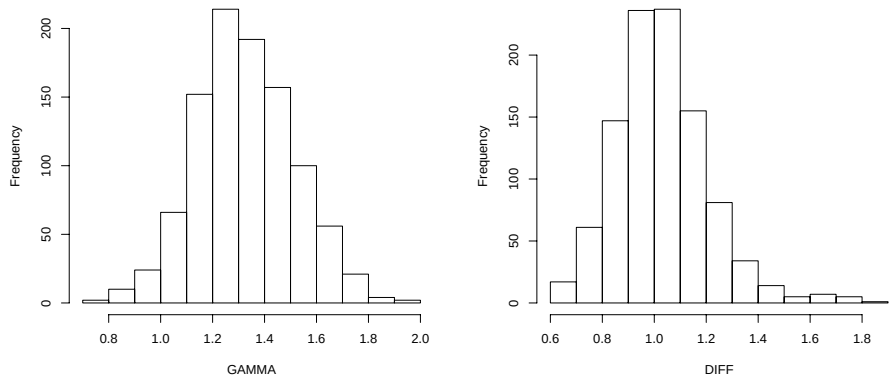

Figure 3: Histogram for γ and $Diff$ MonteCarlo distributions - 1000 replications under principal stratification

5.2.1 Testing parametric assumptions

Within the PS approach, we specify the distribution of the outcome variable conditional on the principal strata, we are also specifying the distribution conditional on observed groups, which usually results as a mixture of parametric distributions. We may question whether such parametric distribution describes the data in a sufficiently satisfying manner. That is, given a specific sample, if we can justify the use of a parametric density function. In order to test whether a parametric density function describes the data accurately enough in a statistical sense, we can make use of nonparametric confidence bands derived from a nonparametric density estimation procedure. For example, to see whether the sample provides some justification for the use of a mixture of log-normal densities (as in the simulated example) for the outcome within the observed group $\{i : T_i = 1, Q_i = R\}$, we can compute the 95bands around a nonparametric density estimate[5]. We then test the parametric mixture hypothesis at a 5% significance level, checking if the parametrically estimated mixture is globally within the confidence bands. If the parametric density has some points out of the bands, the data "rejects" the log-normal mixture as the "true" distribution. In our case the null hypothesis is accepted, consistently with the way data were simulated.

Given our observational setting, all the distribution we specify are conditional on the covariates, so that the parametric distribution that we must compare with the nonparametric bands is the average of the following condi-

[5]Here we use a kernel density estimation with a Gaussian kernel and optimal bandwidth selected using the Silverman rule (Silverman, 1986) and global confidence bands

tional density:

$$\hat{\pi}_{RR|x} \cdot \hat{f}(Y_1^{RR}|x) + \hat{\pi}_{RN|x} \cdot \hat{f}(Y_1^{RN}|x)$$

over the distribution of X. This is consistent with the nonparametric estimation of the density of Y based on the observed data (see figure 4). Besides the

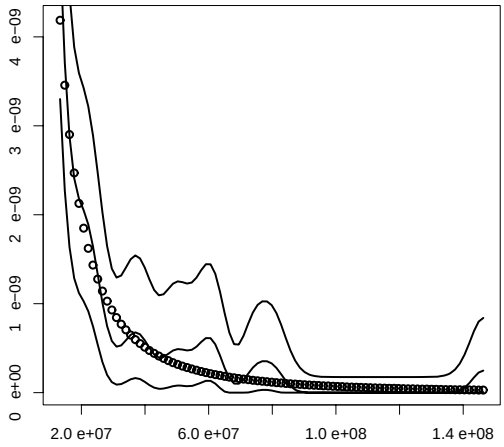

Figure 4: Estimated mixture for group $\{i : T_i = 1, Q_i = R\}$ and nonparametric density estimation with 95% level confidence bands

use of nonparametric confidence bands as a formal diagnostic tool, nonparametric density estimation can also be used as a graphical tool to guide the parametric specification of the mixture. This is particularly true in an experimental setting, where conditioning on the covariates is not required, so that the shape of the nonparametric density can directly suggest the distributional assumptions.

5.2.2 Checking robustness with respect to parametric assumptions

Sofar the likelihood used to estimate causal effect has been consistent with the DGP; we now want to show if the two approaches are robust with respect to misspecification of the parametric assumptions on the distribution of the outcome. With this respect, we present results of a small simulation study, as in section 5.2, where we specify a conditional student t distribution, instead of a normal distribution, for *logY*. In doing that, we maintain the same value for the causal effect within the *RR* stratum, equal to 1 in *log* scale.

Results, reported in Table 4, confirm the lack of robustness of the selection model, as previously noted, among others, in Little (1985) and Copas and Li (1997). As far as the PS estimation is concerned, the effect results positively biased, although the magnitude of the bias is smaller than the one of γ, which is also evident from the comparison of the histograms.

