Universität St.Gallen

Treatment evaluation in the presence of sample selection

Martin Huber

April 2009 Discussion Paper no. 2009-07

Department of Economics University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
Department of Economics
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email vwaabtass@unisg.ch

Publisher: Department of Economics
University of St. Gallen
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35

Electronic Publication: http://www.vwa.unisg.ch

Treatment evaluation in the presence of sample selection[1]

Martin Huber

Author's address:

Martin Huber, Ph.D.
SEW
Varnbüelstrasse 14
9000 St Gallen
Phone +41 71 2299
Fax +41 71 2302
Email Martin.Huber@unisg.ch
Website www.sew.unisg.ch

[1] I have benefited from comments by Joshua D. Angrist, Eva Deuchert, Markus Frölich, Michael Lechner, Giovanni Mellace, Blaise Melly, Rudi Stracke, and by seminar/conference participants in St. Gallen, Engelberg, Bern, Innsbruck, Boston, Mannheim, Geneva, and Shanghai.

Abstract

Sample selection and attrition are inherent in a range of treatment evaluation problems such as the estimation of the returns to schooling or training. Conventional estimators tackling selection bias typically rely on restrictive functional form assumptions that are unlikely to hold in reality. This paper shows identification of average and quantile treatment effects in the presence of the double selection problem (i) into a selective subpopulation (e.g., working - selection on unobservables) and (ii) into a binary treatment (e.g., training - selection on observables) based on weighting observations by the inverse of a nested propensity score that characterizes either selection probability. Root-n-consistent weighting estimators based on parametric propensity score models are applied to female labor market data to estimate the returns to education.

Keywords

Treatment effects, sample selection, inverse probability weighting, propensity score matching.

JEL Classification

C13, C14, C21

1 Introduction

The sample selection problem, which was discussed by Gronau (1974), Heckman (1974), and Vella (1998), among many others, arises whenever the outcome of interest is only observable for some subpopulation that is non-randomly selected even conditional on observed factors. Potential bias due to sample selection related to unobserved characteristics is an issue for a range of treatment evaluation problems, e.g., when estimating the returns to schooling based on a selective subpopulation of working or the effect of school vouchers on college admissions tests, given that some students abstain from the test in a non-random manner.

This paper discusses treatment evaluation under sample selection and attrition when the treatment assignment is non-random and related to observed factors. It considers the case of a double selection problem (i) into the subpopulation for which the outcome is observed (selection on unobservables) and (ii) into the treatment (selection on observables). The main contribution is to show that average and quantile treatment effects are identified by weighting observations by the inverse of a nested propensity score which controls for sample selection bias in the subpopulation with observed outcomes (e.g., working) and treatment selection bias due to non-random treatment assignment.

The present work is related with the literature on inverse probability weighting (IPW), which has long been known as a general approach to tackle selection problems, see Horvitz & Thompson (1952). In the literature on missing data, attrition, and sample selection, Robins & Rotnitzky (1995), Robins, Rotnitzky & Zhao (1995), Rotnitzky & Robins (1995), and Wooldridge (2002, 2007) weight regressions by the inverse of the sample selection propensity score, i.e., the conditional probability to be observed. However, they do not consider selection on unobservables as in this paper. In the treatment evaluation literature relying on the selection on observables or

conditional independence assumption (CIA, see for instance Imbens, 2004), Hirano, Imbens & Ridder (2003) and Firpo (2007) study IPW estimators of average and quantile treatment effects based on weighting by the inverse of the treatment propensity score, the conditional probability to be treated, to control for selection into treatment. Bang & Robins (2005) use IPW in regression models separately for sample selection and treatment selection problems. This paper adds to the literature on IPW by considering both problems within the same model. Identification of treatment effects relies on the inclusion of the (first stage) sample selection propensity score, which is identified using an exclusion restriction, as additional covariate among other observed factors in the (second stage) treatment propensity score.

The paper also contributes to the classic sample selection literature. Under nonparametric identification of the sample selection and treatment propensity scores the identification of treatment effects based on IPW is nonparametric, too. This framework invokes weaker restrictions than the fully parametric selection model in Heckman (1974, 1976, 1979). It is also more general than the semiparametric models of Ahn & Powell (1993), considering a nonparametric sample selection process (e.g., the decision to work), and Newey (2009), considering semiparametric sample selection, who, however, all impose linearity in the outcome equation. Therefore, the selection model discussed in this paper allows for heterogenous effects with respect to observed factors such that the effects may be different for different populations. For this reason the next section discusses identification for various target populations that appear to be interesting for policy interventions. Finally, our model is slightly more general than that of Das, Newey & Vella (2003), who consider a nonparametric sample selection model but still impose additivity of the unobservables which need not be assumed here. Under a parametric specification of the nested propensity score (as considered in the empirical application), identification of treatment effects is semiparametric. This framework is more restrictive with respect to the sample selection process than Ahn

& Powell (1993) and Newey (2009), but more general with respect to effect heterogeneity.

As in the classic sample selection literature, an exclusion restriction is used to identify the sample selection propensity score. Endogeneity only emerges from the sample selection problem. This is distinct from the instrumental variable (IV) literature considering endogenous treatments, see for instance Imbens & Angrist (1994) and Frölich & Melly (2008). Identification in this paper is based on an instrument for sample selection, whereas the IV literature instruments the endogenous treatment directly. Which of the two approaches is accurate, if any, depends on the evaluation problem, the target population, and the data at hand. The framework considered is for example also different to the empirical application in Ahn & Powell (1993), where sample selection and endogeneity in regressors of the outcome equation arises in the same evaluation problem. This requires distinct instruments for selection and the endogenous regressors, whereas we assume conditional exogeneity of the treatment and only instrument selection.

Estimators of the ATE and QTEs naturally arise from the sample analogues of the identification results. Alternatively to IPW estimation, matching estimators (see Rubin, 1973a, 1973b, 1976) on the nested propensity score can be used. Owing to the importance of semiparametric estimation in the empirical treatment evaluation literature, we apply semiparametric IPW and matching (using probit models for the propensity score specifications) to a repeated cross section (1975-1979) from the US Current Population Survey (CPS) previously analyzed by Mulligan & Rubinstein (2008). We estimate the wage differentials between females who went to high school with and without graduation and find that graduating increases average weekly wages by roughly 18 % over dropping out of high school. Furthermore, the graduation effects appear to be larger at higher ranks of the wage distribution. As a robustness check, we also estimate bounds by invoking assumptions previously used by Lechner & Melly (2007), Lee (2009), and Zhang, Rubin & Mealli (2008). $\sqrt{n}$-consistency and asymptotic normality of the semiparametric IPW estimator

is established in the appendix using a GMM framework.

The remainder of the paper is organized as follows. Section 2 introduces a general sample selection model and discusses identification of average and quantile treatment effects for various populations of interest. Section 3 discusses estimation based on IPW, propensity score matching (PSM), and nonparametric bounds and applies it to empirical labor market data from the CPS. Section 4 concludes.

2 Model and identification

2.1 Model

In this section, we introduce a general sample selection model, where the latent outcome is an unknown function of two observed components, the treatment of interest and a vector of covariates, and an unobserved term. Y denotes the latent outcome that is only partially observed conditional on selection, represented by the binary variable S. Let D denote a treatment, which is either 1 (treatment) or 0 (non-treatment). Even though the subsequent discussion focusses on the binary treatment case, it could be easily extended to multiple treatments as discussed in Imbens (2000) and Lechner (2001). Let X, U denote the covariates and the unobserved term, respectively. Throughout the paper we will assume to have an i.i.d. sample of n units before sample selection takes place, indexed by $i = 1, ..., n$. We assume the following model for the latent outcome:

$$Y_i = \varphi(D_i, X_i, U_i), \tag{1}$$

where $\varphi(\cdot)$ is an unknown function. We observe $\{X_i, D_i\}$ for all units in the sample, but Y_i only conditional on $S_i = 1$. The selection indicator S is assumed to be a function of the treatment,

the covariates, an instrument, and an unobserved term:

$$S_i = I\{\zeta(D_i, X_i, Z_i) \geq V_i\}. \tag{2}$$

$I\{\cdot\}$ denotes the indicator function and $\zeta(\cdot)$ is an unknown function. Z represents a one or multi-dimensional instrument which is observable for all units and not directly related with the outcome. V is an unobserved term that is related with U. Due to the dependence of V and U, the observed outcomes are a non-random subsample of latent outcomes. By assumption, S is a function of one element that is excluded in φ, namely the instrument Z. Point identification of treatment effects crucially hinges on this exclusion restriction. Z has to be relevant for S in the sense that it shifts the selection probability considerably conditional on D, X and in general, at least one element of the instrument needs to be continuous.

A classic economic problem to which this model may be applied are the returns to schooling or training. In this case, Y denotes the potential wages which are only observed conditional on employment ($S = 1$) and D represents participation in a training program or educational attainment. X includes other factors that determine wages and are possibly related with D such as work experience. The sample selection problem arises if unobserved factors as motivation affect both the employment decision and potential wages. Identification therefore requires at least one variable (Z) that is related with the employment decision but has no direct effect on wages. In the empirical literature on female wage equations the number of small children in the household and non-wife income have been frequently used as instruments.

2.2 Identification

To identify the causal effects of D, we utilize the potential outcome framework advocated by Rubin (1974), among others. We denote the potential outcome for individual i and some hypothetical

treatment $D = d$ as

$$Y_i^d \quad = \quad \varphi(d, X_i, U_i).$$

The difference $Y_i^1 - Y_i^0$ would identify the individual treatment effect, but is unknown to the researcher, because each individual is either treated or not treated and cannot appear in both states of the world at the same time. As an additional complication, the outcomes are observed for a selective subpopulation. Therefore, effects are only identified when further assumptions are invoked.

If treatment effects were homogenous as assumed in the classic sample selection literature (e.g., Heckman, 1974, 1976, 1979), they would be equal for any individual and population, but this seems implausible for most evaluation problems. Therefore, treatment effects are most likely different for different populations considered. Which target population is most interesting from a policy perspective depends on the particular problem at hand. Lee (2009) and Zhang et al. (2008) consider treatment effects on the subpopulation that is always selected irrespective of the treatment assignment whereas Lechner & Melly (2007) focus on the total population (irrespective of selection). In the subsequent discussion, we will first identify the treatment effects on the subpopulation with observed outcomes, i.e., conditional on being selected, and then show identification for the total population by imposing slightly stronger conditions. After having established our main results, we will also discuss how effects on further target populations can be identified.

