Semiparametric Efficiency in Nonlinear LATE Models

HAN HONG and DENIS NEKIPELOV[1]

This version: June 2007

Key words: Semiparametric efficiency bound, local treatment effect, FTP, child achievement, unemployment benefits

Abstract

In this paper we study semiparametric efficiency for the estimation of a finite-dimensional parameter defined by generalized moment conditions under the local instrumental variable assumptions. These parameters identify treatment effects on the set of compliers under the monotonicity assumption. The distributions of covariates, treatment dummy and the binary instrument are not specified in a parametric form, making the model semiparametric. We derive the semiparametric efficiency bounds for both conditional models and unconditional models. We also develop multi-step semiparametric efficient estimators that achieve the semiparametric efficiency bound.

Using data from the Florida Family Transition Program, we apply the suggested estimation procedure to analyze the effect of the parent's employment on children's achievement for single-parent households who applied for state welfare support. We find that the linear regression estimate of the treatment effect has a substantial attenuation bias as compared to instrument-based methods. In general, parent's employment adversely affects child's achievement. Our result suggests that ignoring the selection effect indeed leads to substantial bias in the estimate of effect of parent's employment.

[1]Departments of Economics, Stanford University and Duke University, USA. The authors acknowledge generous research supports from the NSF. The usual disclaimer applies. The authors have made use of the FTP data without representing positions by the State of Florida and MDRC.

1 Introduction

Semiparametric efficiency is an important issue in the estimation of treatment effect models and models with endogenous regressors. For models with endogenous regressors, many papers in the literature including Chernozhukov and Hansen (2005) and Newey (1990a) among others have developed efficient estimators under conditional mean independence or quantile independence assumptions. Under the strong ignorability assumption, Hahn (1998a) and Hirano, Imbens, and Ridder (2003) derived the semiparametric efficiency bound and developed semiparametric efficient estimators for averaged treatment effect and the averaged treatment effect on the treated, while Firpo (2006) extended the analysis to quantile treatment effects.

An alternative approach to address the endogeneity problem is based on the local instrumental variable (LIV) method developed in a sequence of papers by Imbens and Angrist (1994), Angrist (2004), Angrist, Imbens, and Rubin (1996), Imbens and Rubin (1997) and Abadie, Angrist, and Imbens (2002). The baseline model for this method has a dummy endogenous regressor and a dummy instrument variable. Under the LIV assumption, the instrument variable weakly changes the endogenous regressor in one direction. Abadie, Angrist, and Imbens (2002) showed that the entire distributional causal effect is identified for the *complier* population where the endogenous regressor changes from zero to one as the instrumental variable changes from zero to one. The causal parameters identified by this method are fundamentally different from those identified by models with conditional independence assumptions. In contrast to the strong ignorability assumption, semiparametric efficiency under the LIV ssumption has not been subject to careful studies. An exception is Frolich (2006), who derives the efficiency bound for the average treatment effect of compliers and shows that the propensity score, properly defined in the LIV context, does not affect the efficiency bound.

We make several theoretical contributions in this paper. First we extend the parametric quantile treatment effect parameter model in Abadie, Angrist, and Imbens (2002) to treatment parameters that are defined by general nonlinear conditional moment conditions. Second, we fill an gap in the literature by deriving the semiparametric efficiency bound for these parameters and special them to quantile and linear treatment effect parameters.

To our knowledge we provide new semiparametric efficiency results for local treatment effect parameters. In addition, we also develop semiparametric estimators that achieve the theoretical efficiency bounds. Our semiparametric efficiency calculations include both conditional models and unconditional models, which characterize different treatment effect parameters. The unconditional efficiency bounds include as a special case the mean parameter of Frolich (2006), and also include the treatment effect on the treated compliers, which is related to the average treatment effect of the treated (ATT) when endogeneity is absent. The semiparametric efficiency bounds for the treatment effect of treated compliers are different when the propensity score is unknown, is known, or is correctly specified parametrically. We also simplify the structure of the efficiency analysis compared to the existing literature.

Efficient estimators are developed for both conditional and unconditional models. In the conditional case, we find that the estimator in Abadie, Angrist, and Imbens (2002) does not achieve the semiparametric efficiency bound, and we construct an alternative estimator which allows us to obtain the estimates with the asymptotic variance achieving the efficiency bound. The structure of the model allows us to make use of the binary instrument feature of a conditional moment model, and to reduce the problem of finding semiparametric efficiency bound in the moment-based framework as in Newey (1990b) and Bickel, Klaassen, Ritov, and Wellner (1993). The estimator naturally produces a GMM - based method which produces the estimator achieving the semiparametric efficiency bound. For unconditional models, we described efficient estimators for both the treatment effect of compliers and the treatment effect of treated compliers, for cases when the propensity score is unknown, known, and parametrically specified.

A natural application of our estimation procedure is on the relationship between the employment of a parent in single-parent households and the child achievement at school. We study this relation using data from the sample initially constructed for the analysis of Family Transition Program (FTP), an experimental alternative welfare program launched in the state of Florida in 1994. A major concern in estimation of the relationship between parental employment and child's achievement is the endogeneity of the binary variable of employment. Both the probability of employment of a parent and the children's achievement

should depend on the unobserved intellectual abilities of the parents. This dependence will bias the estimate of the effect of employment on the achievement of children. In particular, parents' employment should adversely affect the achievement of children because working parents will have less control over school attendance and preparation for classes of their children. However, the presence of unobserved ability (which is positively correlated with children's achievement) induces an upward bias in the estimate of the employment effect. This means that a naive estimate that doest not correct for endogeneity will understate the effect of employment.

The structure of the dataset that we use allows us to create a natural binary instrument to correct for endogeneity. The dataset that we use in this study was build from random assignment of people applying for welfare in Florida either to the conventional unemployment aid to families with children - Aid to Families with Dependent children (ADFC), or to the alternative Florida Transition Program (FTP). The latter program limited the amount of of time when unemployed can be supported by welfare, but also provided more intensive training to the participants. The data show that participation in the FTP significantly increases the probability of employment in the subsequent periods. For this reason, the program assignment dummy is a valid instrument for the employment variable in our model. In particular, unobserved individual ability and the program choice are independent due to the random nature of the program assignment, while the more intensive training in the FTP tends to increase employment. We maintain the complier assumption that participation in FTP weakly increase employment status. The set of of compliers in our study will be formed by the people for whom participation in the FTP strictly changes the employment dummy. For this subset of families we produce a consistent and efficient estimate of the effect of employment on children's achievement using our semiparametric methodology.

Section 2 develops the semiparametric efficiency of a complier treatment effect model that generalizes Abadie, Angrist, and Imbens (2002). Section 3 develops efficient estimation methods that achieve the semiparametric efficiency bound, and explicitly quantify the amount of efficiency improvement over existing methods. Section 4 discusses extensions to parameters that are defined unconditionally. Section 3.3 gives regularity assumptions that validate the proposed semiparametric efficient estimator. Section 5 applies the efficient

estimation to the FTP program. Finally section 6 concludes.

2 Semiparametric efficiency bound

2.1 Local treatment effect parameters

The local (complier) treatment effect model (see Imbens and Angrist (1994) and Abadie, Angrist, and Imbens (2002) for example) is defined as the following through a random vector $Y = (Y_1, Y_0)' \in \mathbb{R}^2$, a vector of binary variables $D = (D_1, D_0)' \in \{0, 1\} \times \{0, 1\}$, a binary instrument $Z \in \{0, 1\}$ and a vector of covariates $X \in \mathcal{X} \subset \mathbb{R}^k$. The following assumptions are used by these authors to describe the distributions of the variables under consideration:

Assumption 1 *Almost everywhere in $\mathcal{X}$:*

1. $(Y, D) \perp Z \,|\, X$,
2. $Pr(Z = 1 \mid X) \in (0, 1)$,
3. $E[D_1 \mid X] \neq E[D_0 \mid X]$,
4. $Pr(D_1 \geq D_0 \mid X) = 1$.

Under these four assumptions, in particular the last assumption (monotonicity), the data directly identifies the differences between the cohort that would have been treated for both values of the instrument (always takers) and the cohort that would not have been treated under any circumstances (never takers). The combination of the always taker cohort and the never taker cohort indirectly recovers the compliers which is the cohort that change behavior when the instrument changes.