Student t distribution (5 df)			
Parameter	Mean	Q1	Q3
γ	1.51	1.26	1.79
$\alpha_{T=1,RR}$	12.28	12.11	12.43
$\alpha_{T=0,RR}$	11.00	10.92	11.09
$Diff$	1.28	1.08	1.46
1000 replications			

Table 4: Estimation results under PS with non normality as DGP: MonteCarlo average, 1^{th} and 3^{rd} quartile over 1000 replications

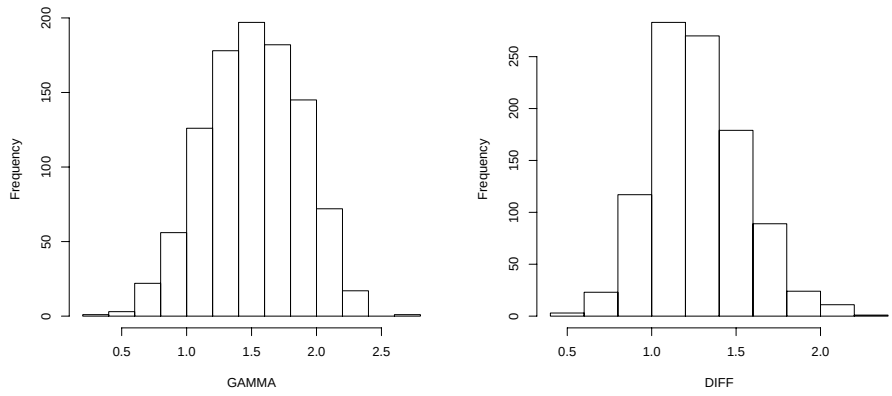

Figure 5: Histogram for γ and $Diff$ MonteCarlo distributions - 1000 replications allowing for non-normality under PS

Even if the PS approach seems to be more robust with respect to misspecification, it leads to biased results. We now want to check, using the testing procedure proposed in previous section, if under the PS approach the misspecified Gaussian model would be rejected. The estimated mixture for group $\{i : T_i = 1, Q_i = R\}$ and the nonparametric density estimation with 95% level confidence bands, reported in Figure 5.2.2, show a substantial lack of fit in the tails of the distribution; in addition, in the central part the estimated

distribution, although within the bands, is far from the non parametric one and squeezed to the lower band.

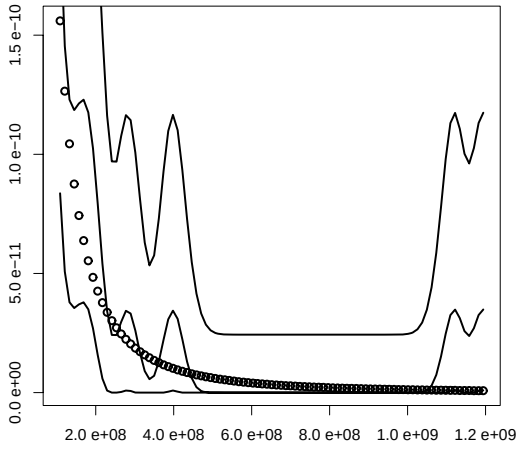

Figure 6: Estimated mixture for group $\{i : T_i = 1, Q_i = R\}$ and nonparametric density estimation with 95% level confidence bands (DGP allowing for non-normality under PS)

6 Allowing for non-monotonicity

As said before, the standard selection model implicity assumes monotonicity and indeed all the equivalence results concerning PS and SM are derived under this assumption. In order to remove this assumption within the SM framework, we need in general to specify a selection model with two versions of Q^* and Y corresponding to the potential outcomes associated with the two treatment levels $T = 0$ and $T = 1$:

$$Y(1) = \alpha^1 + \beta^1 X + u_1^1 \tag{7}$$

$$Q^*(1) = \omega^1 + a^1 X + u_2^1$$

$$Y(0) = \alpha^0 + \beta^0 X + u_1^0 \tag{8}$$

$$Q^*(0) = \omega^0 + a^0 X + u_2^0$$

Monotonicity does not hold because, depending on the values of X and u_2^1 and u_2^0, it can be $Q^*(1) \lesseqgtr Q^*(0)$. The two models (7) and (8) are estimated separately, using observations with $T = 1$ or $T = 0$, respectively.

The causal effect, conditional on X, of T on Y is estimated as

$$(\hat{\alpha}^1 + \hat{\beta}^1 x) - (\hat{\alpha}^0 + \hat{\beta}^0 x),$$

so that the causal estimand is the following:

$$E(Y(1) - Y(0)|X = x),$$

i.e., the effect on Y for the whole population (with given characteristics $X = x$), irrespective of a unit being observed as respondent or non-respondent.

Note that if one want to focus on the following treatment effect

$$E(Y(1) - Y(0)|X = x, Q(1) = 1, Q(0) = 1),$$

i.e., the effect on Y for those who would respond irrespective of being treated or not (our *RR* group), models (7) and (8) do not allow to point identify and estimate it, because the correlations between u_2^1 and u_2^0, between u_1^0 and u_2^1 and between u_1^1 and u_2^0 are not identified with the model assumptions (the correlations never appear in the likelihood function). The effects can only be bounded. It must be pointed out that the group of units such that $Q(1) = 1$ and $Q(0) = 1$ are the only ones for which there is a "direct" evidence on Y under both treatment and control.

Furthermore, as shown by Heckman (1990), there are problems in estimating treatment effects with non and semiparametric versions of selection models (7) and (8), because the intercepts cannot be disentangled from the error correction terms, so that mean outcome levels cannot be estimated and compared. In such cases the "only" way to keep the semiparametric specification and estimate treatment effects is via identification at infinity.

For simulation purposes, we specify a simplified version of models (7) and (8), formulating a single equation for the log-turnover $log(Y)$, so that the effect is constant and captured by the parameter γ:

$$log(Y) = \alpha + \beta X + \gamma T + u_1 \tag{9}$$

$$Q^*(1) = \omega^1 + a^1 X + u_2^1$$

$$Q^*(0) = \omega^0 + a^0 X + u_2^0$$

For the data generating process we assume: $\alpha = 12.5$, $\beta = 1.5$, $\gamma = 1$, $\omega^1 = 0.61$, $a^1 = 0.5$, $\omega^0 = -0.15$, $a^0 = 0.82$, where $u_1 \sim N(0, \sigma_1 = 1)$, $u_2^1 \sim N(0, \sigma_2^1 = 1)$, $u_2^0 \sim N(0, \sigma_2^0 = 1)$, $corr(u_1, u_2^1) = \rho_1 = -0.3$ and $corr(u_1, u_2^0) = \rho_0 = -0.2$.