For the moment, let us assume that we want to learn about the average treatement effect (ATE), denoted as $\Delta_{S=1}$, and quantile treatment effect (QTE), denoted as $\Delta_{S=1}^{\tau}$, on the subpop-

ulation with observed outcomes:

$$\begin{aligned}\Delta_{S=1} &= E[Y^1|S=1] - E[Y^0|S=1], \\ \Delta^{\tau}_{S=1} &= Q^{\tau}_{Y^1|S=1} - Q^{\tau}_{Y^0|S=1}.\end{aligned}$$

τ denotes the rank of the potential outcome distribution at which the QTE is evaluated and is bounded between 0 and 1. E.g., $\tau = 0.5$ yields the median effect of the treatment. $Q^{\tau}_{Y^d|S=1}$ denotes the quantile of the potential outcome for treatment $D = d$ in the subpopulation with observed outcomes and is defined as $\inf_y \Pr(Y^d \leq y|S=1) \geq \tau$.

Briefly speaking, identification in this paper is based on 3 key assumptions: (i) Conditional independence of potential outcomes and treatments in the total population, (ii) the availability of an exclusion restriction to identify the sample selection propensity score, and (iii) conditional independence of observables and unobservables given the sample selection propensity score.

Assumption 1: Conditional independence of treatments and latent potential outcomes.

(1a) $Y^1, Y^0 \perp D|X = x, \forall\ x \in \mathcal{X}$ (conditional independence of the latent outcome),

(1b) $0 < \Pr(D = 1|X) < 1$ (common support of D in X),

The conditional independence assumption (CIA) or selection on observables assumption is frequently imposed in the treatment evaluation literature, see for instance Heckman, Ichimura & Todd (1997) and Lechner (1999). (1a) states that the potential latent outcome is independent of the treatment given the observed covariates X. This implies that all factors jointly affecting the treatment assignment and the latent outcome can be controlled for by conditioning on the covariates. The difference to conventional evaluation studies relying on the CIA is that the outcome is not fully observed. (1b) is a common support assumption and states that the selection probability must not be perfectly predicted conditional on the covariates. If in

addition to Assumption 1 the Stable Unit Treatment Value Assumption (SUTVA, see Rubin, 1990) is satisfied, stating that the potential outcome for any individual is stable in the sense that it takes the same value independent of treatment allocations in the rest of the population, it holds that

$$\begin{aligned} E[Y^1|D &= 0, X = x] = E[Y^1|D = 1, X = x] = E[Y|D = 1, X = x], \\ E[Y^0|D &= 1, X = x] = E[Y^0|D = 0, X = x] = E[Y|D = 0, X = x]. \end{aligned}$$

The ATE conditional on X is $\Delta(x) = E[Y^1|X = x] - E[Y^0|X = x] = E[Y|D = 1, X = x] - E[Y|D = 0, X = x]$. Thus, under Assumption 1, the effect of D on Y *could* be identified conditional on X if the outcome *was* fully observed. However, as unobservables V and U are not independent even conditional on X, the treatment effect is confounded in the subpopulation with observed outcomes. Point identification requires the availability of an instrument Z that predicts selection S but is not related with Y conditional on D, X. We therefore make the following assumption.

Assumption 2: Exclusion restriction.

(2a) $\text{Cov}(Z, S|X, D) \neq 0$ and $Y \perp Z|D, X$ (exclusion restriction),

(2b) $\Pr(S = 1|D = d) > c$, $c > 0$, $d \in \{1, 0\}$ (positive conditional selection probability given D),

(2c) $(U, V) \perp (D, Z)|X, \Pr(S = 1|D, X, Z)$ (conditional independence of unobservables and D, Z given X),

(2d) $F_V(t)$, the cdf of V, is strictly monotonic in the argument t.

Assumption (2a) states that Z shifts S but is independent of the latent outcome given D, X. Direct effects of Z on Y are ruled out. Together with assumption 1, this implies that $F_{(Y|D,X)} = F_{(Y|D,X,Z)}$ for all values of Z, where $F_{(\cdot|\cdot)}$ denotes the conditional cdf. (2b) rules out that being treated or nontreated perfectly predicts non-selection. To see the usefulness of this assumption,

assume the opposite that units with $D = 0$ are never selected independent of the values of X, Z. Obviously, the treatment effect cannot be evaluated as no comparisons with $D = 0$ are available in the subpopulation with observed outcomes.

By (2c), we impose that D, Z are jointly independent of the unobservables U, V given X and the conditional selection probability $\Pr(S = 1|D, X, Z)$. (2c) is for instance violated if U is related to D in the total population conditional on X (and $\Pr(S = 1|D, X, Z)$ which will be kept implicit in the subsequent discussion). Then, the selection bias cannot be controlled for by controlling for X, as unobserved interaction terms of U and D drive the selection probability. To illustrate this issue by means of an example, assume that we are interested in the effects of a training (D) on wages (Y) and that motivation (U) is not observed. Assumption (2c) would be violated if the variance of motivation (and thus, of potential wages) differed for individuals with and without training, but with the same observed factors like age, education, work experience, and others. Albeit strong, equivalent or similar assumptions are crucial for point identification in any selection model of both parametric and general form.

Note that $\Pr(S = 1|D, X, Z) = \Pr(\zeta(D, X, Z) \geq V) = F_V(\zeta(D, X, Z))$. By the monotonicity assumption (2d) it holds that the likelihood to be selected increases monotonically in ζ. Monotonicity is implicitly assumed in any linear index restriction frequently used in the sample selection literature. However, it is a rather strong restriction and its plausibility needs to be evaluated from case to case. E.g., if V reflects ability or motivation and S is employment, it seems reasonable to assume that (2d) holds, as more able and motivated individuals may have a higher intrinsic utility from work and also higher potential wages (extrinsic utility). As a second example, let S denote summer school participation and V ability. If the least able students are likely to participate due to force and the most able students due to personal interest, the monotonicity assumption clearly fails.

By comparing individuals with the same response propensity score under the satisfaction of (2d), we control for V and thus, also for the dependence between V and U. I.e., by fixing V, we rule out confounding of the treatment effect due to attrition related to unobservables. The response propensity score serves as a control function where the exogenous variation comes from Z. Control functions have been applied in semi- and nonparametric sample selection models, e.g., Ahn & Powell (1993) and Das et al. (2003) as well as in nonparametric models with endogeneity, see for example Newey, Powell & Vella (1999), Blundell & Powell (2003), and Imbens & Newey (2003).

For notational ease, let $W \equiv (D, X, Z)$ and $p(W) \equiv \Pr(S = 1|D, X, Z)$. Under Assumption 2, U and D are independent conditional on $p(W)$ and X, which can be shown analogously to the proof of Theorem 1 in Newey (2007). Let $a(U)$ denote any bounded function of U. Note that $\{S = 1\} = \{F_V^{-1}(p(W)) \geq V\}$. Then,

$$\begin{aligned}
E\left[a(U)|D, X, p(W), S = 1\right] &= E\left[E\left[a(U)|V, D, X, Z\right]|D, X, p(W), F_V^{-1}(p(W)) \geq V\right] \\
&= E\left[E\left[a(U)|V, X\right]|D, X, p(W), F_V^{-1}(p(W)) \geq V\right] \\
&= E\left[E\left[a(U)|V, X\right]|X, p(W), F_V^{-1}(p(W)) \geq V\right] \\
&= E\left[E\left[a(U)|V, X, p(W)\right]|X, p(W), S = 1\right] \\
&= E\left[a(U)|X, p(W), S = 1\right],
\end{aligned}$$

where the first equality follows from iterated expectations, the second and third from (2c), and the last from a backward application of the law of iterated expectations.

Thus, as any bounded function of U and D are independent conditional on $p(W)$ and X, sample selection bias among those with observed outcomes can be controlled for by including the sample selection propensity score as additional conditioning variable besides the covariates X. To see this, note that the conditional ATE given X and $p(W)$ in the selected subpopulation is

defined as

$$\begin{aligned}\Delta_{S=1}(x, p(w)) &= \int \varphi(1, x, u) dF_{u|X=x,p(W)=p(w),S=1} \\ &- \int \varphi(0, x, u) dF_{u|X=x,p(W)=p(w),S=1} \\ &= E[Y^1|X = x, p(W) = p(w), S = 1] - E[Y^0|X = x, p(W) = p(w), S = 1].\end{aligned}$$

$E[Y^d|X = x, p(W) = p(w), S = 1]$ is the expected potential outcome for a hypothetical treatment d given X and $p(W)$ in the subpopulation with observed outcomes. By the conditional independence of U and D given $p(W)$ and X, it holds that

$$\begin{aligned}&E[Y^d|X = x, p(W) = p(w), S = 1] = \int \varphi(d, x, u) dF_{u|X=x,p(W)=p(w),S=1} \\ &= \int \varphi(d, x, u) dF_{u|D=d,X=x,p(W)=p(w),S=1} \\ &= E[Y|D = d, X = x, p(W) = p(w), S = 1].\end{aligned}$$

Hence, the expected *potential* outcome is equal to the expected *conditional* outcome given $D = d$. The ATE $\Delta_{S=1}$ is identified by the integration over the marginal distributions of X and $p(W)$ in the subpopulation with observed outcomes.

$$\begin{aligned}&\int\int [E[Y|D = 1, X = x, p(W) = p(w), S = 1] \\ &- E[Y|D = 0, X = x, p(W) = p(w), S = 1]] dF_{x|p(W)=p(w),S=1} dF_{p(w)|S=1} \\ &= \int\int [E[Y^1|X = x, p(W) = p(w), S = 1] \\ &- E[Y^0|X = x, p(W) = p(w), S = 1]] dF_{x|p(W)=p(w),S=1} dF_{p(w)|S=1} \\ &= E[Y^1 - Y^0|S = 1] = \Delta_{S=1}. \qquad (3)\end{aligned}$$

The identification of QTEs requires that the conditional quantiles of interest are unique. I.e., the density in the neighborhood of the quantiles must be bounded away from zero such that each quantile corresponds to exactly one particular rank in the conditional distribution. Furthermore,

for an intuitive interpretation of QTEs, the rank stability assumption has to be satisfied across treatments. It states that individuals occupy the same rank in potential outcome distributions for different treatments, see for instance Firpo (2007) for more discussion.

Let $Q^{\tau}_{Y^d|S=1}(x, p(w))$ denote the τth quantile of the potential outcome Y^d given $X = x, p(W) = p(w)$, and $S = 1$. By Assumption 2,

$$\begin{aligned} F_{Y|D,X,p(W),S=1}(y|d, x, p(w), 1) &= \int I\{\varphi(d, x, u) \leq y\} dF_{u|D=d,X=x,p(W)=p(w),S=1} \\ &= \int I\{\varphi(d, x, u) \leq y\} dF_{u|X=x,p(W)=p(w),S=1} \\ &= Q^{\tau^{-1}}_{Y^d|S=1}(x, p(w)). \end{aligned}$$

The unconditional quantile of the potential outcome is identified as the inverse of the integration over the marginal distributions of X and $p(W)$ given $S = 1$.

$$\int \int Q^{\tau^{-1}}_{Y^d|S=1}(x, p(w)) dF_{x|(p(W)=p(w),S=1} dF_{p(w)|S=1} = Q^{\tau^{-1}}_{Y^d|S=1}. \tag{4}$$

The difference between the quantiles under treatment and non-treatment yields the QTE:

$$\Delta^{\tau}_{S=1} = Q^{\tau}_{Y^1|S=1} - Q^{\tau}_{Y^0|S=1}. \tag{5}$$

Identification of $\Delta_{S=1}, \Delta^{\tau}_{S=1}$ hinges on the common support of the treatment in X and $p(W)$ in the subpopulation with observed outcomes. We therefore impose a further assumption:

Assumption 3: Common support in the treatment propensity score among the selected.