The variables in the model, Y and D are not always completely observable. Only the following transformed variables are observed:

$$\begin{cases} w_1 = g_1(y, d, z, x) = y_1 w_2 + y_0 (1 - w_2) \\ w_2 = g_2(y, d, z, x) = d_0 + z(d_1 - d_0). \end{cases}$$

Due to assumption A1.1, the conditional probabilities of the observable binary variable w_2 can be written as:

$$\mathcal{P}_0(x) = P\left(w_2 = 1 \mid Z = 0,\, X\right) = E\left[D_0 \mid X\right],$$
$$\mathcal{P}_1(x) = P\left(w_2 = 1 \mid Z = 1,\, X\right) = E\left[D_1 \mid X\right].$$

where the second equalities follow from the conditional independence assumption 1.1.

Also define $\mathcal{Q}(x) = E\left[Z \mid X\right]$. Consequently, the conditional probability of the binary treatment w_2 given the instrument in terms of the probabilities of treatment dummies can be expressed as

$$P\left(w_2 = 1 \mid z,\, x\right) = \mathcal{F}(z, x) = \mathcal{P}_1(x)z + \mathcal{P}_0(x)\left(1 - z\right).$$

Taking expectation over z given x gives the conditional probability of w_2 given only x:

$$\mathbf{P}(x) = \mathcal{P}_1(x)\mathcal{Q}(x) + \mathcal{P}_0(x)\left(1 - \mathcal{Q}(x)\right).$$

The object of interest that can be identified under assumption 1 are the distributions of the outcomes Y_1 and Y_0 given $D_1 > D_0$ (implying that $D_1 = 1$ and $D_0 = 0$): for $j = 0, 1,\ f\left(y_j | d_1 > d_0, x\right)$. The subpopulation for which $D_1 > D_0$ is usually referred to as "compliers", for whom random selection into treatment affects the treatment dummy monotonically. Under monotonicity assumption 1 the distributions of compliers can be expressed in terms of the observed conditional distributions:

$$\begin{aligned} &f_{**}\left(w_1 \mid x,\, w_2 = 1\right) = f\left(y_1 = w_1 \mid d_1 > d_0,\, x\right) \\ &= \tfrac{\mathcal{P}_1(x)}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} f\left(w_1 \mid w_2 = 1, z = 1, x\right) - \tfrac{\mathcal{P}_0(x)}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} f\left(w_1 \mid w_2 = 1, z = 0, x\right). \end{aligned} \tag{1}$$

To see this relation, note that under the monotonicity assumption, $\mathcal{P}_0(x)$ is the proportion of always takers ($D_0 = D_1 = 1$) conditional on x while $\mathcal{P}_1(x)$ is the sum of always takers and compliers. $f\left(y_1 \mid w_2 = 1, z = 1, x\right)$ gives the distribution of y_1 conditional on being either an always taker or a complier and the covariate x. $f\left(y_1 \mid w_2 = 1, z = 0, x\right)$ gives the distribution of y_1 conditional on being just an always taker and x. Therefore, $\mathcal{P}_1\left(x\right) f\left(y_1 \mid w_2 = 1, z = 1, x\right)$ can be written as a linear combination for the known distribution of always-takers and the unknown distribution for compliers. Similarly, one can write the joint distribution of y_0 and the event of being either a never taker or a complier,

which is $(1-\mathcal{P}_0(x))\, f\,(y_0 \mid w_2 = 0, z = 0, x)$, as a linear combination of the distributions for never-takers and compliers. This results in:

$$\begin{aligned} & f_{**}\,(w_1 \mid x,\, w_2 = 0) = f\,(y_0 = w_1 \mid d_1 > d_0, x) \\ & = \tfrac{1-\mathcal{P}_0(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)} f\,(w_1 \mid w_2 = 0, z = 0, x) - \tfrac{1-\mathcal{P}_1(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)} f\,(w_1 \mid w_2 = 0, z = 1, x)\,. \end{aligned} \tag{2}$$

We consider a generalization of the linear quantile regression model of Abadie, Angrist, and Imbens (2002) to a parameter vector β determined by a conditional moment equation, for $\forall x$ and $\forall w_2$:

$$\varphi\,(\beta,\, x,\, w_2) = E\,[g\,(w_1,\, w_2, x, \beta) \mid x,\, w_2,\, d_1 > d_0] = 0, \tag{3}$$

for some parametric function $g\,(\cdot)$. The conditional expectation for $w_2 = 1, 0$ is taken with respect to the corresponding conditional density $f_{**}\,(w_1 \mid x,\, w_2)$.

Two direct applications of this general definition are the mean treatment effect of Imbens and Angrist (1994) and the quantile treatment effect of Abadie, Angrist, and Imbens (2002). The mean treatment effect model corresponds to a moment condition:

$$g\,(w_1, w_2, x, \beta) = w_1 - \beta_1 w_2 - (1 - w_2)\,\beta_0 - \beta_2' x.$$

The quantile treatment effect model characterizes the difference in conditional distributions of potential outcomes y_1 and y_0 for compliers through a linear specification of the conditional quantile functions:

$$Q_\tau\,(w_1 | x, w_2, d_1 > d_0) = \beta_0 w_2 + \beta_1' x.$$

The corresponding moment function that defines the QTE parameter is therefore:

$$g\,(w_1, w_2, x, \beta) = 1\left(w_1 \leq \beta_0 w_2 + \beta_1' x\right) - \tau.$$

These models can be extended to allow for a semiparametric component in the conditional moment function. For $\mu\,(x)$ being a nonparametric function of x, we may consider estimating $\mu\,(x)$ and β simultaneously in the moment function:

$$g\,(w_1, w_2, x, \mu\,(x)\,, \beta)\,.$$

For example, the parametric mean treatment effect model can be generalized to a semiparametric partial linear model:

$$g\left(w_1, w_2, x, \mu\left(x\right), \beta\right) = w_1 - \beta_1 w_2 - \left(1 - w_2\right)\beta_0 - \mu\left(x\right).$$

In the rest of the paper we derive semiparametric efficiency bounds for the parameter vector β and develop a semiparametric procedure that achieves the efficiency bound. This framework can be extended to derive the semiparametric efficiency bound for a nonparametric component in the specification of the conditional moment equations.

2.2 Semiparametric efficiency bound for treatment effect parameters

We will use the arguments of Newey (1990a) and Severini and Tripathi (2001) to construct the efficiency bounds for the system of conditional moments. More specifically, given a set of instrument functions of the covariates x, the conditional moments are first transformed into a system of unconditional moments. Then choosing the instrument functions optimally will produce the semiparametric efficiency bound of the conditional moment model.

Theorem 1 *Under Assumption 1, the semiparametric efficiency bound for a k-dimensional parameter β characterizing the subsample of compliers in (3) can be expressed as:*

$$V\left(\beta\right) = E\left(\left(\mathcal{P}_1(x) - \mathcal{P}_0(x)\right)^2 E\left[\frac{\partial\varphi(w_2, x, \beta)}{\partial\beta}\zeta(x, w_2)'\middle| x\right]\overline{\Omega}(x)^{-1}E\left[\zeta(x, w_2)\frac{\partial\varphi(w_2, x, \beta')}{\partial\beta}\middle| x\right]\right)^{-1}.$$