Estimation results with both SM and PS are reported in Table 5, where the very good performance of PS is evident, given that the estimates of $Diff$ are even less disperse around the true parameter value than the estimates of γ (see also Figure 6).

Parameter	Mean	Q1	Q3
γ	1.00	0.79	1.19
$\alpha_{T=1,RR}$	13.27	13.05	13.47
$\alpha_{T=0,RR}$	12.27	12.14	12.40
$Diff$	1.00	0.81	1.18

Table 5: Estimation results allowing for non-monotonicity under generalized SM as data generating process : MonteCarlo average, 1^{th} and 3^{rd} quartile over 1000 replications

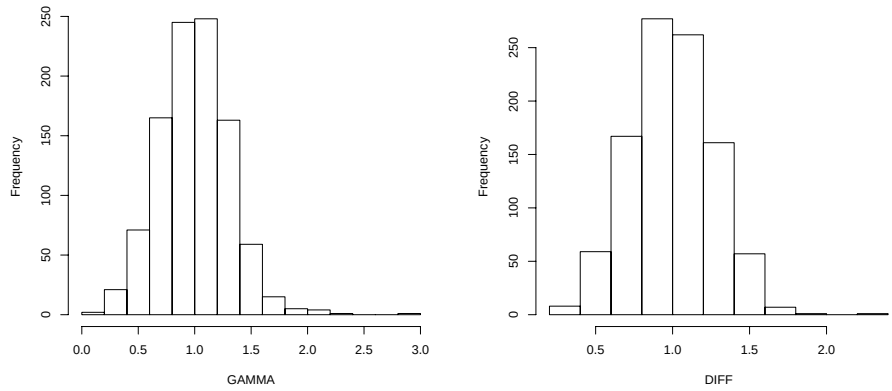

Figure 7: Histogram for γ and $Diff$ MonteCarlo distributions - 1000 replications allowing for non-monotonicity under SM

As in previous sections, we also use the PS framework as DGP, with the following model specification. We still assume a multinomial logit for the distribution of the four principal strata, as in (6), where:

$$\delta_{RR} = 0.5, \quad \gamma_{RR} = 2$$

$$\delta_{RN} = \alpha_{NR} = 0.05, \quad \gamma_{RN} = \gamma_{NR} = 0.6$$

We specify a lognormal model for turnover, as in (5), where:

$$\alpha_{T=1,RR} = 12, \quad \beta_{T=1,RR} = 2, \quad sd_{T=1,RR} = 1$$

$$\alpha_{T=0,RR} = 11, \quad \beta_{T=0,RR} = 2, \quad sd_{T=0,RR} = 1$$

$$\alpha_{T=1,RN} = 13, \quad \beta_{T=1,RN} = 2, \quad sd_{T=1,RN} = 1$$

$$\alpha_{T=0,NR} = 14, \quad \beta_{T=0,NR} = 2, \quad sd_{T=0,RN} = 1$$

Table 6 shows an expected relatively good performance of the PS approach, contrasted by a bad performance of the selection model for which we observe a marked positive bias in the estimated causal effect. We also observed that, while the performance of PS is stable over different generated proportions of the strata[6], the SM crashes when the proportion of the *RR* stratum is low (we found such an evidence for $\pi_{RR} < 0.20$): the estimates of γ are badly biased and the MonteCarlo variabiliy is huge.

Parameter	Mean	Q1	Q3
γ	1.34	1.04	1.60
$\alpha_{T=1,RR}$	12.02	11.92	12.10
$\alpha_{T=0,RR}$	11.00	10.91	11.09
$Diff$	1.04	0.90	1.12

Table 6: Estimation results allowing for non-monotonicity under PS as data generating process : MonteCarlo average, 1^{th} and 3^{rd} quartile over 1000 replications

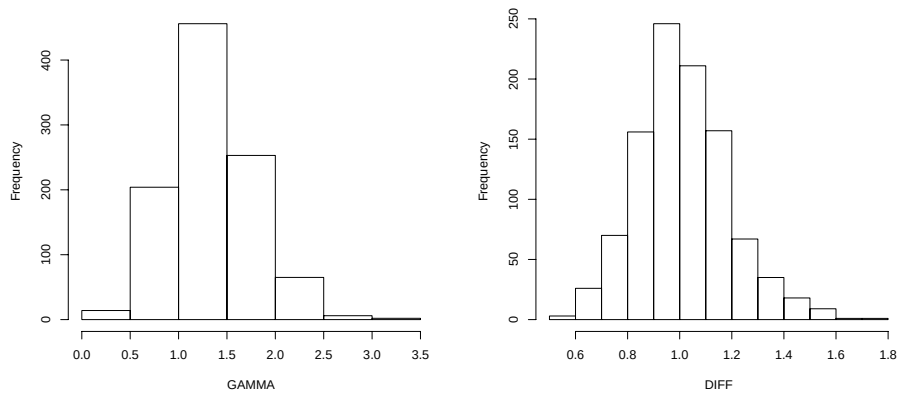

Figure 8: Histogram for γ and $Diff$ MonteCarlo distributions - 1000 replications allowing for non-monotonicity under PS

[6]Remember that, as previously noted, estimation of mixture distributions may be difficult if $\pi_{RR} = \pi_{RN}$ or $\pi_{RR} = \pi_{NR}$.