(3a) $c < \Pr(D = 1|X = x, p(W) = p(w), S = 1) < 1 - c, \forall\ x \in \mathcal{X}, \forall\ p(w) \in \mathcal{P}, c > 0$

(common support of D in X and $p(W)$).

Assumption (3) states that the treatment propensity score conditional on being observed is bounded away from zero and one. It is obvious that Assumption (2b) is a necessary condition for Assumption (3) to hold. E.g, if the outcomes of individuals with $D = 1$ were never observed,

$\Pr(D = 1|X, p(W), S = 1)$ would always be zero. Assumption (2b) is, however, not sufficient for (3). Consider the case that all individuals receiving treatment $D = 1$ and having characteristics $X = x$ are selected, which is not ruled out by (2b). I.e. $D = 1, X = x$ implies $p(W) = 1$, independent of Z. If $p(W) < 1$ for $D = 0$ and any other value of Z given $X = x$, it follows that $\Pr(D = 1|X = x, p(W) = 1) = 1$. Thus, $p(W) = 1$ perfectly predicts that $D = 1$ conditional on $X = x$ in the subpopulation with observed outcomes such that the common support assumption fails. Unless the selection probability conditional on $D = 1, X = x$ is not smaller than one, identification requires that there exists some combination of $(D = 0, Z = z)$ with $p(W) = 1$ given $X = x$.

At this point, let us consider the special case that Assumption (3) is satisfied and $p(W) = 1$ for some triples (D, X, Z). Obviously, selection bias is not an issue for these observations as $E[Y|D = d, X = x, p(W) = 1, S = 1] = E[Y|D = d, X = x, p(W) = 1]$. This allows identifying local treatment effects for the subpopulation with $p(W) = 1$, given that there is variation in the treatment state. It remains a priori unclear why this particular population should be of any policy interest. However, if one is willing to impose the strong restriction of treatment effect homogeneity across selection probabilities, i.e. $\Delta_{S=1}(x, p(w)) = \Delta_{S=1}(x)\ \forall\ p \in \mathcal{P}$, treatment effects can be identified for other populations as well conditional on common support in X. Identification based on $p(W) = 1$ is known as 'identification at infinity' and was discussed by Heckman (1990) and Andrews & Schafgans (1998). However, in empirical applications, observation with selection probabilities close to one might be rare and effect homogeneity in $p(W)$ is a strong assumption that might not hold in reality. We therefore concentrate on a more general identification strategy using the whole distribution of $p(W)$.

After having established the identifying assumptions, we will now propose expressions for $\Delta_{S=1}, \Delta^{\tau}_{S=1}$ based on inverse probability weighting (IPW) which can be used to build sample

analogues required for estimation. Let $\pi(X, p(W))$ denote the treatment propensity score, i.e., the probability of being treated conditional on X and $p(W)$, $\pi(X, p(W)) \equiv \Pr(D = 1|X, p(W))$. To control for selection into treatment, we will henceforth condition on $\pi(X, p(W))$ instead of X and $p(W)$. Rosenbaum & Rubin (1983) have shown that conditioning on the treatment propensity score is equivalent to conditioning on the covariates directly, as both are balancing scores in the sense that they adjust the distributions of covariates in the groups of treated and controls. However, conditioning on $\pi(X, p(W))$ will have the advantage that practical problems related to the nonparametric estimation based on high dimensional covariates, e.g., empty cells for particular combinations of covariate values, can be circumvented.

PROPOSITION 1 (Identification of mean effects on the selected subpopulation).

Under Assumptions 1, 2, and 3, the ATE in the subpopulation with observed outcomes is identified by

$$\Delta_{S=1} = E\left[\frac{D \cdot Y}{\pi(X, p(W))}|S = 1\right] - E\left[\frac{(1 - D) \cdot Y}{1 - \pi(X, p(W))}|S = 1\right]. \tag{6}$$

Proof: See Appendix A.

The ATE on the selected subpopulation is identified by reweighing the observed outcomes by the inverse of the conditional treatment probability given X and $p(W)$. An analogous approach identifies the quantiles and the QTE.

PROPOSITION 2 (Identification of quantiles in the selected subpopulation).

Under Assumptions 1, 2, and 3, $Q^{\tau}_{Y^1|S=1}$ is an implicit function of

$$E\left[\frac{D}{\pi(X, p(W))} \cdot I\{Y \leq Q^{\tau}_{Y^1|S=1}\}|S = 1\right] = F_{Y^1|S=1}(Q^{\tau}_{Y^1|S=1}) = \tau. \tag{7}$$

Proof: See Appendix B.

It follows that

$$Q^{\tau}_{Y^1|S=1} = \arg\operatorname{zero}_y E\left[\frac{D}{\pi(X, p(W))} \cdot (I\{Y < y\} - \tau)|S = 1\right],$$

which is a first order condition to

$$Q^{\tau}_{Y^1|S=1} = \arg\min_y E\left[\frac{D}{\pi(X, p(W))} \cdot \rho_\tau(Y - y)|S = 1\right]. \tag{8}$$

$\rho_\tau(a) \equiv u \cdot (\tau - I\{a < 0\})$ denotes the check function, an asymmetric loss function suggested by Koenker & Bassett (1978) for quantile regression. An equivalent identification result holds for $Q^{\tau}_{Y^0|S=1}$ and it follows that $\Delta^{\tau}_{S=1} = Q^{\tau}_{Y^1|S=1} - Q^{\tau}_{Y^0|S=1}$. Based on reweighing observed outcomes by the inverse of the nested propensity score, we identify the ATE and QTEs in the selected subpopulation.

As noted by Newey (2007), without further assumptions, effects cannot be identified for other groups than the selected subpopulation, as Y is not even observed when $S = 0$. However, under particular common support conditions and conditional homoscedasticity of Y, the IPW framework even allows identifying the ATE on the total population ($\Delta = E[Y^1] - E[Y^0]$), i.e., irrespective of selection. To this end, we make the following two assumptions:

Assumption 4: Common support in the sample selection and treatment propensity scores.

(4a) $\Pr(S = 1|D = d, X = x, Z = z) > c,\ c > 0, \forall\ x \in \mathcal{X}, \forall\ Z \in \mathcal{Z}$ (positive sample selection propensity score),

(4b) $c < \Pr(D = 1|X = x, p(W) = p(w)) < 1 - c, \forall\ x \in \mathcal{X}, \forall\ p(w) \in \mathcal{P}$ (common support in the treatment propensity score), $c > 0$,

(4a) states that the sample selection propensity score is bounded away from zero, which is stronger than (2b). Effects on the total population could not be identified if there existed individuals with a sample selection propensity score equal to zero as this would rule out suitable comparisons in the subpopulation with observed outcomes. (4b) states that there must be

common support in the treatment propensity score in the population.

Assumption 5: Separability of observed and unobserved terms.

(5a) $Y = \varphi(D, X) + U$ (separability).

Assumption 5 decreases the generality of our model, but ensures homoscedasticity of Y given (D, X), which is required for the subsequent proposition. Nonparametric sample selection models with additive unobserved terms have also been considered in Das et al. (2003).

PROPOSITION 3 (Identification of mean effects on the total population).

Under Assumptions 1, 2, 4, and 5, the ATE on the total population is identified by

$$\Delta \quad = \quad E\left[\frac{S \cdot D \cdot Y}{p(W) \cdot \pi(X, p(W))}\right] - E\left[\frac{S \cdot (1-D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))}\right]. \tag{9}$$

Proof: See Appendix C.

The ATE on the total population is identified based on reweighing observations (additionally to the inverse treatment propensity score) by the inverse of the sample selection propensity score, i.e., by using the relative likelihood of a particular triple (D, X, Z) to appear in the total population as weighting function. It may seem surprising that identification is possible even though outcomes are only partially observed and the observed outcomes do generally not allow inferring on the unobserved outcomes. I.e., $E[Y|D = d, X = x, p(W) = p(w), S = 1] \neq E[Y|D = d, X = x, p(W) = p(w), S = 0]$ due to different conditional distributions of the unobserved term U. However, Assumptions (2c) and (5) imply that $\Delta_{S=1}(x, p(w)) = \Delta_{S=0}(x, p(w))$. To see this, note that by Assumption (2c), $F_{U|D=d,X=x,p(W)=p(w),S=s} = F_{U|X=x,p(W)=p(w),S=s}$ for $s \in \{0, 1\}$

such that

$$\begin{aligned}
\Delta_{S=1}(x, p(w)) &= \int [\varphi(1, x) + u] dF_{u|X=x, p(W)=p(w), S=1} \\
&- \int [\varphi(0, x) + u] dF_{u|X=x, p(W)=p(w), S=1}, \\
\Delta_{S=0}(x, p(w)) &= \int [\varphi(1, x) + u] dF_{u|X=x, p(W)=p(w), S=0} \\
&- \int [\varphi(0, x) + u] dF_{u|X=x, p(W)=p(w), S=0}.
\end{aligned}$$

$\Delta_{S=1}(x, p(w))$ and $\Delta_{S=0}(x, p(w))$ only differ with respect to the integrals over different conditional distributions of U given $S = 1$ and $S = 0$, which cancel out in the subtractions due to the additivity assumption. Thus, $\Delta_{S=1}(x, p(w)) = \Delta_{S=0}(x, p(w))$. Therefore, reweighing the conditional treatment effects in the subpopulation with observed outcomes according to the distribution of (D, X, Z) in the total population identifies Δ.

It seems useful to confront our results to Wooldridge (2002, 2007) who discusses IPW M-estimation of missing data models. Wooldridge considers the estimation of the general objective function $m(A; \theta)$, where A denotes a data matrix and θ is the parameter of interest. The latter is identified by the moment condition $E\left[\frac{S}{p(W)} m(A; \theta)\right] = 0$. By defining $m(A; \theta)$ as $\left(\frac{D}{\pi(X,p(W))} - \frac{(1-D)}{1-\pi(X,p(W))}\right) \cdot (Y - \theta)$, it follows that $E\left[\frac{S}{p(W)} \cdot \left(\frac{D}{\pi(X,p(W))} - \frac{(1-D)}{1-\pi(X,p(W))}\right) \cdot (Y - \theta)\right] = 0$ such that θ identifies the ATE on the total population. At a first glance, our results appear to be a special case.