Denote $\omega_{w_2,z}(x) = V\left(g\left(w, x, \beta\right) \mid w_2, z, x\right)$ and $\gamma_{w_2,z}(x) = E\left(g\left(w, x, \beta\right) \mid w_2, z, x\right)$. We can then express the elements of the matrix $\overline{\Omega}(x)$ in the following way:

$$\overline{\Omega}_{11}(x) = \left(\frac{\mathcal{P}_1(x)\omega_{11}(x)}{\mathcal{Q}(x)} + \frac{\mathcal{P}_0(x)\omega_{10(x)}}{1-\mathcal{Q}(x)} + \frac{\gamma_{11}^2(x)\mathcal{P}_1(x)\mathbf{P}(x)}{\mathcal{P}_0(x)\mathcal{Q}(x)(1-\mathcal{Q}(x))}\left[1 - \frac{\mathcal{P}_1(x)\mathcal{P}_0(x)}{\mathbf{P}(x)}\right]\right)$$

$$\overline{\Omega}_{22}(x) = \left(\frac{(1-\mathcal{P}_1(x))\omega_{01}(x)}{\mathcal{Q}(x)} + \frac{(1-\mathcal{P}_0(x))\omega_{00}(x)}{1-\mathcal{Q}(x)} + \frac{\gamma_{00}^2(x)(1-\mathcal{P}_0(x))(1-\mathbf{P}(x))}{\mathcal{Q}(x)(1-\mathcal{Q}(x))(1-\mathcal{P}_1(x))}\left[1 - \frac{(1-\mathcal{P}_0(x))(1-\mathcal{P}_1(x))}{1-\mathbf{P}(x)}\right]\right)$$

and

$$\begin{aligned}\overline{\Omega}_{21}(x) = \overline{\Omega}_{12}(x) &= \left(\frac{\mathcal{P}_1(x)(1-\mathcal{P}_0(x))}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\gamma_{11}(x)\gamma_{00}(x)\right) \\ &= L_{12}\end{aligned}.$$

In this theorem we have used the notation:

$$\zeta\left(w_2, x\right) = \left(\frac{w_2}{\mathbf{P}(x)},\ \frac{1-w_2}{1-\mathbf{P}(x)}\right)'.$$

This structure of the variance bound shows several visible features. First of all, the semiparametric efficiency bound will grow if the fraction of compliers $\mathcal{P}_1\left(x\right) - \mathcal{P}_0\left(x\right)$ in the sample decreases. Moreover, the efficiency bound will be higher if the binary instrument is taking one of the values most of the time, in which case $\mathcal{Q}\left(x\right)$ is closer to 0 or 1. In addition, the proof and the estimation section show that the structural of the variance reflects the optimal instrument function as $\mathcal{M}\left(x\right)\zeta\left(x, w_2\right)$ where the weight function is defined as

$$\mathcal{M}(x) = E\left[\frac{\partial\varphi(w_2,x,\beta)}{\partial\beta}\zeta(x,w_2)'\middle| x\right]\overline{\Omega}(x)^{-1}\mathrm{diag}\left\{\frac{\mathbf{P}(x)}{\mathcal{Q}(x)},\ \frac{1-\mathbf{P}(x)}{1-\mathcal{Q}(x)}\right\}.$$

3 Efficient estimation

In this section we describe an estimator that achieves the semiparametric efficiency bound that makes use of the knowledge of the efficiency variance bound and the efficient score function of the model. The connection betweeen the efficient estimators and the structure of the efficient influence function is exploited in Bickel, Klaassen, Ritov, and Wellner (1993) and Murphy and van der Vaart (1997). In particular, the linear quantile treatment effect estimator of Abadie, Angrist, and Imbens (2002) has a limiting variance that is strictly larger than the semiparametric variance bound.

3.1 Efficiency improvement over existing methods

We have seen that the parameters of the treated and non-treated distributions form a conditional moment equation:

$$\int g\left(w_1,\ w_2,\ x,\ \beta\right) f\left(w_1 \mid w_2, d_1 > d_0,\ x\right) = 0.$$

The idea of the estimator is closely related to the identification argument. First of all, any given set of instrument functions, denoted

$$A\left(w_2, x\right) = \mathcal{M}\left(x\right)\zeta\left(x, w_2\right)$$

and

$$\mathcal{A}(w_2, x) = (\mathcal{P}_1(x) - \mathcal{P}_0(x)) \left(\frac{\mathcal{Q}(x) w_2}{\mathbf{P}(x)} + \frac{(1 - \mathcal{Q}(x))(1 - w_2)}{1 - \mathbf{P}(x)} \right) A(w_2, x), \tag{4}$$

can be used to transform the conditional moment equations (3) into unconditional ones:

$$E[E[\mathcal{A}(x, w_2) g(w_1, w_2, x, \beta) \mid d_1 > d_0, x, w_2]] = 0,$$

where the outer expectation is taken with respect to the marginal distribution of w_2 and x. For a given $A(x, w_2)$, we conjecture the form of the efficient estimator from the identification arguments. It is then shown that efficiency bound is achieved when the optimal $A(x, w_2)$ is estimated consistently.

The identification condition in (2) translates into the following implications on the conditional moment functions, for $\tilde{g} = A(x, w_{w_2}) g(w, x, \beta)$ and $\hat{g} = \mathcal{A}(x, w_{w_2}) g(w, x, \beta)$:

$$\begin{aligned}
&E(\hat{g} \mid w_2, d_1 > d_0, x) \\
&= \tfrac{\mathcal{P}_1(x) w_2}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} E(\hat{g} \mid w_2 = 1, z = 1, x) - \tfrac{\mathcal{P}_0(x) w_2}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} E(\hat{g} \mid w_2 = 1, z = 0, x) \\
&+ \tfrac{(1 - \mathcal{P}_0(x))(1 - w_2)}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} E(\hat{g} \mid w_2 = 0, z = 0, x) - \tfrac{(1 - \mathcal{P}_1(x))(1 - w_2)}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} E(\hat{g} \mid w_2 = 0, z = 1, x).
\end{aligned} \tag{5}$$

Using (4) this can be expressed in terms of $\tilde{g}$:

$$\begin{aligned}
&E(\tilde{g} \mid w_2, d_1 > d_0, x) \\
&= \tfrac{\mathcal{P}_1(x) \mathcal{Q}(x) w_2}{\mathbf{P}(x)} E(\hat{g} \mid w_2 = 1, z = 1, x) - \tfrac{\mathcal{P}_0(x) \mathcal{Q}(x) w_2}{\mathbf{P}(x)} E(\hat{g} \mid w_2 = 1, z = 0, x) \\
&+ \tfrac{(1 - \mathcal{P}_0(x))(1 - \mathcal{Q}(x)(1 - w_2))}{1 - \mathbf{P}(x)} E(\hat{g} \mid w_2 = 0, z = 0, x) \\
&- \tfrac{(1 - \mathcal{P}_1(x))(1 - \mathcal{Q}(x))(1 - w_2)}{1 - \mathbf{P}(x)} E(\hat{g} \mid w_2 = 0, z = 1, x).
\end{aligned} \tag{6}$$

Using the Bayes rule and the law of iterated expectation, we can further write

$$\begin{aligned}
&E[E(\tilde{g} \mid w_2, d_1 > d_0, x)] \\
&\quad = E\left\{ \left(w_2 z - \tfrac{\mathcal{Q}(x)}{1 - \mathcal{Q}(x)} w_2 (1 - z) + (1 - w_2)(1 - z) - \tfrac{(1 - \mathcal{Q}(x))}{\mathcal{Q}(x)} (1 - w_2) z \right) \hat{g} \right\}.
\end{aligned}$$

The conditional probability in this moment condition, $\mathcal{Q}(x)$, is not observed. However, it can be consistently estimated in a first step, using for example either a kernel regression or

a sieve based estimator. This estimate, $\hat{\mathcal{Q}}(x)$, can then be used to form a sample analog of the above moment conditions given any estimated instrument function $\hat{A}(x, w_2)$, or $\hat{\mathcal{M}}(x)$:

$$\begin{aligned}&\tfrac{1}{N}\textstyle\sum_{k=1}^{N}\hat{\psi}_k(\beta) = \tfrac{1}{N}\textstyle\sum_{k=1}^{N}\psi_k\left(\beta, \hat{\mathcal{Q}}, \hat{\mathcal{M}}\right)\\ &\quad = \frac{1}{N}\sum_{k=1}^{N}\left\{w_{2k}z - \frac{\hat{\mathcal{Q}}(x_k)w_{2k}(1-z_k)}{1-\hat{\mathcal{Q}}(x_k)} + (1-z_k)(1-w_{2k}) - \frac{\left(1-\hat{\mathcal{Q}}(x_k)\right)z_k(1-w_{2k})}{\hat{\mathcal{Q}}(x_k)}\right\}\tilde{g}_k.\end{aligned}$$

The proof of theorem 3 shows that this estimator achieves the efficiency bound under suitable regularity conditions, and the following asymptotic representation holds:

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\widehat{\psi}_k(\beta) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\mathcal{M}(x_k)\left\{\chi_k(\beta) + E\left[\frac{\partial\chi_k(\beta)}{\partial Q}\middle| x_k\right](z_k - \mathcal{Q}(x_k))\right\} + o_p(1).$$

In the above we have used the following notation to isolate the instrument matrix:

$$\psi_k\left(\beta, \hat{\mathcal{Q}}\right) = \mathcal{M}(x_k)\chi_k\left(\beta, \hat{\mathcal{Q}}\right). \tag{7}$$

We now compare the efficient variance to the variance of the estimator obtained using the approach in Abadie, Angrist, and Imbens (2002). To describe their estimator we start with a distance function $\rho(\cdot)$ whose first order condition can produce a moment condition that is implied by the conditional moment model. Define their weight function:

$$\kappa(w_2, z, x) = 1 - \frac{w_2(1-z)}{1-\mathcal{Q}(x)} - \frac{(1-w_2)z}{\mathcal{Q}(x)}.$$

For some consistent estimate of the probability $\mathcal{Q}(x)$ the estimator for β will solve:

$$\widetilde{\beta} = argmin\left\{\frac{1}{N}\sum_{k=1}^{N}\widehat{\kappa}(w_{2k}, z_k, x_k)\rho(w_k, x_k, \beta)\right\}$$

This optimization usually leads to a moment equation in the form:

$$\psi(\beta) = \left(1 - \frac{w_2(1-z)}{1-\mathcal{Q}(x)} - \frac{(1-w_2)z}{\mathcal{Q}(x)}\right)h(w_2, x, \beta)\,g(w, x, \beta).$$

In the above $h(w_2, x, \beta)$ is an instrument function that can also depend on β. In estimation we replace the functions under consideration with their empirical analogs. In this case

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\widehat{\psi}_k(\beta) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\left\{\psi_k(\beta) + E\left[\frac{\partial\psi_k(\beta)}{\partial Q}\middle| x_k\right](z_k - \mathcal{Q}(x_k))\right\} + o_p(1),$$

where $\widehat{\psi}$ is similar to ψ with $\mathcal{Q}(x)$ replaced by $\hat{\mathcal{Q}}(x)$. Note that we can write

$$E\left[\frac{\partial\psi_k(\beta)}{\partial Q}|x\right] = \widetilde{\theta}(x)'\tilde{D}(x)^{-1}E\left[\frac{\partial\chi_k(\beta)}{\partial Q}|x\right],$$

where χ is defined implicitly in (7), and

$$\tilde{D}(x) = \text{diag}\left\{\mathbf{P}(x),\ 1-\mathbf{P}(x)\right\}, \qquad \tilde{\theta}(x) = \begin{pmatrix} h(1,x,\beta_0)' \\ h(0,x,\beta_0)' \end{pmatrix}.$$

To compute asymptotic variance associated with the empirical moment equation note that:

$$V\left(E\left[\frac{\partial\psi_k(\beta)}{\partial Q}\middle| x_k\right](z_k - \mathcal{Q}(x_k))\right) = V\left\{\widetilde{\theta}(x_k)'\tilde{D}(x_k)^{-1}E\left[\frac{\partial\chi_k(\beta)}{\partial Q}|x_k\right](z_k - \mathcal{Q}(x_k))\right\}.$$

Moreover:

$$cov\left(\psi_k(\beta),\ E\left[\frac{\partial\psi_k(\beta)}{\partial Q}\middle| x_k\right](z_k - \mathcal{Q}(x_k))\right) = -V\left(E\left[\frac{\partial\psi_k(\beta)}{\partial Q}\middle| x_k\right](z_k - \mathcal{Q}(x_k))\right).$$

Finally:

$$V(\psi_k(\beta)) = E\left\{\widetilde{\theta}(x)'\tilde{D}(x)^{-1}V(\chi_k(\beta)\mid x)\tilde{D}(x)^{-1}\widetilde{\theta}(x)\right\},$$

which suggests that

$$V\left(\widehat{\psi}_k(\beta)\right) = V\left(\widetilde{\theta}(x)\tilde{D}(x)^{-1}\widehat{\chi}_k(\beta)\right)$$

We can express the Jacobi matrix for this model as:

$$J = E\left\{\widetilde{\theta}(x)'\tilde{D}(x)^{-1}\theta(x)\right\}.$$

This gives the following expression for the asymptotic variance:

$$V\left(\widetilde{\beta}\right) = E\left\{\tilde{\theta}(x)'\tilde{D}(x)^{-1}\theta(x)\right\}^{-1}E\left\{\tilde{\theta}(x)'\tilde{D}(x)^{-1}\overline{\Omega}(x)\tilde{D}(x)^{-1}\tilde{\theta}(x)\right\}$$
$$\times E\left\{\tilde{\theta}(x)'\tilde{D}(x)^{-1}\theta(x)\right\}^{-1}.$$

Next we note that

$$V\left(\widehat{\beta}\right)^{-1} - V\left(\widetilde{\beta}\right)^{-1}$$

can be written as the variance-covariance of the residual vector of the set of regression where the dependent variables are $\bar{\Omega}(x)^{-1/2}\theta(x)$ and the regressors are

$$\tilde{\theta}(x)'\tilde{D}(x)^{-1}\bar{\Omega}(x)^{1/2}.$$

Therefore, This result implies that $V\left(\widetilde{\beta}\right) - V\left(\widehat{\beta}\right)$ is a positive semi-definite matrix and thus the variance in Abadie, Angrist, and Imbens (2002) is smaller than that for the efficient estimator.

3.2 Efficient propensity score weighting estimator

To summarize, the following multi-step procedure leads to a semiparametric efficient estimator under suitable regularity conditions:

In the first stage, we first use a kernel based or sieve based nonparametric estimator to obtain estimates $\widehat{\mathcal{P}}_1(x)$, $\widehat{\mathcal{P}}_0(x)$, $\widehat{\mathcal{Q}}(x)$ of the conditional probabilities $\mathcal{P}_1(x)$, $\mathcal{P}_0(x)$ and $\mathcal{Q}(x)$.

In step two, using an initial choice of an instrument matrix $\tilde{A}(x, w_2)$ of dimension $d_\beta \times d_g$, construct an initial estimate $\bar{\beta}$ that equates

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\Bigg\{w_{2k}z - \frac{\hat{\mathcal{Q}}(x_k)w_{2k}(1-z_k)}{\left(1-\hat{\mathcal{Q}}(x_k)\right)} + (1-z_k)(1-w_{2k}) \\ - \frac{\left(1-\hat{\mathcal{Q}}(x_k)\right)z_k(1-w_{2k})}{\hat{\mathcal{Q}}(x_k)}\Bigg\}\tilde{A}(x, w_2)\, g\left(w_k, x_k, \bar{\beta}\right) = 0.\end{aligned} \tag{8}$$

In step three, $\bar{\beta}$ is used to estimate the optimal instrument nonparametrically. For this purpose we need to estimate

$$\hat{\omega}_{w_2,z}\left(x, \bar{\beta}\right) = \widehat{V}\left(g\left(w,\, x,\, \bar{\beta}\right)|w_2, z, x\right)$$

and $\hat{\gamma}_{w_2,z}\left(x, \bar{\beta}\right) = \widehat{E}\left(g\left(w,\, x,\, \bar{\beta}\right)|w_2, z, x\right)$ for $w_2 = 0, 1$ and $z = 0, 1$. Then an estimate of $\bar{\Omega}(x)$ and $\mathcal{M}(x)$ can be analytically computed, as

$$\hat{\mathcal{M}}(x) = \begin{pmatrix} \frac{\partial\varphi\left(1,x,\bar{\beta}\right)}{\partial\beta'} \\ \frac{\partial\varphi\left(0,x,\bar{\beta}\right)}{\partial\beta'} \end{pmatrix}' \widehat{\bar{\Omega}}\left(x;\bar{\beta}\right)^{-1}\mathrm{diag}\left\{\frac{\hat{\mathbf{P}}(x)}{\hat{\mathcal{Q}}(x)},\ \frac{1-\hat{\mathbf{P}}(x)}{1-\hat{\mathcal{Q}}(x)}\right\}.$$

Finally, the efficient $\hat{\beta}$ is obtained through a similar sample moment condition as the one that leads to $\bar{\beta}$, except that we replace $\tilde{A}(x, w_2)$ by $\hat{\mathcal{M}}(x)\zeta(w_2, x)$. The particular form of $\varphi(w_2, x, \beta)$ is model specific. For example, for quantile treatment effect parameters:

$$\frac{\partial\varphi(w_2, x, \beta)}{\partial\beta} = f_{**}\left(w_2\beta_0 + x'\beta_1 | w_2, x\right)\begin{pmatrix} w_2 \\ x \end{pmatrix}.$$

Section 3.3 formally provides regularity condition for the asymptotic distribution.