7 Adding complications

In previous sections we have shown in a simple "though not trivial" setting the different features and performances of the PS and SS approaches. From now on we will deal with several complications, analyzing how parametric or nonparametric identification can be achieved in the two cases. Generally speaking, complications may refer to treatment selection on unobservables (non compliance), non ignorable missingness of the outcome, "censoring due to death".

As an illustrative example, we maintain the context of financial aids to firms, but, first of all, we remove unconfoundedness, so selection depends on unobservables, and we consider a binary instrumental variable Z, which represents the firm's knowledge of the law on financial aids[7]. Furthermore, we introduce "censoring due to death" as an additional (to the non response problem) different source of selection: firms can end their activity before their post-treatment turnover can be observed.

In order to formalize this new setting we must introduce some notation. We denote as S the survival indicator at the time of the interview. The variable Z plays the role of treatment assignment, so for unit i in the study $T_i(z)$,$S_i(z)$, $Q_i(z)$, $Y_i(z)$ ($z = 0, 1$) are, respectively, the potential treatment received indicator, the potential survival indicator, the potential response indicator and the potential outcome variable[8].

Throughout this example we assume randomization of Z conditional on pre-treatment covariates X; this implies that:

$$\text{Assumption } 2: \quad Z \perp T(0), T(1), S(0), S(1), Q(0), Q(1), Y(0), Y(1)|X$$

Principal strata are now defined by the joint value of the six indicators $T(0)$, $T(1)$, $S(0)$, $S(1)$, $Q(0)$, $Q(1)$. It must be noted that the post-treatment non response indicator is usually not used to characterize principal strata, but only the joint values of the other relevant post-treatment variables contribute to the definition of the strata (Frangakis *et al.*, 2003). As a consequence, each latent stratum is actually the union of four of our strata (Q being a binary variable). This collapsing of strata can be justified by a specific assumption on the non response mechanism, such as Latent Ignorability (Frangakis and Rubin, 1999). We prefer not to use such ignorability assumption a priori, because it does not seem defensible in the case of missing turnover.

By construction, because $Z = 0$ implies $T = 0$ (if a firm is not aware of financial aid regulation it cannot have access to financial aid) and so $T = 1$ implies $Z = 1$, we can state that there are no "defiers" ($T(0) = 1$ and $T(1) = 0$) and there are no "alwaystakers" ($T(0) = 1$ and $T(1) = 1$). So, we can

[7] In Italy regulation on financial aids to firms is manifold and complex, so it is not unusual that a firm is unaware of it and possibly the knowledge is not directly related to firm's performance.

[8] Note that, in what follows, and in the tables in particular, we will assume to be inside a cell defined by the values of the covariates, so that all assumptions will hold conditional on the covariates.

only have "complier" firms ($T(0) = 0$ and $T(1) = 1$) and "nevertaker" firms ($T(0) = 0$ and $T(1) = 0$).

If we indicate with $Zobs_i$ the observed value of the instrument, the observed data are

$$\left(T(Zobs_i), S(Zobs_i), Q(Zobs_i), Y(Zobs_i), X_i\right) \qquad i = 1, \ldots, N$$

In the following Table, principal strata and corresponding observed groups are reported. We denote with a dot a missing value of turnover due to nonresponse and with $\nexists$ a missing value of Y due to death in order to distinguish between the two different selection mechanisms.

We can see that, without additional restrictions, the principal strata are 18, but the only one containing information on the effect of T on Y is $G = 9$. Indeed, this is the only stratum where firms are observed receiving and not receiving the treatment T and both $Y(0)$ and $Y(1)$ can be observed and compared (they survive and respond under both treatment and control).

Note that also for stratum $G = 18$ can both $Y(0)$ and $Y(1)$ be observed, although their comparison would not inform us on the effect of T on Y, but rather on the effect of Z on Y.

Because we have not assumed any kind of ignorability of not response so far, the stratum and the causal effect we focus on is different from the estimand of other works (see, for example, Frangakis *et al.*, 2003; Mattei and Mealli, 2007).

Against the latent strata, the observed groups, defined by the values of Z, $Tobs$, $Sobs$ and $Qobs$, are only 9, so that each observed group can be generated by more than one latent stratum. As an example, the observed groups $\{0, 0, 1, 1\}$ and $\{1, 1, 1, 1\}$ may result not only from latent stratum $G = 9$, but also from latent strata $G = 3, 5, 7, 8, 14$ and 17.

The general aim of our inference is, first of all, to estimate the proportions of the strata and, second, to estimate the parameters of the induced mixture distributions within observed strata, in order to obtain the estimate of the outcome distribution conditional on the principal strata. In order to identify these quantities, different strategies can be exploited, some of which are assumptions aiming at reducing the number of latent strata. Plausibility of these assumptions depends on the empirical context and actual data; so the examples presented here should not be considered as having general validity.