However, the framework of Wooldridge (2002, 2007) is somewhat different because it does not consider sample selection on unobservables such that the sample selection propensity score $p(W)$ does not enter the objective function $m(A; \theta)$. I.e., V, the unobserved term in S must not be related with U, the unobserved factor in Y, whereas instrument Z may be related with U. In the selection on unobservables framework treated in this paper (which also underlies the classic sample selection literature) Z must not be related with V and U, but V may be related with

U, see Fitzgerald, Gottschalk & Moffitt (1998) for a discussion of these distinct assumptions. For the same reason, our sample selection problem also differs from Robins & Rotnitzky (1995), Robins et al. (1995), and Rotnitzky & Robins (1995), who consider IPW adjusted regression under selection on observables.

Furthermore, we can link our work to identification based on IPW under the CIA, see for instance Hirano et al. (2003) and Firpo (2007). The validity of the CIA in the absence of sample selection implies that the treatment effect is unconfounded conditional on the treatment propensity score with respect to X alone. In our framework, we need to condition on both X and $p(W)$ to control for selection into the subpopulation with observed outcomes *and* into the treatment.

2.3 Further target populations

We have discussed the identification of treatment effects on the subpopulation with observed outcomes and on the total population. However, depending on the evaluation problem, different target populations might be relevant from a policy perspective. E.g., Lee (2009) and Zhang et al. (2008) focus on the subpopulation of those being selected irrespective of the treatment assignment. Let S^d denote the potential sample selection indicator for treatment $D = d$. If one is willing to assume that the sample selection increases uniformly in the treatment (see for instance Lechner and Melly, 2007, and Lee, 2009), i.e., $\Pr(S^1 \geq S^0) = 1$, then those observations with $(S = 1, D = 0)$ are always selected irrespective of the treatment assignment, satisfying $(S^1 = 1, S^0 = 1)$. The always selected, or 'always takers' in the notation of Imbens & Angrist (1994), are the *nontreated* individuals in the subpopulation with observed outcomes.

Hirano et al. (2003) discuss the identification of weighted ATEs based on IPW, which provides a general framework for the identification of treatment effects on different target populations.

Translated to our sample selection framework their results imply that

$$\Delta_{g|S=1} = \frac{1}{E[g|S=1]} \cdot E\left[\frac{D \cdot Y \cdot g}{\pi(X,p(W))} - \frac{(1-D)\cdot Y \cdot g}{1-\pi(X,p(W))}|S=1\right],$$

where g is a general weighting function. For the the always selected, the weight to be used is the propensity not to receive the treatment, $1-\pi(X,p(W))$, because reweighing the conditional effect $\Delta_{S=1}(x,p(w))$ and integrating over the distributions of X and $p(W)$ in the selected sample yields the ATE on the always selected, denoted as $\Delta_{S=1,D=0}$:

$$\begin{aligned}\Delta_{S=1,D=0} &= \int\int \Delta_{S=1}(x,p(w))dF_{x|p(W)=p(w),D=0,S=1}dF_{p(w)|D=0,S=1} \\ &= \int\int \Delta_{S=1}(x,p(w))(1-\pi(x,p(w)))dF_{x|p(W)=p(w),S=1}dF_{p(w)|S=1}/ \\ &\quad \int\int (1-\pi(x,p(w)))dF_{x|p(W)=p(w),S=1}dF_{p(w)|S=1}.\end{aligned}$$

Therefore, $\Delta_{S=1,D=0}$ is identified by

$$\Delta_{S=1,D=0} = \frac{1}{\Pr(D=0|S=1)} \cdot E\left[D\cdot Y \cdot \frac{1-\pi(X,p(W))}{\pi(X,p(W))} - (1-D)\cdot Y|S=1\right],$$

where $\Pr(D=0|S=1) = E[1-\pi(X,p(W))|S=1]$. All observations $(S=1,D=1)$ are reweighed by $\frac{1-\pi(X,p(W))}{\pi(X,p(W))}$ such that they are comparable to the always selected $(S=1,D=0)$ in terms of the treatment propensity score. Similarly, the quantile $Q^{\tau}_{Y^1|S=1,D=0}$ is an implicit function of

$$E\left[\frac{D}{\Pr(D=0|S=1)} \cdot \frac{1-\pi(X,p(W))}{\pi(X,p(W))} \cdot I\{Y \leq Q^{\tau}_{Y^1|S=1,D=0}\}|S=1\right],$$

see also the discussion on the identification of quantile treatment effects on the treated (QTET) in Firpo (2007). An equivalent result holds for $Q^{\tau}_{Y^0|S=1,D=0}$, which implies the identification of $\Delta^{\tau}_{S=1,D=0}$. Note that Assumption 3 can be relaxed to $c < \Pr(D=1|X,p(W),S=1)$, $c>0$, which suffices for the exclusion of arbitrarily large weights $\frac{1-\pi(X,p(W))}{\pi(X,p(W))}$.

By the same logic, the ATE on those with $(S = 1, D = 1)$ is identified by weighting with $\pi(X, p(W))$. Given that uniformity of S in D holds, this group is made up by two subpopulations, namely the always selected $(S^1 = 1, S^0 = 1)$ and those individuals who are selected under treatment, but would not be under non-treatment $(S^1 = 1, S^0 = 0)$. In the spirit of Imbens & Angrist (1994), we refer to this latter group as compliers, where compliance means that the selection state reacts on the treatment assignment. E.g., when evaluating the returns to a training, the compliers are those who switch into employment when being placed into a training. Evaluating the effects on the potential wages of individuals who change their labor market behavior in the light of the treatment may be of great policy relevance and compliers appear to be an interesting population in many other problems, too. We can identify the ATE on the compliers, denoted as $\Delta_{S^1=1,S^0=0}$, by making the following observation:

$$\begin{aligned}
\Delta_{S=1} &= \Delta_{S=1,D=1} \cdot \Pr(D = 1|S = 1) + \Delta_{S=1,D=0} \cdot \Pr(D = 0|S = 1), \text{ where} \\
\Delta_{S=1,D=1} &= \Delta_{S^1=1,S^0=1} \cdot \Pr(S^1 = 1, S^0 = 1|S = 1, D = 1) \\
&+ \Delta_{S^1=1,S^0=0} \cdot (1 - \Pr(S^1 = 1, S^0 = 1|S = 1, D = 1)) \\
&= \Delta_{S=1,D=0} \cdot \frac{\Pr(S = 1|D = 0)}{\Pr(S = 1|D = 1)} + \Delta_{S^1=1,S^0=0} \cdot \left(1 - \frac{\Pr(S = 1|D = 0)}{\Pr(S = 1|D = 1)}\right).
\end{aligned}$$

The first and second equalities follow from the law of total probability. The third equality holds because of $\Pr(S^1 \geq S^0) = 1$ such that the always selected are one subpopulation in $(S = 1, D = 1)$. Their fraction is $\frac{\Pr(S=1|D=0)}{\Pr(S=1|D=1)}$, i.e., the share of individuals that would even be selected without treatment among those selected under the treatment. Therefore, the remaining fraction $1 - \frac{\Pr(S=1|D=0)}{\Pr(S=1|D=1)}$ must be made up by compliers, see also Lee (2009). This allows identifying the

ATE on the compliers by

$$\begin{aligned}\Delta_{S^1=1,S^0=0} &= \Delta_{S=1,D=1} \cdot \left(1 - \frac{\Pr(S=1|D=0)}{\Pr(S=1|D=1)}\right)^{-1} \\ &- \Delta_{S=1,D=0} \cdot \frac{\Pr(S=1|D=0)}{\Pr(S=1|D=1)} \cdot \left(1 - \frac{\Pr(S=1|D=0)}{\Pr(S=1|D=1)}\right)^{-1}.\end{aligned}$$

The framework of weighted treatment effects could be used to identify the effects on further target populations, but this is beyond the scope of this paper. The empirical application will focus on the subpopulation with observed outcomes.

3 Empirical application

In this section we estimate a female wage equation using a subsample of the US Current Population Survey (CPS) data previously analyzed by Mulligan & Rubinstein (2008). The sample consists of a repeated cross section that covers the years 1975 to 1979 and contains information on white females aged between 25 and 54. Individuals are classified as working ($S = 1$) if they work 35+ hours per week paid at least 50 weeks during the year. Self-employed and persons in the military, agriculture, or private household sectors as well as individuals with inconsistent reports on earnings or with allocated earnings are excluded from the sample with observed wages, see Mulligan & Rubinstein (2008) for further details. The outcome of interest (Y) is the wife's hourly wage, which is computed based on total annual earnings which are deflated by the US Consumer Price Index (CPI).

We are interested in the ATE and QTEs of graduating from high school ($D = 1$) vs. receiving 9 to 11 years of high school education without graduation ($D = 0$) on wages on those who went to high school and work. The evaluation sample consists of 67,848 observations, thereof 52,354 high school graduates and 15,494 high school drop outs. 20,148 graduates and 3,598 drop outs are observed to work according to the definition of Mulligan & Rubinstein (2008). Additional

to education, the data include information on potential work experience, the marital status, and regional dummies, which serve as covariate vector X. Finally, the the number of children aged 0-6 interacted with the marital status are used as exclusion restrictions Z.