3.3 Regularity conditions and asymptotic distribution

In this section we state a set of sufficient regularity conditions for the semiparametric efficient estimator. We will focus mainly on the reweighting estimator described in previous section 3.2. Similar conditions can be given for the conditional expectation projection estimator described in the next section 3.4 and for the unconditional parameters estimators described in section 4. We follow much of the recent literature and describe regularity conditions in terms of sieve nonparametric estimators for conditional probabilities and conditional expectations. Most of these conditions are well understood in the recent literature (e.g. Ai and Chen (2003), Chen, Linton, and Van Keilegom (2003) and Newey (1994)). Therefore we only highlight the essential elements.

Let $\{q_l(X), l = 1, 2, ...\}$ denote a sequence of known basis functions that can approximate any square-measurable function of X arbitrarily well. Also let

$$q^{k(n)}(X) = \left(q_1(X), ..., q_{k(n)}(X)\right)' \quad \text{and}$$
$$Q = \left(q^{k(n)}(X_1), ..., q^{k(n)}(X_n)\right)'$$

for some integer $k(n)$, with $k(n) \to \infty$ and $k(n)/n \to 0$ when $n \to \infty$. A first stage nonparametric estimator for $\mathcal{Q}(x)$ is then defined as

$$\hat{\mathcal{Q}}(X) = \sum_{j=1}^{n} Z_j q^{k(n)}(X_j)\left(Q'Q\right)^{-1} q^{k(n)}(X).$$

An estimator of the instrument function $\hat{\mathcal{M}}(x)$ depends on a preliminary parameter estimate $\tilde{\beta}$ and nonparametric estimates of the quantities that define $\bar{\Omega}(x)$, which include $\hat{\omega}_{kl}\left(x, \tilde{\beta}\right)$,

$k, l = 0, 1$, $\hat{\gamma}_{kl}\left(x, \tilde{\beta}\right)$, $k, l = 0, 1$, $\hat{\mathcal{P}}_k(x)$, $k = 0, 1$, $\hat{\mathbf{P}}(x)$ and the Jacobi matrix term

$$\hat{E}\left[\frac{\partial \hat{\varphi}(w_2, x, \tilde{\beta})}{\partial \beta}\zeta(x, w_2)'\middle| x\right],$$

which can be nonparametrically estimated by

$$\hat{E}\left[\frac{\partial \hat{\varphi}(w_2, x, \tilde{\beta})}{\partial \beta}\zeta(x, w_2)'\middle| x\right] = \sum_{j=1}^{n} \hat{\mathcal{W}}_j\left(\tilde{\beta}\right) q^{k(n)}(X_j)\left(Q'Q\right)^{-1} q^{k(n)}(x)$$

where

$$\hat{\mathcal{W}}_j\left(\tilde{\beta}\right) = \frac{g\left(w_j, x_j, \tilde{\beta} + h\right) - g\left(w_j, x_j, \tilde{\beta} - h\right)}{2h}\zeta(x_j, w_{2j}).$$

To state the regularity conditions, we make use of the definitions of the weighted sup norm metric $||h||_{\infty,\omega}$ and the from Chen, Hong, and Tarozzi (2005). Let $\hat{\mathcal{Q}}_0(x)$ denote the true $\mathcal{Q}(x)$, and similarly for other estimated quantities. The first set of assumptions concerns the consistent of the parameter estimate. The assumption that $\mathcal{Q}_0(x)$ is bounded away from 0 and 1 is convenient but strong.

Assumption 2 *The following conditions hold:*

1. $E\left[E\left[\mathcal{A}(x, w_2)\, g(w_1,\, w_2, x, \beta) \mid d_1 > d_0, x, w_2\right]\right] = 0$ *if and only if* $\beta = \beta_0$.
2. $\mathcal{Q}_0() \in \mathcal{H} = \{\mathcal{Q}() : 0 < \underline{q} \leq \mathcal{Q}(x) \leq \overline{q} < 1\}$ *for some* $\gamma > 0$;
3. $||\hat{\mathcal{Q}}(\cdot) - \mathcal{Q}_0(\cdot)||_{\infty,\omega} \xrightarrow{p} 0$, $||\hat{A}(\cdot, k) - A_0(\cdot, k)||_{\infty,\omega} \xrightarrow{p} 0$, *for* $k = 0, 1$;
4. $E_a\left[\sup_{\beta \in B} ||g(w_i, X_i; \beta)||^2 \left(1 + ||X_i||^2\right)^{\omega}\right] < \infty$;
5. *there is a non-increasing function* $b(\cdot)$ *such that* $b(\delta) \to 0$ *as* $\delta \to 0$ *and*

$$E_a\left[\sup_{||\beta - \widetilde{\beta}|| < \delta} ||g(w_i, X_i; \beta) - g(w_i, X_i, \widetilde{\beta})||^2\right] \leq b(\delta).$$

for all small positive value δ.

Theorem 2 *Under assumption 2,* $\hat{\beta} - \beta_0 = o_p(1)$.

The next two sets of assumptions pertains to asymptotic normality and efficiency of $\hat{\beta}$.

Assumption 3 *The following conditions hold:*

1. *There exist a constant* $\epsilon \in (0,1]$ *and a small* $\delta_0 > 0$ *such that*

$$E_a\left[\sup_{||\beta-\widetilde{\beta}||<\delta} ||m(Z_i;\beta) - m(Z_i,\widetilde{\beta})||^2\right] \leq const.\delta^{\epsilon}$$

   for any small positive value $\delta \leq \delta_0$*;*

2. *The class of nonparametric functions* $\mathcal{Q}(\cdot)$ *and* $A(\cdot,\cdot)$ *is managable in the sense of condition 3.3 of Theorem 3 of Chen, Linton, and Van Keilegom (2003).*

3. $||\hat{\mathcal{Q}}(\cdot) - \mathcal{Q}_0(\cdot)||_{\infty,\omega} = o_p\left(n^{-1/4}\right)$, $||\hat{A}(\cdot,k) - A_0(\cdot,k)||_{\infty,\omega} = o_p\left(n^{-1/4}\right)$, *for* $k = 0,1$*;*

4. $E_a\left[\sup_{\beta\in B} ||g(w_i,X_i;\beta)|| \left(1+||X_i||^2\right)^{2\omega}\right] < \infty$;

5. $E_a\left[\sup_{|\beta-\beta_0|\leq\delta} ||\frac{\partial E[g(w_i,X_i;\beta)|w_{2i},x_i]}{\partial\beta}||^2 \left(1+||X_i||^2\right)^{\omega}\right] < \infty$;

The next assumption makes sure that the linear approximation of the sample moment condition (8) between the estimated $\hat{\mathcal{Q}}(x)$ and true $\mathcal{Q}(x)$ is asymptotically normal. For this purpose define

$$\begin{aligned}\delta_0(x) = &-\frac{1}{1-\mathcal{Q}_0(x)} E\left(w_2 A(x,w_2)\, g(w,x,\beta_0)\,|z=0,x\right)\\ &+\frac{1}{\mathcal{Q}_0(x)} E\left((1-w_2) A(x,w_2)\, g(w,x,\beta_0)\,|z=1,x\right).\end{aligned}$$

Also define $\delta_{k(n)}(X)$ and $\mathcal{Q}_{k(n)}(x)$ to be the projections of $\delta_0(X)$ and $\mathcal{Q}_0(X)$ onto the linear space spanned by $q^{k(n)}(X)$. For example:

$$\mathcal{Q}_{k(n)}(x) = q^{k(n)}(X)\left(Eq^{k(n)}(X)\, q^{k(n)}(X)'\right)^{-1} Eq^{k(n)}(X)\,\mathcal{Q}_0(X)'.$$

Assumption 4 *The following conditions hold:*

$nE\left[||\delta_0(X) - \delta_{k(n)}(X)||^2\right] \cdot E\left[||\mathcal{Q}_0(X) - \mathcal{Q}_{k(n)}(x)||^2\right] \longrightarrow 0.$

$E\left[||\delta_{k(n)}(X)\left(\mathcal{Q}_0(X) - \mathcal{Q}_{k(n)}(x)\right)||^2\right] \longrightarrow 0.$

$E\delta_0(X)\, q^{k(n)}(X)'\left((Q'Q/n)^{-1} - (EQ'Q/n)^{-1}\right)\sum_{i=1}^{n} q^{k(n)}(X_i)\left(Z_i - \mathcal{Q}_{k(n)}(X_i)\right)/n = o_p(1)$

Theorem 3 *Under assumptions 2 and 3 and 4 the obtained M-estimates are consistent, asymptotically normal and achieve the variance lower bound. In other words:*

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)\xrightarrow{d}N\left(0\,,\,V\left(\beta\right)\right).$$

for $V(\beta)$ given in theorem 1.