Principal Strata(T,S,Q)							Turnover		Observed Groups			
	COMPLIERS											
G	T(0)	T(1)	S(0)	S(1)	Q(0)	Q(1)	Y(0)	Y(1)	Z	$Tobs$	$Sobs$	$Qobs$
1	0	1	0	0	∄	∄	∄	∄	0	0	0	∄
									1	1	0	∄
2	0	1	0	1	∄	0	∄	·	0	0	0	∄
									1	1	1	0
3	0	1	0	1	∄	1	∄	OBS	0	0	0	∄
									1	1	1	1
4	0	1	1	0	0	∄	·	∄	0	0	1	0
									1	1	0	∄
5	0	1	1	0	1	∄	OBS	∄	0	0	1	1
									1	1	0	∄
6	0	1	1	1	0	0	·	·	0	0	1	0
									1	1	1	0
7	0	1	1	1	0	1	·	OBS	0	0	1	0
									1	1	1	1
8	0	1	1	1	1	0	OBS	·	0	0	1	1
									1	1	1	0
9	0	1	1	1	1	1	OBS	OBS	0	0	1	1
									1	1	1	1
	NEVER TAKERS											
10	0	0	0	0	∄	∄	∄	∄	0	0	0	∄
									1	0	0	∄
11	0	0	0	1	∄	0	∄	·	0	0	0	∄
									1	0	1	0
12	0	0	0	1	∄	1	∄	OBS	0	0	0	∄
									1	0	1	1
13	0	0	1	0	0	∄	·	∄	0	0	1	0
									1	0	0	∄
14	0	0	1	0	1	∄	OBS	∄	0	0	1	1
									1	0	0	∄
15	0	0	1	1	0	0	·	·	0	0	1	0
									1	0	1	0
16	0	0	1	1	0	1	·	OBS	0	0	1	0
									1	0	1	1
17	0	0	1	1	1	0	OBS	·	0	0	1	1
									1	0	1	0
18	0	0	1	1	1	1	OBS	OBS	0	0	1	1
									1	0	1	1

For example, we can assume monotonicity with respect to S (financial aids cannot reduce the surviving probability):

$$\text{Assumption } 3: \quad S(1) \geq S(0)$$

which implies that four principal strata do not exist, as shown in Table 7.

	Principal Strata(T,S,Q)						Turnover	
	COMPLIERS							
G	T(0)	T(1)	S(0)	S(1)	Q(0)	Q(1)	Y(0)	Y(1)
1	0	1	0	0	$\nexists$	$\nexists$	$\nexists$	$\nexists$
2	0	1	0	1	$\nexists$	0	$\nexists$	·
3	0	1	0	1	$\nexists$	1	$\nexists$	OBS
6	0	1	1	1	0	0	·	·
7	0	1	1	1	0	1	·	OBS
8	0	1	1	1	1	0	OBS	·
9	0	1	1	1	1	1	OBS	OBS
	NEVER TAKERS							
10	0	0	0	0	$\nexists$	$\nexists$	$\nexists$	$\nexists$
11	0	0	0	1	$\nexists$	0	$\nexists$	·
12	0	0	0	1	$\nexists$	1	$\nexists$	OBS
15	0	0	1	1	0	0	·	·
16	0	0	1	1	0	1	·	OBS
17	0	0	1	1	1	0	OBS	·
18	0	0	1	1	1	1	OBS	OBS

Table 7: Principal strata under Assumptions 2 and 3.

An additional way of reducing the number of strata is by imposing some kind of exclusion restriction. In our context it can be reasonably assumed that for nevertakers (NT) survival is not affected by the knowledge of law:

$$\text{Assumption } 4: \quad S(1) = S(0)|NT$$

which implies the exclusion of two latent strata (see Table 8). A consequence of Assumption 4 is that $Z \perp S(Z)|NT, X$.

Furthermore, we can assume that compliers (C) have the same response behavior irrespective of the knowledge of law:

$$\text{Assumption } 5: \quad Q(1) = Q(0)|C$$

leaving out other two latent strata (see Table 9). As above, a consequence of Assumption 5 is that $Z \perp Q(Z)|C, X$.

We still have 9 observed groups against 10 latent strata, so we cannot non-parametrically identify the proportion of the strata; we could however derive nonparametric bounds on these proportions and consequently derive bounds for the causal effect within the $G = 9$ group.

Principal Strata(T,S,Q)							Turnover	
	COMPLIERS							
G	T(0)	T(1)	S(0)	S(1)	Q(0)	Q(1)	Y(0)	Y(1)
1	0	1	0	0	∄	∄	∄	∄
2	0	1	0	1	∄	0	∄	·
3	0	1	0	1	∄	1	∄	OBS
6	0	1	1	1	0	0	·	·
7	0	1	1	1	0	1	·	OBS
8	0	1	1	1	1	0	OBS	·
9	0	1	1	1	1	1	OBS	OBS
	NEVER TAKERS							
10	0	0	0	0	∄	∄	∄	∄
15	0	0	1	1	0	0	·	·
16	0	0	1	1	0	1	·	OBS
17	0	0	1	1	1	0	OBS	·
18	0	0	1	1	1	1	OBS	OBS

Table 8: Principal strata under Assumptions 2, 3 and 4.

Bounds can be sharpened by adding assumptions that allow to point identify the proportion of strata; for example assuming monotonicity of non response for never takers eliminates stratum $G = 17$, so that we are left with the same number of observed and latent groups. Alternatively, restrictions on turnover distribution would also make bounds sharper. One of such restriction could be a form of stochastic dominance: the distribution of $Y(z)$ for a particular stratum stochastically dominates the distribution of $Y(z)$ for another stratum. For example, we could assume that the distribution of turnover $Y(1)$ for compliers, who survive under both treatment and control, stochastically dominates the distribution of $Y(1)$ for compliers who survive only under treatment:

$$F(Y(1) = y|G = 9) \leq F(Y(1) = y|G = 3)$$

where F denotes the distribution function of turnover. All these forms of stochastic dominance across strata differ from the ones used, for example, by Manski (2003), where the assumption involve the comparison of the distribution of different potential outcomes.