The estimation of the ATE and QTE on the working (denoted as $\Delta_{S=1}, \Delta^{\tau}_{S=1}$) based on IPW proceeds in three steps: (1) Obtain the predicted sample selection propensity score $\hat{p}(W)$ by regressing S on D, X, Z, (2) obtain the predicted treatment propensity score $\hat{\pi}(X, \hat{p}(W))$ by regressing D on X and $\hat{p}(W)$, (3) obtain the ATE and QTE estimates $\hat{\Delta}_{S=1}, \hat{\Delta}^{\tau}_{S=1}$ by plugging the treatment propensity score into the normalized sample analogues of the identification results in equations (6) and (8). E.g., the normalized ATE estimator is

$$\hat{\Delta}_{S=1} = \sum_{i|S=1}^{n} \frac{D_i \cdot Y_i}{\hat{\pi}(X_i, \hat{p}(W_i))} \Big/ \sum_{j|S=1}^{n} \frac{D_j}{\hat{\pi}(X_j, \hat{p}(W_j))} - \sum_{i|S=1}^{n} \frac{(1 - D_i) \cdot Y_i}{1 - \hat{\pi}(X_i, \hat{p}(W_i))} \Big/ \sum_{j|S=1}^{n} \frac{(1 - D_j)}{1 - \hat{\pi}(X_j, \hat{p}(W_j))},$$

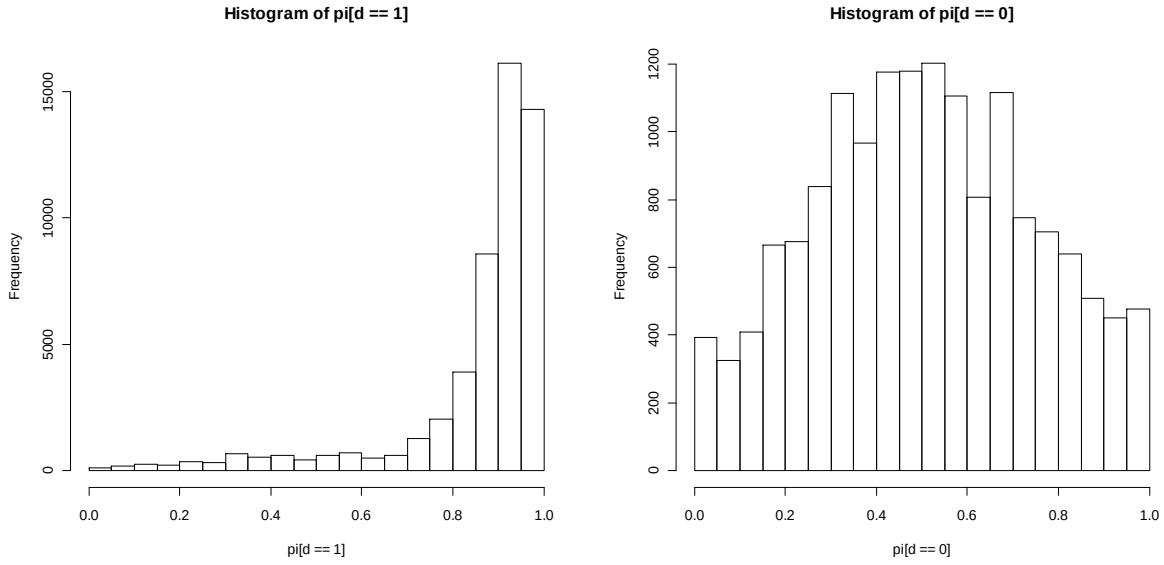
where $\hat{p}(W)$ and $\hat{\pi}(X, \hat{p}(W))$ are the propensity score estimates to work and to graduate from high school, respectively, and and $\hat{\Delta}$ denotes the ATE estimate. The normalizations $\sum_{j|S=1}^{n} \frac{D_j}{\hat{\pi}(X_j, \hat{p}(W_j))}$ and $\sum_{j|S=1}^{n} \frac{(1-D_j)}{1-\hat{\pi}(X_j, \hat{p}(W_j))}$ guarantee that the weights add up to unity, see the discussion in Imbens (2004) for further details.

We use probit models for the propensity scores which results in a semiparametric IPW estimator that is shown to be asymptotically normal in the appendix. Albeit parametric, the propensity score models are quite flexible, as we include higher order and interaction terms. $p(W)$ is a function of the treatment dummy (high school graduation), marital status and the number of children aged 0-6 along with interactions, number of kids aged 0-6 squared, a potential work experience cubic interacted with education dummies, and the regional dummies. $\pi(X, p(W))$ is a function of $p(W)$, the potential work experience cubic, marital status, and the region. In the sample selection equation, the coefficients on high school graduation, the marital status, the number of children, the region, and interaction terms between experience and high

school graduation are significant at the 5% level, and in the treatment equation, all coefficients, including the one on the predicted sample selection propensity score, are significant at the 1% level.

The common support or overlap in the treatment propensity score distributions of treated and non-treated units is of crucial importance. An insufficient overlap would point to a lack of appropriate comparisons across treatment groups. The histograms of $\hat{\pi}(X_i, \hat{p}(W_i))$ for $D = 1$ and $D = 0$ presented in Figure 1 reveal that the common support is quite satisfactory. In fact, both treatment groups contain observations over the entire theoretical support of the propensity score. As some propensity score estimates are close to the boundaries, we trim these values

Figure 1: Estimated treatment propensity scores for $D = 1$ and $D = 0$



to a maximum of 0.99 and a minimum of 0.01 to avoid arbitrarily large weights which might entail instability of the IPW estimator, see Khan & Tamer (2007). Furthermore, wage outliers are trimmed when estimating the ATE and left unchanged when estimating QTEs in the same manner as in Mulligan & Rubinstein (2008).

In addition to IPW, we estimate the ATE on the observed subpopulation using PSM. To be

specific, we use two-nearest-neighbor caliper matching (see Sekhon, 2007) where the caliper defines the maximally acceptable distance in any match's propensity score. This procedure eliminates those matches that are not comparable in terms of their treatment probabilities, i.e., lie outside the support. We set the caliper to 0.25 standard deviations of the estimated treatment propensity score, but due to the decent common support, no observations have to be dropped. After-matching balance tests indicate that balance is considerably increased, suggesting that treated and nontreated matches are comparable with respect to the distribution of the covariates and the estimated sample selection propensity score. We use 999 bootstrap replications to compute standard errors and p-values of the IPW estimators. PSM standard errors are estimated by the (within treatment group) matching-based variance estimator suggested by Abadie & Imbens (2006), which, however, does not account for uncertainty in the estimation of the propensity scores.

Table 1 provides the ATE estimates ($\hat{\Delta}_{S=1}$) and standard errors (s.e.) of the semiparametric IPW and PSM procedures. The highly significant effects suggest that graduating from high school increases the average hourly wage by 1.93 to 1.98 USD or roughly 18%. The estimate of the parametric two step heckit procedure (see Heckman, 1976), which is also provided in the table, is somewhat lower (1.63 USD).

To assess the credibility of our results, we compare them to the worst case bounds (see, Manski 1989 and 1994 for an introduction to partial identification) on $\Delta_{S=1}$ when neither controlling for

Table 1: ATE estimates (increase of hourly wage in USD)

	IPW	PSM	heckit		worst case bounds	bounds w. assumptions
$\hat{\Delta}_{S=1}$	1.980	1.934	1.626	identified set	[-29.040, 13.026]	[1.440, 2.061]
(s.e.)	0.173	0.450	0.115	(s.e.'s of bounds)	(1.815, 0.395)	(0.158, 0.695)
p-value	0.000	0.000	0.000	confidence regions	[-32.597, 13.800]	[1.130, 3.423]

Note: Standard errors (s.e.'s) of IPW and worst case bounds are based on 999 bootstrap replications.

S.e.'s of bounds with assumptions are based on 999 subsampling draws.

S.e.'s of PSM are computed using the Abadie & Imbens (2006) variance estimator.

S.e.'s of heckit is based on asymptotic theory.

sample selection nor treatment selection. Note that

$$\begin{aligned}
\Delta^{UB}_{S=1} &= E[Y|D=1,S=1]\cdot \Pr(D=1|S=1) + UB\cdot(1-\Pr(D=1|S=1)) \\
&- LB\cdot \Pr(D=1|S=1) - E[Y|D=0,S=1]\cdot(1-\Pr(D=1|S=1)) \\
&\geq \Delta_{S=1} \geq [Y|D=1,S=1]\cdot \Pr(D=1|S=1) + LB\cdot(1-\Pr(D=1|S=1)) \\
&- UB\cdot \Pr(D=1|S=1) - E[Y|D=0,S=1]\cdot(1-\Pr(D=1|S=1)) = \Delta^{LB}_{S=1},
\end{aligned}$$

where $\Delta^{UB}_{S=1}$ and $\Delta^{LB}_{S=1}$ denote the upper and lower bound of the ATE. UB and LB are the upper and lower bound of hourly wages, which are set to the maximum and minimum observed wages in the data, $\max(Y|S=1), \min(Y|S=1)$, respectively. For estimation, we simply take the sample analogues of $\Pr(D=1|S=1)$ and $E[Y|D=d]$. Not surprisingly, the estimated bounds are quite uninformative as the admissible ATEs range from -29.040 to 13.026. We bootstrap the lower and upper bound 999 times in order to estimate their standard errors and compute the confidence interval $[\hat{\Delta}^{LB}_{S=1} - 1.96\cdot\hat{\sigma}_{LB}, \hat{\Delta}^{UB}_{S=1} + 1.96\cdot\hat{\sigma}_{UB}]$, where $\hat{\Delta}^{LB}_{S=1}, \hat{\Delta}^{UB}_{S=1}, \hat{\sigma}_{LB}, \hat{\sigma}_{UB}$ are the estimates of the ATE bounds and their respective standard errors. This confidence interval covers the true ATE on the working with at least 0.95 probability.

To tighten the bounds we assume the CIA to hold conditional on $\pi(X) = \Pr(D=1|X)$, which

is implied by Assumption 1, and impose the uniformity assumption of Lechner & Melly (2007) and Lee (2009). I.e., $\Pr(S^1 \geq S^0) = 1$ (see also the last section), implying that everyone working without graduation would also work with graduation. Lee bounds the treatment effect for the always selected with $S^0 = 1, S^1 = 1$, which are those who work irrespective of education. Under the CIA and the uniformity assumption $E[Y|D = 0, \pi(X)]$ is equal to the expected potential outcome for the always selected under non-graduation, $E[Y^0|\pi(X), S^0 = 1, S^1 = 1]$. $E[Y|D = 1, \pi(X)]$ is a weighted average of outcomes of always selected and compliers, i.e., individuals who work with graduation but would not without graduation ($S^0 = 0, S^1 = 1$). I.e.,

$$\begin{aligned} E[Y|D=1,\pi(X)] &= E[Y|D=1,\pi(X),S^0=1,S^1=1]\cdot(1-c) \\ &+ E[Y|D=1,\pi(X),S^0=0,S^1=1]\cdot c, \\ &= E[Y^1|\pi(X),S^0=1,S^1=1]\cdot(1-c)+E[Y^1|\pi(X),S^0=0,S^1=1]\cdot c, \end{aligned}$$

where c denotes the probability to be a complier given the propensity score, $Pr(S^0 = 0, S^1 = 1|\pi(X))$.

Thus, the expected potential outcome for the always selected under graduation, $E[Y^1|\pi(X), S^0 = 0, S^1 = 1]$, can be bounded by taking the expectation of the upper or lower share of $Y|D = 1, S = 1, \pi(X)$ that corresponds to the probability to be an always selected, which is $1 - c = 1 - \frac{Pr(S=1|D=1,\pi(X))-Pr(S=1|D=0,\pi(X))}{Pr(S=1|D=1,\pi(X))}$, see Lee (2009) for further details. The upper and lower bounds on the ATE for the always selected, Δ_a, are identified by

$$\begin{aligned} \Delta_a^{UB} &= \int \{E[Y|D=1,S=1,\pi(X)=\pi(x),Y\geq Q_Y^c] \\ &- E[Y|D=0,S=1,\pi(X)=\pi(x)]\}dF_{\pi(X)|S=1}, \\ \Delta_a^{LB} &= \int \{E[Y|D=1,S=1,\pi(X)=\pi(x),Y\leq Q_Y^{1-c}] \\ &- E[Y|D=0,S=1,\pi(X)=\pi(x)]\}dF_{\pi(X)|S=1}, \end{aligned}$$

where Q_Y^τ denotes the τth quantile of Y.