Finally, we need to give a set of primitive conditions for condition 3 in assumption 3 regarding the estimation of the instrument function $A\left(\cdot,k;\tilde{\beta}\right)$, $k=0,1$, where the dependence of $A(\cdot)$ on the initial estimate $\tilde{\beta}$ is explicitly noted. Of course, in the first stage initial estimation of $\tilde{\beta}$, such assumption is not needed.

Assumption 5 *In addition to everything in assumptions 2 and 3 except the second part of condition 3 of both assumptions, assume the following conditions hold.*

1. $0<\underline{p}\leq\mathcal{P}_0\left(x\right)\leq\overline{p}<1$. $0<\underline{p}\leq\mathcal{P}_1\left(x\right)\leq\overline{p}<1$.
2. $||\hat{\mathcal{P}}_k\left(\cdot\right)-\mathcal{P}_k^0\left(\cdot\right)||_{\infty,\omega}=o_p\left(n^{-1/4}\right)$, *for* $k=0,1$.
3. $\sup_{|\beta-\beta_0|\leq\delta_n}||\hat{\omega}_{jk}\left(\cdot,\beta\right)-\omega_{jk}^0\left(\cdot,\beta\right)||_{\infty,\omega}=o_p\left(n^{-1/4}\right)$, *for* $j,k=0,1$.
4. $\sup_{|\beta-\beta_0|\leq\delta_n}||\hat{\gamma}_{jk}\left(\cdot,\beta\right)-\gamma_{jk}^0\left(\cdot,\beta\right)||_{\infty,\omega}=o_p\left(n^{-1/4}\right)$, *for* $j,k=0,1$.
5. $\varphi\left(\beta,\,x,\,w_2\right)$ *is twice continuously differentiable in* β *uniformly over* x *and* w_2.
6. $\frac{1}{\sqrt{n}h}+h^2=o\left(n^{-1/4}\right)$.

The last condition, in particular, requires that $h\to0$ at a rate that is slower than $n^{-1/4}$ but faster than $n^{-1/8}$. h^2 characterizes the bias of the two sided numerical derivative under condition 5.

Proposition 1 *Assumption 5 implies condition 3 of assumption 3.*

The proofs of the theorems and propositions in this section are in the appendix and follow immediately from the assumptions.

3.4 Conditional expectation projection estimator

The estimation method described in the previous sections 3.1 and 3.2 is based on a sample average of the properly reweighted moment conditions, where the weights are related to the conditional probabilities $\mathcal{Q}(x)$, $\mathcal{P}_1(x)$ and $\mathcal{P}_0(x)$, all of which needs to be estimated nonparametrically. Borrowing from the terminology from treatment effect estimation under the unconfoundedness (i.e. strong ignorability) assumption, we will call this the inverse propensity score weighting estimator. In fact, in the exogenous case when $\mathcal{P}_1(x) = 1$ and $\mathcal{P}_0(x) = 0$, this is identical to the inverse probability weighting estimator for strongly ignorable conditional treatment effect models.

There exists an alternative estimator that relies on direct estimation of the conditional expectation $E[g(w, x, \beta) | w_2, x, D_1 > D_0]$ for each candidate parameter β instead of on reweighting the moment conditions using the inverse of $\hat{\mathcal{Q}}(x)$. To describe this estimator, begin with rewriting the identification condition (5) as

$$\begin{aligned}
&E(\tilde{g} \mid w_2, d_1 > d_0,\, x) \\
&= \tfrac{w_2}{\mathcal{P}_1(x)-\mathcal{P}_0(x)} E(w_2\tilde{g} \mid z = 1, x) - \tfrac{w_2}{\mathcal{P}_1(x)-\mathcal{P}_0(x)} E(w_2\tilde{g} \mid z = 0, x) \\
&+ \tfrac{(1-w_2)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)} E((1-w_2)\,\tilde{g} \mid z = 0, x) - \tfrac{(1-\mathcal{P}_1(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)} E((1-w_2)\,\tilde{g} \mid z = 1, x).
\end{aligned}$$

For a given instrument matrix $\mathcal{M}(x)$, this suggests estimating β by equating to zero the following sample analog:

$$\begin{aligned}
\tfrac{1}{N}\sum_{k=1}^{N} \phi_k(\beta) = \tfrac{1}{N}\sum_{k=1}^{N} \Bigg\{ &\tfrac{w_{2k}}{\hat{\mathcal{P}}_1(x_k)-\hat{\mathcal{P}}_0(x_k)} \widehat{E}(w_{2k}\tilde{g} \mid z = 1, x_k) \\
&- \tfrac{w_{2k}}{\hat{\mathcal{P}}_1(x_k)-\hat{\mathcal{P}}_0(x_k)} \widehat{E}(w_{2k}\tilde{g} \mid z = 0, x_k) + \tfrac{1-w_{2k}}{\hat{\mathcal{P}}_1(x_k)-\hat{\mathcal{P}}_0(x_k)} \widehat{E}((1-w_{2k})\,\tilde{g} \mid, z = 0, x_k) \\
&- \tfrac{1-w_{2k}}{\hat{\mathcal{P}}_1(x_k)-\hat{\mathcal{P}}_0(x_k)} \widehat{E}((1-w_{2k})\,\tilde{g} \mid z = 1, x_k) \Bigg\},
\end{aligned}$$

where each of the conditional expectation terms above are estimated nonparametrically at every given parameter value β. For example,

$$\widehat{E}(w_{2k}\tilde{g} \mid z = 1, x_k) = \frac{\widehat{E}(w_{2k} z \tilde{g} \mid x_k)}{\widehat{Q}(x_k)}.$$

Both conditional expectations can be estimated using a variety of nonparametric regression methods such as sieve expansion or kernel smoothing.

It is easy to show that the asymptotic linear influence function corresponding to the moment condition $\frac{1}{N}\sum_{k=1}^{N}\phi_k(\beta)$ for a given $\mathcal{M}(x)$ including the optimal one coincides with the semiparametric efficient function. First of all, similar to before, estimating $\mathcal{P}_1(x) - \mathcal{P}_0(x)$ has no impact on the asymptotic variance due to the conditional nature of the moment restrictions. Using the representation theorem of Newey (1994), we can for example expand the first component as:

$$\begin{aligned}
&\tfrac{1}{\sqrt{N}}\sum_{k=1}^{N}\tfrac{w_{2k}}{\hat{\mathcal{P}}_1(x_k)-\hat{\mathcal{P}}_0(x_k)}\tfrac{\widehat{E}(w_{2k}z_k\tilde{g}|x_k)}{\widehat{\mathcal{Q}}(x_k)} \\
&= \tfrac{1}{\sqrt{N}}\sum_{k=1}^{N}\left\{\tfrac{\mathbf{P}(x_k)w_{2k}z_k\tilde{g}}{(\mathcal{P}_1(x_k)-\mathcal{P}_1(x_k))\mathcal{Q}(x_k)} - \tfrac{\mathbf{P}(x_k)\mathcal{P}_1(x_k)E(\tilde{g}|w_{2k}=1,z_k=1,x_k)}{(\mathcal{P}_1(x_k)-\mathcal{P}_1(x_k))\mathcal{Q}(x_k)}\left(z_k-\mathcal{Q}(x_k)\right)\right. \\
&\qquad\left. + \tfrac{w_{2k}-\mathbf{P}(x)}{(\mathcal{P}_1(x)-\mathcal{P}_0(x))}E\left(w_{2k}\tilde{g}\mid z_k=1,x_k\right)\right\} + o_p(1).
\end{aligned}$$

Similar calculations can be applied to the other three terms.