In order to achieve point identification of the effect on turnover for the $G = 9$ group, restrictions on covariates' effect can also be exploited. For example, assume that latent stratun $G = 17$ does not exist, so that the proportions of the 9 latent strata can be identified and estimated by a method of moment approach. Remind that the effect we are interested in is an average treatment effect[9], $E(Y(1) - Y(0)|G = 9)$, and suppose we only have one covariate X with three levels (for example three different sectors of economic

[9]Note that, if we assume latent ignorability conditional on the principal strata defined by T and S, then the effect for the $G = 9$ group can be extended to the $G = 6$ group of non respondents, i.e. to

Principal Strata(T,S,Q)							Turnover		Observed Groups			
	COMPLIERS											
G	T(0)	T(1)	S(0)	S(1)	Q(0)	Q(1)	Y(0)	Y(1)	Z	$Tobs$	$Sobs$	$Qobs$
1	0	1	0	0	∄	∄	∄	∄	0	0	0	∄
									1	1	0	∄
2	0	1	0	1	∄	0	∄	·	0	0	0	∄
									1	1	1	0
3	0	1	0	1	∄	1	∄	OBS	0	0	0	∄
									1	1	1	1
6	0	1	1	1	0	0	·	·	0	0	1	0
									1	1	1	0
9	0	1	1	1	1	1	OBS	OBS	0	0	1	1
									1	1	1	1
	NEVER TAKERS											
10	0	0	0	0	∄	∄	∄	∄	0	0	0	∄
									1	0	0	∄
15	0	0	1	1	0	0	·	·	0	0	1	0
									1	0	1	0
16	0	0	1	1	0	1	·	OBS	0	0	1	0
									1	0	1	1
17	0	0	1	1	1	0	OBS	·	0	0	1	1
									1	0	1	0
18	0	0	1	1	1	1	OBS	OBS	0	0	1	1
									1	0	1	1

Table 9: Principal strata and observed groups under Assumptions 2, 3, 4 and 5

activity). We focus on the two observed groups $(0, 0, 1, 1)$ and $(1, 1, 1, 1)$, for which we can obtain the six turnover sample means $\bar{Y}^{X=i}_{(0,0,1,1)}$ and $\bar{Y}^{X=i}_{(1,1,1,1)}$, $i = 1, 2, 3$. Each sample average can be written as a weighted average of two latent strata turnover averages:

$$\begin{aligned}\bar{Y}^{X=i}_{(0,0,1,1)} &= \bar{Y}^{1}_{G=3,X=i} \cdot \hat{\pi}_{G=3,X=i} + \bar{Y}^{1}_{G=9,X=i} \cdot \hat{\pi}_{G=9,X=i} \\ \bar{Y}^{X=i}_{(1,1,1,1)} &= \bar{Y}^{0}_{G=9,X=i} \cdot \hat{\pi}_{G=9,X=i} + \bar{Y}^{0}_{G=18,X=i} \cdot \hat{\pi}_{G=18,X=i}\end{aligned}$$

This system does not have a unique solution, because there are 6 equations and 12 unknowns ($\bar{Y}^{1}_{3,X=i}$, $\bar{Y}^{1}_{9,X=i}$, $\bar{Y}^{0}_{18,X=i}$,$\bar{Y}^{0}_{9,X=i}$, $i = 1, 2, 3$). However, if we assume that the effect of X on turnover is constant across strata, $\bar{Y}^{T}_{G,X=i} = \bar{Y}^{T}_{G,X=1} + k_i$, $i = 2, 3$, the number of unknowns becomes equal to the number of equations, i.e. six. Similar strategies have been suggested by Jo (2002), to remove exclusion restriction in an IV context, and by Frangakis (2006).

An alternative strategy, that does not necessarily require to match the number of latent and observed groups, is to parametrize the distribution

the complier firms surviving under both treatment and control.

of the outcome conditional on the principal strata, within cells defined by the covariates. As a result, observed distributions are mixtures of parametric distributions analogously to the specification used in section 5: mixture weights as well as parameters of the distributions can be identified and estimated by maximum likelihood exploiting standard finite mixture model theory (McLachlan and Peel, 2000).

While specifying these distributions conditional on covariates, one can also impose restrictions on some parameters, which in general differ depending on the values of the covariates. These restrictions may a) improve estimation in terms of efficiency and b) ease identifiability when this is driven by parametric specification.

7.1 Specifying complications with selection models

In the presence of complications, as the ones introduced in the PS framework (treatment selection on unobservables, censoring due to death and non response), a standard fully parametrized selection model would be specified as follows:

$$\begin{aligned} Y &= \alpha + \beta X + \gamma T + u_1 \\ Q^* &= \omega + aX + bT + u_2 \\ S^* &= \psi + \theta X + \eta T + u_3 \\ T^* &= \phi + \delta X + u_4 \end{aligned} \tag{10}$$

where Y is observed only for $Q^* > 0$ and $S^* > 0$. Multivariate parametric distribution of the error terms would allow identification of model (10), even if the absence of instruments will generally create problems in the estimation process. In case some credible sources of exogenous variation can be found, in the form of an instrumental variable Z, this would be in general included in the model as below:

$$\begin{aligned} Y &= \alpha + \beta X + \gamma T + u_1 \\ Q^* &= \omega + aX + bT + u_2 \\ S^* &= \psi + \theta X + \eta T + u_3 \\ T^* &= \phi + \delta X + \lambda Z + u_4 \end{aligned} \tag{11}$$

Note that the assumptions of randomization and exclusion restrictions of Z with respect to Q, S and Y are all implicitly assumed in model (11). As a consequence, this specification does not allow to pose exclusion restrictions only for specific subgroups of units.