As we want to estimate the bounds for the entire population with observed outcomes ($S = 1$), we also need to bound the counterfactual of $E[Y|D = 1, S = 1, \pi(X)]$ which is

$$E[Y^0|D = 1, S = 1, \pi(X)] = (1 - c) \cdot E[Y^0|\pi(X), S^0 = 1, S^1 = 1] + c \cdot E[Y^0|\pi(X), S^0 = 0, S^1 = 1].$$

Due to the uniformity assumption the counterfactual for the always selected, $E[Y^0|\pi(X), S^0 = 1, S^1 = 1]$, is $E[Y|D = 0, \pi(X), S^0 = 1, S^1 = 1] = E[Y|D = 0, S = 1\pi(X) = \pi(x)]$ and observed. However, $E[Y^0|\pi(X), S^0 = 0, S^1 = 1]$ is unknown as complier outcomes are not observed for $D = 0$. We define the upper bound of $E[Y^0|\pi(X), S^0 = 0, S^1 = 1]$ as $E[Y|D = 0, S = 1, \pi(X) = \pi(x)]$, assuming that observed compliers would on average not earn more without graduation than the always selected. The latter would be employed with and without graduation and are therefore likely to be more motivated and/or able than the compliers. Zhang et al. (2008) argue that ability tends to be positively correlated with wages and thus, this assumption appears to be plausible. Also Lechner & Melly (2007) assume positive selection with respect to wages.

We define the lower bound of $E[Y^0|\pi(X), S^0 = 0, S^1 = 1]$ as the minimum wage that is observed among working, $\min(Y|S = 1)$. Then, the upper and lower bounds of the ATE on the always selected and compliers, Δ_{ac}, are identified by

$$\begin{aligned}
\Delta_{ac}^{UB} &= \int \{E[Y|D = 1, S = 1, \pi(X) = \pi(x)] \\
&- E[Y|D = 0, S = 1, \pi(X) = \pi(x)] \cdot (1 - c) - \min(Y|S = 1) \cdot c\} dF_{\pi(X)|S=1}, \\
\Delta_{ac}^{LB} &= \int \{E[Y|D = 1, S = 1, \pi(X) = \pi(x)] \\
&- E[Y|D = 0, S = 1, \pi(X) = \pi(x)]\} dF_{\pi(X)|S=1}.
\end{aligned}$$

Finally, $\Delta_{S=1}$ is partially identified by

$$\Delta_{ac}^{UB} \cdot \Pr(D=1|S=1) + \Delta_{a}^{UB} \cdot (1 - \Pr(D=1|S=1)) \geq \Delta_{S=1}$$

$$\geq \Delta_{ac}^{LB} \cdot \Pr(D=1|S=1) + \Delta_{a}^{LB} \cdot (1 - \Pr(D=1|S=1)).$$

We estimate $\pi(X)$ using a probit model and denote the estimated propensity score as $\hat{\pi}(X)$. Also $Pr(S=1|D=d, \pi(X))$ is estimated by a probit regression of $S|D=d$ on $\hat{\pi}(X)$ and a constant. $E[Y|D=1, S=1, \pi(X)=\pi(x), Y \geq Q_Y^c]$ and $E[Y|D=1, S=1, \pi(X)=\pi(x), Y \leq Q_Y^{1-c}]$ are estimated by averaging over the predictions of linear quantile regressions of $Y|D=1$ on the polynomial $\sum_{p=0}^{3} \hat{\pi}(X)^p$ and $E[Y|D=d, S=1, \pi(X)=\pi(x)]$ by averaging over the predictions of a linear mean regression of $Y|D=d$ on $\sum_{p=0}^{3} \hat{\pi}(X)^p$. $\Delta_a^{UB}, \Delta_a^{LB}, \Delta_{ac}^{UB}, \Delta_{ac}^{LB}$ are estimated by matching on $\hat{\pi}(X)$. To compute the confidence intervals we draw 999 subsamples without replacement, see Politis, Romano & Wolf (1999), of subsample size 20,000. Under the CIA and the uniformity assumption the identified set is quite informative and positive. The ATE's lower bound is significantly different from zero. Notably, the IPW and PSM point estimates lie within the estimated bounds and therefore do not contradict the results obtained from partial identification.

Finally, Table 2 reports the QTE estimates based on IPW for the 0.1th to 0.9th quantile of potential wages. The effects vary importantly across different parts of the wage distribution. The results suggest that those with comparably large hourly wages benefit most while those with little wages benefit least from a high school graduation, given that the rank stability assumption holds.

Table 2: QTE estimates (increase of hourly wage in USD)

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\hat{\Delta}^{\tau}_{S=1}$	1.119	1.148	1.436	1.900	2.199	2.199	2.443	2.452	2.671
(s.e.)	0.225	0.226	0.213	0.212	0.231	0.218	0.225	0.216	0.229
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: Standard errors (s.e.'s) of IPW are based on 999 bootstrap replications.

4 Conclusion

This paper discusses the identification and estimation of average and quantile treatment effects in the presence of sample selection, attrition, and non-response related to unobservables. It considers the case of a double selection problem (i) into the subpopulation for which the outcome is observed (selection on unobservables) and (ii) into the treatment (selection on observables). The main contribution of the paper is nonparametric identification based on weighting observations by the inverse of a nested propensity score which controls for selection bias related to being observed and being assigned to the treatment. This approach requires a continuous instrument for sample selection which needs to be - just as the treatment - conditionally independent of the unobserved factors in the model. Estimators based on inverse probability weighting (IPW) naturally arise from the sample analogues of the identification results. Alternatively to IPW, propensity score matching (PSM) estimators on the nested propensity score may also be applied.

In contrast to most parametric and semiparametric models, the sample selection framework considered is of rather general form. It does not require a tight specification of the relation between the selection probability, the covariates, and the outcome and allows for effect heterogeneity with respect to the observed covariates and the sample selection propensity score. Therefore, the paper shows identification of average and quantile treatment effects for various target

populations, namely the selected subpopulation (whose outcomes are observed), the entire population (irrespective of selection), and the always selected (who are selected irrespective of the treatment).

We apply IPW and PSM to US labor market data previously analyzed by Mulligan & Rubinstein (2008) to determine the effect of high school graduation vs. no high school graduation on the wages of white females. The estimates suggest that graduation increases the hourly wage of working graduates and non-graduates on average by 18 % in the period considered (1975 to 1979). We also estimate worst case bounds and tighter bounds based on particular assumptions concerning the sample selection process but do not obtain contradictory results.

A Appendix

A.1 Proof of Proposition 1

Under Assumptions 1, 2, and 3, $\Delta_{S=1}$, the ATE on the subpopulation with observed outcomes, is identified by

$$\Delta_{S=1} \quad = \quad E\left[\frac{D\cdot Y}{\pi(X,p(W))} \mid S=1\right] - E\left[\frac{(1-D)\cdot Y}{1-\pi(X,p(W))} \mid S=1\right].$$

Proof:

$$\begin{aligned}
& E\left[\frac{D\cdot Y}{\pi(X,p(W))} \mid S=1\right] - E\left[\frac{(1-D)\cdot Y}{(1-\pi(X,p(W)))} \mid S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{D\cdot Y}{\pi(X,p(W))} - \frac{(1-D)\cdot Y}{(1-\pi(X,p(W)))}|X,p(W),S=1\right]|p(W),S=1\right] \mid S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{Y}{\pi(X,p(W))}|D=1,X,p(W),S=1\right]\cdot\pi(X,p(W))\right.\right. \\
- & \left.\left. E\left[\frac{Y}{(1-\pi(X,p(W)))}|D=0,X,p(W),S=1\right]\cdot(1-\pi(X,p(W)))|p(W),S=1\right] \mid S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y|D=1,X,p(W),S=1\right] - E\left[Y|D=0,X,p(W),S=1\right]|p(W),S=1\right] \mid S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y^1|X,p(W),S=1\right] - E\left[Y^0|X,p(W),S=1\right]|p(W),S=1\right] \mid S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[\Delta_{S=1}(X,p(W))|p(W),S=1\right] \mid S=1\right] = \Delta_{S=1}.
\end{aligned}$$

The first equality follows from the law of iterated expectations, the fourth from Assumptions 1 and 2. $\Delta_{S=1}(X,p(W))$ denotes the conditional ATE given X and $p(W)$ in the selected subpopulation. Finally, the last equality is a backward application of the law of iterated expectations.

A.2 Proof of Proposition 2

Under Assumptions 1, 2, and 3, $Q^{\tau}_{Y^1|S=1}$, the τth quantile of $Y^1|S=1$, is an implicit function of the following expression:

$$E\left[\frac{D}{\pi(X,p(W))}\cdot I\{Y\leq Q^{\tau}_{Y^1|S=1}\} \mid S=1\right] = F_{Y^1|S=1}(Q^{\tau}_{Y^1|S=1}) = \tau.$$

Proof:

$$
\begin{aligned}
& E\left[\frac{D}{\pi(X,p(W))}\cdot I\{Y\leq Q^{\tau}_{Y^1|S=1}\}\mid S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{D}{\pi(X,p(W))}\cdot I\{Y\leq Q^{\tau}_{Y^1|S=1}\}\mid X,p(W),S=1\right]\mid p(W),S=1\right]\mid S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{I\{Y\leq Q^{\tau}_{Y^1|S=1}\}}{\pi(X,p(W))}\mid D=1,X,p(W),S=1\right]\cdot\pi(X,p(W))\mid p(W),S=1\right]|S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[I\{Y\leq Q^{\tau}_{Y^1|S=1}\}\mid D=1,X,p(W),S=1\right]\mid p(W),S=1\right]|S=1\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[I\{Y^1\leq Q^{\tau}_{Y^1|S=1}\}\mid X,p(W),S=1\right]\mid p(W),S=1\right]|S=1\right] \\
= & E\left[I\{Y^1\leq Q^{\tau}_{Y^1|S=1}\}\mid S=1\right]=\tau.
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the fourth from Assumptions 1 and 2. The fifth equality is a backward application of the law of iterated expectations. An equivalent result holds for $Q^{\tau}_{Y^0|S=1}$. Therefore, the QTE $\Delta^{\tau}_{S=1}$ is identified by $Q^{\tau}_{Y^1|S=1}-Q^{\tau}_{Y^0|S=1}$.