When summing these four components, we note that the last terms in each of the components cancel out due to the implications of the conditional moment restrictions that

$$E\left(w_{2k}\tilde{g}\mid z_k=1,x_k\right) = E\left(w_{2k}\tilde{g}\mid z_k=0,x_k\right)$$

and

$$E\left(\left(1-w_{2k}\right)\tilde{g}\mid z_k=1,x_k\right) = E\left(\left(1-w_{2k}\right)\tilde{g}\mid z_k=0,x_k\right).$$

Therefore, it is easy to check that the sum of the four influence function is identical to the semiparametric efficient influence function when the instrument is chosen optimally. For the sake of brevity we omit the regularity conditions for the conditional expectation projection estimator.

4 Unconditional parameters

Often times researches can be mainly interested in parameters that are defined unconditionally. For example, under the unconfoundedness assumption where the latent outcome

is conditionally independent of the treatment status given exogenous covariates, the semiparametric efficiency literature has focused on the average treatment effect and the average treatment effect on the treated, both are defined unconditionally with repect to the exogenous covariates X.

Under the unconfoundedness assumption, one can also specify a model where the average treatment effect or effect on the treated conditional on each covariate is constant or a known parametric function of the covariates, similar to what is analyzed in the previous section and in Abadie, Angrist, and Imbens (2002). However, most of the literature has focused on analyzing the average treatment effect or effect on the treated without requiring that this effect is a constant conditional on every value of the exogenous covariate.

When X is not a constant, the conditional model and the unconditional model imply very different parameters of interest. For example, the semiparametric efficiency bound for an average treatment effect that is assumed to be constant across all possible covariate X is tighter than that for the average treatment effect defined unconditionally with respect to the covariates X. This section investigates efficient estimator for unconditionally defined treatment effect parameters under the LIV monotonicity assumption

4.1 Semiparametric efficiency of mean treatment effects

This section will restrict attention to mean effect parameters in order to illustrate the ideas. However, the results are readily extensible to general moment conditions in section 4.3. Specifically, we consider the average treatment effect on compliers (ATEC):

$$\beta \equiv \beta_1 - \beta_0 = E\left(Y_1 - Y_0 | D_1 > D_0\right),$$

and the average treatment effect on the treated compliers (ATTC):

$$\gamma \equiv \gamma_1 - \gamma_0 = E\left(Y_1 - Y_0 | w_2 = 1, D_1 > D_0\right).$$

These parameters reduce to the usual notation of average treatment effect (ATE) and the effect on the treated (ATT) under strong ignorability when $P\left(D_1 > D_0\right) = 1$. The efficiency bound for ATEC has been derived by Frolich (2006) although we develop a simplified derivation. Our results for ATTC are new and are applicable when the propensity score

$\mathcal{Q}(x)$ is unknown, known or parametrically specified. The first theorem considers unknown propensity scores.

Theorem 4 *The semiparametric efficient bound for β is given by the variance of the following efficient influence function:*

$$\frac{1}{P(D_1 > D_0)}\left\{ \frac{z}{\mathcal{Q}(x)}(w_1 - E(w_1|z=1,x)) + E(w_1|z=1,x) \right.$$
$$- \frac{1-z}{1-\mathcal{Q}(x)}(w_1 - E(w_1|z=0,x)) - E(w_1|z=0,x)$$
$$- \left(\frac{z}{\mathcal{Q}(x)}(w_2 - E(w_2|z=1,x)) + E(w_2|z=1,x) \right.$$
$$\left. \left. - \frac{1-z}{1-\mathcal{Q}(x)}(w_2 - E(w_2|z=0,x)) - E(w_2|z=0,x) \right) \beta \right\}$$

while the semiparametric efficiency bound for γ is given by the variance of the following efficient influence function:

$$\frac{1}{P(w_2 = 1, D_1 > D_0)}\left\{ w_1 - \frac{1-z}{1-\mathcal{Q}(x)}(w_1 - E(w_1|z=0,x)) - E(w_1|z=0,x) \right.$$
$$\left. - \left(w_2 - \frac{1-z}{1-\mathcal{Q}(x)}(w_2 - E(w_2|z=0,x)) - E(w_2|z=0,x) \right) \gamma \right\}.$$

Obviously, under strong ignorability when $z = w_2, P(D_1 > D_0) = 1$, both of these reduce to the corresponding influence functions derived in Hahn (1998a). In fact the only difference (other than the factor outside the bracket) is in the coefficient in front of β and γ, which become 1 and z under strong ignorability.

The literature has also been concerned with the semiparametric efficiency when the so-called propensity score, in our case $\mathcal{Q}(x)$, is either known or parametrically specified. We will still leave $\mathcal{P}_1(x) - \mathcal{P}_0(x)$ nonparametrically specified, even though cases when this is known or parametrically specified can be analyzed too.

From the proof of Theorem 4, it is clear that $\mathcal{Q}(x)$ does not even enter the definition of the moment conditions (13) and (15) that define β. Consequently, any knowledge of $\mathcal{Q}(x)$ will have no impact on the efficiency bound for β.

Such knowledge, however, will improve on the efficiency bound for γ, as described in the following theorem.

Theorem 5 *When the propensity score $\mathcal{Q}(x;\alpha)$ is correctly specified up to a finite dimensional parameter α, the semiparametric efficiency bound for γ is the variance of the following efficient influence function*

$$\begin{aligned}\frac{1}{P(w_2=1,D_1>D_0)}\Big\{ & z(w_1-E(w_1|z_1=1,x))+\mathcal{Q}(x)E(w_1|z_1=1,x) \\ & -\frac{1-z}{1-\mathcal{Q}(x)}Q(x)[w_1-E(w_1|z=0,x)]-\mathcal{Q}(x)E(w_1|z=0,x) \\ & -\Big\{ z(w_2-E(w_2|z_1=1,x))+\mathcal{Q}(x)E(w_2|z_1=1,x) \\ & -\frac{1-z}{1-\mathcal{Q}(x)}Q(x)[w_2-E(w_2|z=0,x)]-\mathcal{Q}(x)E(w_2|z=0,x)\Big\}\gamma \\ & +Proj[(z-\mathcal{Q}(x))\kappa(x)\mid S_\alpha(z;x)]\Big\}.\end{aligned}$$

In the above we have used the definition that

$$\begin{aligned}\kappa(x)=&E(w_1|z=1,x)-E(w_1|z=0,x) \\ &-(E(w_2=1|z=1,x)-E(w_2=1|z=0,x))\gamma,\end{aligned}$$

and the efficient influence function of the parametric propensity score model:

$$S_\alpha(z;x)=\frac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\frac{\partial\mathcal{Q}}{\partial\alpha}(x,\alpha).$$

In the above Proj denotes the linear projection operator:

$$\text{Proj}[(z-\mathcal{Q}(x))\kappa(x)\mid S_\alpha(z;x)]=S_\alpha(z;x)Var(S_\alpha(z;x))^{-1}Cov(S_\alpha(z;x),\kappa(x)).$$

In fact, Theorem 4 can be considered a special case of this influence function when

$$\text{Proj}[(z-\mathcal{Q}(x))\kappa(x)\mid S_\alpha(z;x)]$$

is replaced by just $(z-\mathcal{Q}(x))\kappa(x)$. In another special case, the efficient influence function for a model where the propensity score $\mathcal{Q}(x)$ is known is the same as in theorem 5, except that the last term $\text{Proj}[(z-\mathcal{Q}(x))\kappa(x)\mid S_\alpha(z;x)]$ is replaced by 0.