Parametric assumptions, especially related to the distribution of error terms in the selection equations (u_2, u_3 and u_4) can be relaxed by imposing additional restrictions on X, because in this case instrumental variables are required also for Q and S (Das *et al.*, 2003). Consequently, non or semiparametric identification is gathered by exploiting assumptions that are in general different from the ones we have shown previously within the PS framework.

8 Concluding remarks

In this paper we have shown two approaches for dealing with "endogenous selection" problems when estimating causal effects, namely selection models and principal stratification.

Within a parametric setting, we have investigated the different nature of their parametric hypotheses and, in a relatively simple framework, we have demonstrated, by simulating under different parametric specifications, the better performance and robustness of PS. Specifically, under a SM data generating process, both approaches provide estimates of the causal effect that are consistent with the DGP. On the other hand, simulating under a PS model, we find a substantial failure of SM in estimating the causal effect.

In a more complicated setting, which includes more than one post-treatment complications, we have shown how PS is able to suggest alternative identification strategies, not always easily translated into a selection model. Identifiability, within the PS approach, can be achieved if the number of different (sampling) statistics (containing non overlapping pieces of information) is not less than the number of unknowns, characterizing the distribution of potential outcomes and principal strata. With this respect, identifiability can be pursued following different strategies. To summarize, identification can be achieved by:

1. reduction of the number of strata by construction or by assumptions (monotonicity with respect to some post-treatment variables, exclusion restrictions on the intermediate variables);
2. restrictions on potential outcome distributions (e.g., exclusion restrictions on the outcome, stochastic dominance);
3. restrictions based on covariates (e.g., relative or absolute effects assumed to be constant across values of some covariates);
4. restrictions on the parametric specification of the outcome distribution (e.g., assuming specific parametric distribution for the outcome, incorporating assumptions in 1, 2, 3). By restricting some of the parameters one may: a) improve estimation in terms of efficiency if identifiability can be nonparametrically achieved, b) obtain identifiability when it cannot be nonparametrically achieved.

As most of the assumptions are conditional on covariates, it must be noted that sometimes parametric hypotheses are required, because, in finite samples it is infeasible to work within cells defined by the covariates, in particular if they are continuous. Parametrization is also required if the number of strata is large, which happens when post-treatment variables are a lot and/or they can assume many different values. However, the assumptions embedded in a parametric model are more explicit, in terms of the behavior of units, than the ones characterizing SM. Furthermore, the parametric hypotheses imply directly assumptions on the distribution of the observed data. These

hypotheses are, thus, easier to interpret and to judge in terms of plausibility than SM assumptions, which concern error terms, and, finally, they can be more easily empirically tested.

References

[1] Ahn, H. and Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism, Journal of Econometrics, 58, 329.

[2] Angrist, J.D., Imbens, G. W. and Rubin, D.B. (1996). Identification of causal effects using instrumental variables, (with discussion), Journal of the American Statistical Association, 91, 444–472.

[3] Barnard, J., Frangakis, C.E., Hill, J.L., and Rubin, D.B. (2003). A Principal Stratification approach to broken randomized experiments: a case study of School Choice vouchers in New York City, Journal of the American Statistical Association (with discussion), 98: 299-323.

[4] Choi, K. (1990). The semiparametric estimation of the sample selection model using series expansion and the propensity score, mimeo, University of Chicago.

[5] Christofides, L.N., Li, Q., Liu, Z. and Min, I. (2003). Recent two-stage sample selection procedures with an application to the gender wage gap, Journal of Business & Economic Statistics, 21: 396–405

[6] Copas, J. B. and Li, H. G. (1997). Inference for non-random samples (with discussion), Journal of the Royal Statistical Society, Series B, 59: 55–95.

[7] Cosslet, S. (1991). Semiparametric estimation of a regress ion model with sample selectivity, in: Nonparametric and semiparametric methods in econometrics and statistics, W.A. Barnett, J. Powell and G.Tauchen (eds.), Cambridge University Press.

[8] Das, M. and Newey, W.K. and Vella, F. (2003). Nonparametric estimation of sample selection models, Review od Economic Studies, 52: 33–58.

[9] Forcina A. (2006). Causal effects in teh presencce of non compliance: a latent variable interpretation, Metron, LXIV : 1–7.

[10] Frangakis, C.E. (2006). Comment on Forcina (2006), Metron, LXIV : 8–14.

[11] Frangakis, C. E. and Rubin, D. B. (1999). Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncomplianze and Subsequent Missing Outcomes, Biometrika, 86: 365–379.

[12] Frangakis, C.E. and Rubin, D.B.(2002). The defining role of "principal stratification and effects" for comparing treatments adjusted for posttreatment variables: from treatment noncompliance to surrogate endpoints, Biometrics, 58: 191–199.

[13] Gallant, A. and Nychka, D. (1987). Seminonparametric maximum likelihood estimation, Econometrica, 55: 363–390.

[14] Green, P. and Silverman,B. (1994). Nonparametric regression and gereralized linear models, Chapman and Hall.

[15] Gronau, R. (1974). Wage comparison - A selectivity bias, The Journal of Political Economy, 82: 1119-1143.