A.3 Proof of Proposition 3

Under Assumptions 1, 2, 4, and 5, Δ, the ATE on the total population, is identified by

$$
\Delta = E\left[\frac{S\cdot D\cdot Y}{p(W)\cdot\pi(X,p(W))}\right]-E\left[\frac{S\cdot(1-D)\cdot Y}{p(W)\cdot(1-\pi(X,p(W)))}\right].
$$

Proof:

$$
\begin{aligned}
& E\left[\frac{S\cdot D\cdot Y}{p(W)\cdot\pi(X,p(W))}\right]-E\left[\frac{S\cdot(1-D)\cdot Y}{p(W)\cdot(1-\pi(X,p(W)))}\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{S\cdot D\cdot Y}{p(W)\cdot\pi(X,p(W))}-\frac{S\cdot(1-D)\cdot Y}{p(W)\cdot(1-\pi(X,p(W)))}|X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{D\cdot Y}{p(W)\cdot\pi(X,p(W))}-\frac{(1-D)\cdot Y}{p(W)\cdot(1-\pi(X,p(W)))}|S=1,X,p(W)\right]\cdot p(W)|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[\frac{D\cdot Y}{\pi(X,p(W))}-\frac{(1-D)\cdot Y}{(1-\pi(X,p(W)))}|S=1,X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\Bigg[\underset{X}{E}\Bigg[E\left[\frac{Y}{\pi(X,p(W))}|D=1,S=1,X,p(W)\right]\cdot\pi(X,p(W)) \\
- & E\left[\frac{Y}{(1-\pi(X,p(W)))}|D=0,S=1,X,p(W)\right]\cdot(1-\pi(X,p(W)))|p(W)\Bigg]\Bigg] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y|D=1,S=1,X,p(W)\right]-E\left[Y|D=0,S=1,X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[E\left[Y^1|S=1,X,p(W)\right]-E\left[Y^0|S=1,X,p(W)\right]|p(W)\right]\right] \\
= & \underset{p(W)}{E}\left[\underset{X}{E}\left[\Delta_{S=1}(X,p(W))|p(W)\right]\right]=\underset{p(W)}{E}\left[\underset{X}{E}\left[\Delta(X,p(W))|p(W)\right]\right]=\Delta,
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the sixth from Assumptions 1 and 2. The eighth equality follows from Assumption (2.c) by which $F_{U|D=d,X=x,p(W)=p(w),S=s} = F_{U|X=x,p(W)=p(w),S=s}$ and Assumption (5) which imposes additivity of observed and unobserved terms. Both together imply that $\Delta_{S=1}(X,p(W))$, the conditional ATE given X and $p(W)$ in the selected subpopulation, is equal to $\Delta_{S=0}(X,p(W))$ and thus, $\Delta(X,p(W))$. Finally, the last equality is a backward application of the law of iterated expectations.

A.4 Asymptotic distribution of the IPW estimator using parametric propensity score models

This section shows $\sqrt{n}$-consistency and asymptotic normality of IPW estimators using parametric models for the selection into the subpopulation with observed outcomes and into treatment. The properties are discussed in a GMM framework that is similar to the one considered by Lechner

(2009) for dynamic treatment evaluation.

It is assumed that the nested propensity scores p, π for sample selection and treatment assignment are known up to a finite number of coefficients. I.e., $\beta \equiv (\beta_s, \beta_d)$, where β_s denotes the coefficients on $W \equiv D, X, Z$ in $p = p(W, \beta_s)$ and β_d the coefficients on X, p in $\pi = \pi(X, p(W, \beta_s), \beta_d)$. Furthermore, there exists a $\sqrt{n}$ -consistent, asymptotically normal estimator $\hat{\beta}$, for instance a two step ML estimator of a nested probit or logit model with likelihood functions $L_s(s, \beta_s), L_d(d, \beta_d, \beta_s)$. Note that $\hat{\beta}_d$, the coefficient estimates characterizing the treatment propensity score, are a function of the sample selection propensity score which itself is a function of $\hat{\beta}_s$ (which is $\sqrt{n}$-consistent) rather than of the true value β_s.

Murphy & Topel (1985) show that under certain regularity conditions the two step ML estimator of $\hat{\beta}_d$ is $\sqrt{n}$-consistent and asymptotically normal.[1] Let k, g denote the score functions, i.e., the first derivatives of the likelihood functions with respect to the coefficients: $\begin{pmatrix} k(x,z,s,d,\beta_s) \\ g(x,z,s,d,\beta) \end{pmatrix} = \begin{pmatrix} \partial L_s(s,\beta_s)/\partial\beta_s \\ \partial L_d(d,\beta_d,\beta_s)/\partial\beta_d \end{pmatrix}$. Using a GMM framework, the estimators of the unknown values of β_d, β_s satisfy the conditions

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} k(X_i, Z_i, S_i, D_i; \hat{\beta}_s) &= 0. \\ \frac{1}{n}\sum_{i=1}^{n} g(X_i, Z_i, S_i, D_i; \hat{\beta}) &= 0. \end{aligned}$$

These conditions allow predicting the sample selection and treatment propensity scores and

[1] Murphy & Topel (1985) prove that

$$\begin{aligned} \sqrt{n}(\hat{\beta}_d - \beta_d) &\rightarrow N(0, \Sigma), \\ \Sigma &= R_2^{-1} + R_2^{-1}[R_3' R_1^{-1} R_3 - R_4' R_1^{-1} R_3 - R_3' R_1^{-1} R_4] R_2^{-1}, \\ R_1 &= -E\frac{\partial^2 L_s(s,\beta_s)}{\partial\beta_s \partial\beta_s'}, R_2 = -E\frac{\partial^2 L_d(d,\beta_d,\beta_s)}{\partial\beta_d \partial\beta_d'}, \\ R_3 &= -E\frac{\partial^2 L_d(d,\beta_d,\beta_s)}{\partial\beta_s \partial\beta_d'}, R_4 = E\frac{\partial L_s(s,\beta_s)}{\partial\beta_s}\left(\frac{\partial L_d(d,\beta_d,\beta_s)}{\partial\beta_d}\right)', \end{aligned}$$

where ‘′’ denotes transposed.

will serve as one part of the final GMM estimator that will also incorporate a moment condition related to the treatment effects. Now consider the ATE estimator

$$\hat{\Delta}_{S=1} = \frac{1}{\sum_{j=1}^{n} S_j} \cdot \sum_{i|S=1}^{n} \frac{D_i \cdot Y_i}{\pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)} - \frac{1}{\sum_{j=1}^{n} S_j} \cdot \sum_{i|S=1}^{n} \frac{(1 - D_i) \cdot Y_i}{1 - \pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)},$$

which is the sample analogue of equation (6). It is straightforward to rewrite the estimator as

$$\hat{\Delta}_{S=1} = \frac{1}{n} \sum_{i=1}^{n} \lambda_i(x, z, s, d, \hat{\beta}) \cdot Y_i,$$

with

$$\begin{aligned} \lambda_i(x, z, s, d, \hat{\beta}) &= n \cdot S_i \cdot \frac{1}{\sum_{j=1}^{n} S_j} \cdot \left(\frac{D_i}{\pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)} - \frac{1 - D_i}{1 - \pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)} \right) \\ &= \frac{S_i}{\hat{\chi}} \cdot \left(\frac{D_i = d}{\pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)} - \frac{1 - D_i}{1 - \pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)} \right), \end{aligned}$$

where $\hat{\chi}$ denotes the unconditional probability to be observed, $\hat{\chi} \equiv (\sum_{j=1}^{n} S_j)/n$. This allows us to formulate the estimator of $\Delta_{S=1}$ as the value $\hat{\Delta}_{S=1}$ satisfying

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i, Z_i, S_i, D_i; \hat{\beta}, \hat{\Delta}_{S=1}) = \frac{1}{n} \sum_{i=1}^{n} \lambda_i(x, z, s, d; \hat{\beta}) \cdot (Y_i - \hat{\Delta}_{S=1}) = 0,$$

which constitutes the second ingredient of the GMM estimator.

As in Lechner (2009), one particularity of this otherwise standard GMM problem (see Hansen, 1982, and Newey and McFadden, 1994) is that some of the moment conditions depend only on a subset of unknown parameters. I.e., the moment conditions g related to β do not depend on $\Delta_{S=1}$ and furthermore, $L_s(s, \beta_s)$ does not depend on β_d. The regularity conditions required for consistency and asymptotic normality in this framework of sequential estimators were established by Newey (1984): Data must be generated from stationary and ergodic processes, the moment functions and the respective derivatives must exist and must

be measurable and continuous, the parameters must be finite and not at the boundary of the parameter space, and the derivatives of the moment conditions w.r.t. the parameters must have full rank. Furthermore, the sample moments must converge to their population counterparts with decreasing variances and to uniquely identified values of the unknown parameters.

Applying the results of Newey (1984) and using the partitioned inverse formula on the matrix of derivatives (w.r.t. to the unknown parameters $\beta_s, \beta_d, \Delta_{S=1}$) of the moment conditions, the asymptotic variance of the ATE estimator is equal to

$$\begin{aligned}
\text{asVar}(\sqrt{n}\hat{\Delta}_{S=1}) &= H_{\Delta_{S=1}}^{-1} E[\{h(\cdot) + H_{\beta_d} G_{\beta_d}^{-1} g(\cdot) - (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s}) K_{\beta_s}^{-1} G_{\beta_d}^{-1} k(\cdot)\} \\
&\quad \times \{h(\cdot) + H_{\beta_d} G_{\beta_d}^{-1} g(\cdot) - (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s}) K_{\beta_s}^{-1} G_{\beta_d}^{-1} k(\cdot)\}'] H_{\Delta_{S=1}}^{-1} \\
&= V_{hh} + H_{\beta_d} G_{\beta_d}^{-1} V_{gh} - (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s}) K_{\beta_s}^{-1} G_{\beta_d}^{-1} V_{kh} \\
&\quad + H_{\beta_d} G_{\beta_d}^{-1} V_{gg} G_{\beta_d}^{-1\prime} H_{\beta_d}' + V_{hg} G_{\beta_d}^{-1\prime} H_{\beta_d}' - (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s}) K_{\beta_s}^{-1} G_{\beta_d}^{-1} V_{kg} G_{\beta_d}^{-1\prime} H_{\beta_d}' \\
&\quad + (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s}) K_{\beta_s}^{-1} G_{\beta_d}^{-1} V_{kk} G_{\beta_d}^{-1\prime} K_{\beta_s}^{-1\prime} (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s})' \\
&\quad - V_{hk} G_{\beta_d}^{-1\prime} K_{\beta_s}^{-1\prime} (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s})' - H_{\beta_d} G_{\beta_d}^{-1} V_{gk} G_{\beta_d}^{-1\prime} K_{\beta_s}^{-1\prime} (H_{\beta_s} G_{\beta_d} - H_{\beta_d} G_{\beta_s})',
\end{aligned}$$

where ‘′’ denotes transposed and

$$\begin{aligned}
H_{\Delta_{S=1}} &\equiv E\frac{\partial h(\cdot)}{\partial \Delta_{S=1}} = 1, H_{\beta_d} \equiv -E\frac{\partial h(\cdot)}{\partial \beta_d} = -E\left[\frac{\partial \lambda_i(\cdot)}{\partial \beta_d} Y_i\right], H_{\beta_s} \equiv -E\frac{\partial h(\cdot)}{\partial \beta_s} = -E\left[\frac{\partial \lambda_i(\cdot)}{\partial \beta_s} Y_i\right], \\
G_{\beta_d} &\equiv E\frac{\partial g(\cdot)}{\partial \beta_d}, G_{\beta_s} \equiv E\frac{\partial g(\cdot)}{\partial \beta_s}, K_{\beta_s} \equiv E\frac{\partial k(\cdot)}{\partial \beta_s}, \\
V_{hh} &\equiv E[h(\cdot)^2] = \text{Var}[\lambda_i(x, z, s, d, \hat{\beta}) \cdot Y_i], V_{gg} \equiv E[g(\cdot)g(\cdot)'], V_{kk} \equiv E[k(\cdot)k(\cdot)'], \\
V_{gh} &\equiv E[g(\cdot)h(\cdot)], V_{hg} \equiv V_{gh}', V_{kh} \equiv E[k(\cdot)h(\cdot)], V_{hk} \equiv V_{kh}', V_{kg} \equiv E[k(\cdot)g(\cdot)], V_{gk} \equiv V_{kg}'.
\end{aligned}$$

Ignoring the estimation of the nested propensity score would amount to assuming that $\frac{\partial \lambda_i(\cdot)}{\partial \beta_d} = 0$ and $\frac{\partial \lambda_i(\cdot)}{\partial \beta_s} = 0$ such that $\text{asVar}(\sqrt{n}\hat{\Delta}_{S=1})$ =$\text{Var}[\lambda_i(x, z, s, d, \hat{\beta}) \cdot Y_i]$. As acknowledged by Lechner (2009), the full variance might be smaller or larger than $\text{Var}[\lambda_i(x, z, s, d, \hat{\beta}) \cdot Y_i]$, depending on $\frac{\partial \lambda_i(\cdot)}{\partial \beta}$

and on the correlation of the moment conditions. The asymptotic variance can be consistently estimated by using the sample analogues of the terms in the formula or by bootstrapping.

We conclude this section by establishing a condition for the estimation of quantile functions required to estimate QTEs. To estimate $Q^{\tau}_{Y^1|S=1}$, note that the sample analogue of equation (8)is

$$\begin{aligned} &= \arg\min_y \frac{1}{\sum_{j=1}^n S_j} \cdot \sum_{i|S=1}^{n} \frac{D_i \cdot \rho_\tau(Y_i - y)}{\pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)}, \\ &= \arg\min_y \frac{1}{n} \sum_{i=1}^{n} \left[n \cdot S_i \cdot \frac{1}{\sum_{j=1}^n S_j} \cdot \frac{D_i \cdot \rho_\tau(Y_i - y)}{\pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)} \right]. \end{aligned}$$

This implies the first order condition

$$\frac{1}{n} \sum_{i=1}^{n} h_\tau(x_i, z_i, s_i, D_i = 1; \hat{\beta}, \hat{Q}^{\tau}_{Y^1|S=1}) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{S_i}{\hat{\chi}} \cdot \frac{D_i \cdot (I\{Y_i < \hat{Q}^{\tau}_{Y^1|S=1}\} - \tau)}{\pi(X_i, p(W_i, \hat{\beta}_s), \hat{\beta}_d)} \right] = 0,$$

which immediately serves as condition for GMM estimation. The asymptotic variance of the asymptotically normal estimator $\hat{Q}^{\tau}_{Y^1|S=1}$ can be obtained in a similar way as outlined for the ATE estimator and equivalent results hold for $\hat{Q}^{\tau}_{Y^0|S=1}$. As a consequence, the difference $\hat{\Delta}^{\tau}_{S=1} = \hat{Q}^{\tau}_{Y^1|S=1} - \hat{Q}^{\tau}_{Y^0|S=1}$ is asymptotically normal, too. $\hat{\Delta}^{\tau}_{S=1}$ involves independent terms, see for instance the argumentation in Firpo (2007). Therefore, the asymptotic variance of $\hat{\Delta}^{\tau}_{S=1}$ can be easily obtained from the asymptotic variances of $\hat{Q}^{\tau}_{Y^1|S=1}, \hat{Q}^{\tau}_{Y^0|S=1}$ as the covariance term is zero.

References

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

Ahn, H. & Powell, J. (1993), 'Semiparametric estimation of censored selection models with a nonparametric selection mechanism', *Journal of Econometrics* **58**, 3–29.

Andrews, D. & Schafgans, M. (1998), 'Semiparametric estimation of the intercept of a sample selection model', *Review of Economic Studies* **65**, 497–517.

Bang, H. & Robins, J. (2005), 'Doubly robust estimation in missing data and causal inference models', *Biometrics* **61**, 962–972.

Blundell, R. & Powell, J. (2003), Endogeneity in nonparametric and semiparametric regression models, *in* L. H. M. Dewatripont & S. Turnovsky, eds, 'Advances in Economics and Econometrics', Cambridge University Press, Cambridge, pp. 312–357.

Das, M., Newey, W. K. & Vella, F. (2003), 'Nonparametric estimation of sample selection models', *Review of Economic Studies* **70**, 33–58.

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**, 259–276.

Fitzgerald, J., Gottschalk, P. & Moffitt, R. (1998), 'An analysis of sample attrition in panel data: The michigan panel study of income dynamics', *The Journal of Human Resources* **33**, 251–299.

Frölich, M. & Melly, B. (2008), Unconditional quantile treatment effects under endogeneity.

Gronau, R. (1974), 'Wage comparisons-a selectivity bias', *Journal of Political Economy* **82**, 1119–1143.

Hansen, L. (1982), 'Large sample properties of generalized method of moment estimators', *Econometrica* **50**, 1029–1054.

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

Heckman, J. J. (1976), 'The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models', *Annals of Economic and Social Measurement* **5**, 475–492.

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

Heckman, J. J. (1990), 'Varieties of selection bias', *American Economic Review, Papers and Proceedings* **80**, 313–318.

Heckman, J. J., Ichimura, H. & Todd, P. (1997), 'Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme', *Review of Economic Studies* **64**, 605–654.

Hirano, K., Imbens, G. W. & Ridder, G. (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**, 1161–1189.

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

Imbens, G. W. (2000), 'The role of the propensity score in estimating dose-response functions', *Biometrika* **87**, 706–710.

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

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

Imbens, G. W. & Newey, W. (2003), 'Identification and estimation of triangular simultaneous equations models without additivity', *Econometrica* **77**, 1481–1512.

Khan, S. & Tamer, E. (2007), 'Irregular identification, support conditions, and inverse weight estimation', *unpublished manuscript* .

Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**, 33–50.

Lechner, M. (1999), 'Earnings and employment effects of continuous off-the-job training in east germany after unification', *Journal of Business and Economic Statistics* **17**, 74–90.

Lechner, M. (2001), Identification and estimation of causal effects of multiple treatments under the conditional independence assumption, *in* M. Lechner & F. Pfeiffer, eds, 'Econometric Evaluations of Active Labor Market Policies in Europe', Heidelberg: Physica.

Lechner, M. (2009), 'Sequential causal models for the evaluation of labor market programs', *Journal of Business and Economic Statistics* **27**, 71–83.

Lechner, M. & Melly, B. (2007), 'Earnings effects of training programs', *IZA Discussion Paper no. 2926* .

Lee, D. S. (2009), 'Training, wages, and sample selection: Estimating sharp bounds on treatment effects', *Review of Economic Studies* **76**, 1071–1102.

Manski, C. F. (1989), 'Anatomy of the selection problem', *The Journal of Human Resources* **24**, 343–360.

Manski, C. F. (1994), The selection problem, *in* C. Sims., ed., 'Advances in Econometrics: Sixth World Congress', Cambridge University Press, pp. 143–170.

Mulligan, C. B. & Rubinstein, Y. (2008), 'Selection, investment, and women's relative wages over time', *Quarterly Journal of Economics* **123**, 1061–1110.

Murphy, K. M. & Topel, R. H. (1985), 'Estimation and inference in two-step econometric models', *Journal of Business and Economic Statistics* **3**, 88–97.

Newey, W. K. (1984), 'A method of moments interpretation of sequential estimators', *Economics Letters* **14**, 201–206.

Newey, W. K. (2007), 'Nonparametric continuous/discrete choice models', *International Economic Review* **48**, 1429–1439.

Newey, W. K. (2009), 'Two-step series estimation of sample selection models', *Econometrics Journal* .

Newey, W. K. & McFadden, D. (1994), Large sample estimation and hypothesis testing, *in* R. Engle & D. McFadden, eds, 'Handbook of Econometrics', Elsevier, Amsterdam.

Newey, W., Powell, J. & Vella, F. (1999), 'Nonparametric estimation of triangular simultaneous equations models', *Econometrica* **67**, 565–603.

Politis, D. N., Romano, J. P. & Wolf, M. (1999), *Subsampling*, Springer-Verlag, New York.

Robins, J. & Rotnitzky, A. (1995), 'Semiparametric efficiency in multivariate regression models with missing data', *Journal of American Statistical Association* **90**, 122–129.

Robins, J., Rotnitzky, A. & Zhao, L. (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of American Statistical Association* **90**, 106–121.

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

Rotnitzky, A. & Robins, J. (1995), 'Semiparametric regression estimation in the presence of dependent censoring', *Biometrika* **82**, 805–820.

Rubin, D. B. (1973*a*), 'Matching to remove bias in observational studies', *Biometrics* **29**, 159–183.

Rubin, D. B. (1973*b*), 'The use of matched sampling and regression adjustment to remove bias in observational studies', *Biometrics* **29**, 185–203.

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

Rubin, D. B. (1976), 'Multivariate matching methods that are equal percent bias reducing, i: Some examples', *Biometrics* **32**, 109–120.

Rubin, D. B. (1990), 'Formal modes of statistical inference for causal effects', *Journal of Statistical Planning and Inference* **25**, 279–292.

Sekhon, J. S. (2007), 'Multivariate and propensity score matching software with automated balance optimization: The matching package for r', *forthcoming in the Journal of Statistical Software* .

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

Wooldridge, J. (2002), 'Inverse probability weigthed m-estimators for sample selection, attrition and stratification', *Portuguese Economic Journal* **1**, 141–162.

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

Zhang, J., Rubin, D. B. & Mealli, F. (2008), Evaluating the effects of job training programs on wages through principal stratification, *in* D. Millimet, J. Smith & E. Vytlacil, eds, 'Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics', Vol. 21, Elsevier Science Ltd., pp. 117–145.