[16] Haavelmo, T. (1944). The probability approach in econometrics, Econometrica, 15: 413-419.

[17] Haerdle, W. *et al.* (2006). Non- and semiparametric models, Springer.

[18] Hastie, T. and Tibshirani, R. (1990). Generalized additive models, Chapman and Hall.

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

[20] Heckman, J. (1990). Varieties of selection bias, American Economic Review, 80: 313–318.

[21] Holland, P. (1986). Statistics and causal inference, Journal of American Statistical Association, 81: 945–970.

[22] Honore, B. E., Kyriazidou, E, Udry, C. (1997). Estimation of type-3 tobit models using symmetrics trimming and pairwise comparisons, Journal of Econometrics, 76: 107–128.

[23] Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications (with comment and rejoinder), Journal of Educational and Behavioral Statistics, 27, 385420.

[24] Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models, Econometrica, 61: 387–421.

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

[26] Imbens, G.W. and Rubin, D.B. (1997). Bayesian inference for causal eects in randomized experiments with noncompliance, Annals of Statistics, 25: 305–327.

[27] Lee, L.F. (1982). Some approaches in the correction of selectivity bias, Review of Economic Studies, 49: 355–372.

[28] Lee, L.F. (1983). Generalized econometric models with selectivity, Econometrica, 51: 507–512.

[29] Lee, L.F. (1994). Semiparametric two-stage estimation of sample selection models subject to tobit-type selection rules, Journal of Econometrics, 61: 305–344.

[30] Li, Q., Wooldridge, J. (2002). Semiparametric estimationof partially linear models for dependent data with generated regressors, Econometric Theory, 18: 625–645.

[31] Little, R.J.A. (1985). A note about models for selectivity bias, Econometrica, 53, 1569–1474.

[32] McLachlan, G., Peel, D. (2000). Finite Mixture Models, Wiley series in probability and statistics.

[33] Manski, C.F. (2003). Partial Identication of Probability Distributions, Springer-Verlag.

[34] Mattei, A. and Mauro, V. (2007). Valutazione di politiche per le imprese artigiane, Rapporto di ricerca, IRPET.

[35] Mattei, A. and Mealli, F. (2007). Application of the Principal Stratification Approach to the Faenza Randomized Experiment on Breast Self-Examination, Biometrics, forthcoming.

[36] Mealli, F. and Rubin, D. B. (2003). Assumptions allowing the estimation of direct causal eects: Commentary on "health, wealth, and wise? test for direct causal paths between health and socioeconomic status" by Adams et al., Journal of Econometrics, 112: 79–87.

[37] Mealli F., G. Imbens, S. Ferro, A. Biggeri (2004). Analyzing a Randomized Trial on Breast Self Examination with Noncompliance and Missing Outcomes, Biostatistics, 5: 207-222.

[38] Newey, W. (1990). Semiparametric efficiency bounds, Journal of Applied Econometrics, 5: 99–135.

[39] Olsen, R. (1980). A least square correction for selectivity bias, Econometrica, 48: 1815–1820.

[40] Pagan, A. and Ullah, A. (1997). Nonparametric econometrics, Cambridge University Press.

[41] Pagan, A. and Vella, F. (1989). Diagnostic tests for models based on individual data: a survey, Journal of Applied Econometrics, 4: S29-S59.

[42] Powell, J.L., Stock, J.H. and Stoker, T.M. (1989). Semiparametric estimation of index coefficients, Econometrica, 57(6): 1403-1430.

[43] Robinson, P.M. (1988). Root n-consistent semiparametric regression, Econometrica, 56: 931–954.

[44] Rosembaum, P. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment, Journal of the Royal Statistical Society, Series A, 147: 656–666.

[45] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Educational Psycology, 66: 688-701.

[46] Rubin, D.B. (1978). Bayesian inference for causal effects, Annals of Statistics, 6: 34–58.

[47] Rubin, D.B. (2004). Direct and indirect causal effects via potential outcomes, Scandinavian Journal of Statistics, 31: 161–170.

[48] Schafgans, M. and Zinde-Walsh, V. (2002). On Intercept Estimation in the Sample Selection Model, Econometric Theory, 18: 40–50.

[49] Silverman, B.W. (1986). Density estimation for statistics and data analysis, Vol. 26 of Monographs on Statistics and Applied Probability, Chapman and Hall.

[50] Tinbergen, J. (1930). Determination and interpretation of supply curves: an example. Zeitschrift fur Nationalokonomie, reprinted in: The Foundations of Econometric Analysis, D.F. Hendry and M.S. Morgan (eds.), Cambridge, U.K.: Cambridge University Press, 1997.

[51] Vella, F. (1998). Estimating models with sample selection bias: A survey, The Journal of Human Resources, 33: 127–169.

[52] Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result, Econometrica, 70: 331341.

[53] Wooldridge, J. M. (1994). Selection corecctions with a censored selection variable, mimeo, Michigan State University.

[54] Wooldridge, J.M. (2002). Econometric analysis of cross section and panel data, MIT Press.

[55] Zhang, J.L. and Rubin, D.B. (2003). Estimation of causal effects via principal stratication when some outcomes are truncated by death, Journal of Educational and Behavioral Statistics, 28: 353–368.

[56] Zhang J., Rubin D.B., Mealli F. (2006). Evaluating The Effects of Job Training Programs on Wages through Principal Stratification, in: Modeling and Evaluating Treatment Effects in Econometrics, Advances in Econometrics vol. 21, D. Millimet, J. Smith, and E. Vytlacil (eds.), Elsevier Science Ltd, UK.