4.2 Efficient estimation of unconditional parameters

It is easy to show that an efficient estimator can be derived from the principle of conditional expectation projection that follows the identification condition. Consider first the case of the average treatment effect on compliers (ATEC) $\beta = E[Y_1 - Y_0 \mid D_1 > D_0]$. Combining equations for the means of the distributions of treated and non-treated observations for compliers, we obtain the following unconditional moment equation:

$$E\left\{\beta\left(\mathcal{P}_1(x) - \mathcal{P}_0(x)\right) - \left(E[w_1 \mid z = 1, x] - E[w_1 \mid z = 0, x]\right)\right\} = 0.$$

Efficient semiparametric estimator is obtained from the sample analog of this moment equation and takes the form:

$$\widehat{\beta} = \frac{\frac{1}{N}\sum_{k=1}^{N}\left(\widehat{E}[w_{1k} \mid z_k = 1, x_k] - \widehat{E}[w_{1k} \mid z_k = 0, x_k]\right)}{\frac{1}{N}\sum_{k=1}^{N}\left(\widehat{\mathcal{P}}_1(x_k) - \widehat{\mathcal{P}}_0(x_k)\right)}.$$

Conditional expectations in this expression can be estimated nonparametrically by kernel or sieve-based mathod. Semiparametric efficiency of this estimator can be established by the same projection arguments that we used before to establish efficiency of the estimator for the conditional moment-based model.

Similarly to the ATEC we can estimate the the average treatment effect for the treated (ATTC) as $\gamma = E[Y_1 - Y_0 \mid w_2 = 1, D_1 > D_0]$. The ATTC can be written in terms of the unconditional moment equation:

$$E\left\{\left[\gamma\left(\mathcal{P}_1(x) - \mathcal{P}_0(x)\right) - \left(E[w_1 \mid z = 1, x] - E[w_1 \mid z = 0, x]\right)\right]\mathcal{Q}(x)\right\} = 0.$$

By the same principle as the ATEC we express the efficient estimator as an empirical analog:

$$\widehat{\gamma} = \frac{\frac{1}{N}\sum_{k=1}^{N}\hat{\mathcal{Q}}(x_k)\left(\widehat{E}[w_{1k} \mid z_k = 1, x_k] - \widehat{E}[w_{1k} \mid z_k = 0, x_k]\right)}{\frac{1}{N}\sum_{k=1}^{N}\hat{\mathcal{Q}}(x_k)\left(\widehat{\mathcal{P}}_1(x_k) - \widehat{\mathcal{P}}_0(x_k)\right)}.$$

Using projection argument of Newey (1994) we can easily verify that this estimator achieves the semiparametric efficiency bound, when each of the conditional expectations and conditional probabilities above are estimated nonparametrically using either kernel of sieve based methods.

If the $\mathcal{Q}(x)$ is specified as a parametric function or is a known function, $\mathcal{Q}_\alpha(x)$, then the efficient estimator for γ becomes

$$\widehat{\gamma} = \frac{\frac{1}{N}\sum_{k=1}^{N} \mathcal{Q}_{\hat{\alpha}}(x_k)\left(\widehat{E}\left[w_{1k} \mid z_k = 1, x_k\right] - \widehat{E}\left[w_{1k} \mid z_k = 0, x_k\right]\right)}{\frac{1}{N}\sum_{k=1}^{N} \hat{\mathcal{Q}}_{\hat{\alpha}}(x_k)\left(\widehat{\mathcal{P}}_1(x_k) - \widehat{\mathcal{P}}_0(x_k)\right)}.$$

where $\hat{\alpha}$ is the parametric MLE, or the known α_0 if $\mathcal{Q}(x)$ is fully known.

4.3 General separable unconditional model for compliers

The treatment effect models considered in this section have a straightforward generalization to the separable conditional moment restrictions expressed in terms of unobservable outcome variables. Consider a problem where a finite-dimensional parameter $\beta \in \mathbb{R}^k$ is given by the following unconditional moment equation described in terms of unobservable variables Y_1 and Y_0:

$$\varphi(\beta) = E\left[g_1(Y_1, x, \beta) - g_0(Y_0, x, \beta) \,\middle|\, D_1 > D_0\right] = 0. \tag{9}$$

In particular, when $g_1(Y_1, \beta) = Y_1 - \beta$ and $g_0(Y_0, \beta) = Y_0 + \beta$ parameter β defines the average treatment effect for compliers. On the other hand, $g_1(Y_1, \beta) = 1(Y_1 \leq \beta_1) - \tau$ and $g_0(Y_0, \beta) = 1(Y_0 \leq \beta_0) + \tau$ define a complier analog of the average quantile treatment effect parameter proposed in Firpo (2006).

Note that we can represent this moment equation for compliers in terms of distributions for the entire population. Using the Bayes's rule we find that this equation is equivalent to:

$$E\Bigg[\left(\mathcal{P}_1(x) - \mathcal{P}_0(x)\right)\Bigg(\mathcal{Q}(x)E\left[g_1(Y_1, x, \beta) \mid w_2 = 1, D_1 > D_0, x\right] \\ - \left(1 - \mathcal{Q}(x)\right)E\left[g_0(Y_0, x, \beta) \mid w_2 = 0, D_1 > D_0, x\right]\Bigg)\Bigg] = 0,$$

which can be redefined in terms of only observable variables in the form:

$$E\Bigg[\left(\mathcal{P}_1(x) - \mathcal{P}_0(x)\right)\left(\frac{\mathcal{Q}(x)w_2}{\mathbf{P}(x)} + \frac{(1-\mathcal{Q}(x))(1-w_2)}{1-\mathbf{P}(x)}\right) \\ \times E\left[w_2 g_1(w_1, x, \beta) - (1 - w_2)\, g_0(w_1, x, \beta) \,\middle|\, w_2, x, D_1 > D_0\right]\Bigg] = 0.$$

This equation in general defines an over-identified system of moments for β. Using a constant matrix A (which we can then choose optimally) we can transform this vector of moments to an exactly identified system. The Jacobi matrix J for this system given A is computed in the standard way. The following theorem describes the structure of the efficient influence function for this model.

Theorem 6 *In the model given by the general moment condition (9) the efficient influence function, corresponding to finite-dimensional parameter β can be expressed as:*

$$\begin{aligned}\Phi(w,x,z) &= -J^{-1}A\tfrac{z-\mathcal{Q}(x)}{1-\mathcal{Q}(x)}\Big\{w_2 g_1(w_1,x,\beta)+(1-w_2)\,g_0(w_1,x,\beta)\\ &-\tfrac{1}{\mathcal{Q}(x)}\Big[(1-\mathcal{Q}(x))\,E\left[(1-w_2)\,g_0(w_1,x,\beta)\mid z=1,x\right]+\mathcal{Q}(x)E\left[w_2 g_1(w_1,x,\beta)\mid z=0,x\right]\Big]\Big\}\\ &= -J^{-1}A\,\phi(w,x,z).\end{aligned}$$

The structure of the efficient influence function in this case is similar to that in the ATE model which we considered earlier in this section. We can further choose the matrix A such that it minimizes the variance of the efficient influence function. In particular, given that the Jacobi matrix can be expressed as:

$$J = A\,\frac{\partial\varphi(\beta)}{\partial\beta'},$$

the semiparametric efficiency bound for this model when A is chosen optimally takes the form:

$$V(\beta) = \left(\frac{\varphi(\beta)}{\partial\beta}E\left[\phi(w,x,z)\,\phi(w,x,z)'\right]\frac{\varphi(\beta)}{\partial\beta'}\right)^{-1}.$$

An optimally weighted GMM estimator based on the following nonparametrically estimated moment condition:

$$\frac{\frac{1}{N}\sum_{k=1}^{N}\left(\widehat{E}\left[g_1(w_{1k},x_k,\beta)\mid z_k=1,x_k\right]-\widehat{E}\left[g_0(w_{1k},x_k,\beta)\mid z_k=0,x_k\right]\right)}{\frac{1}{N}\sum_{k=1}^{N}\left(\widehat{\mathcal{P}}_1(x_k)-\widehat{\mathcal{P}}_0(x_k)\right)}=0,$$

can easily be shown to achieve the efficiency bound derived in the Theorem 6.

Similarly, it is immediate to develop semiparametric efficiency bounds for a nonlinear treatment effect parameter for *treated compliers*, defined as:

$$\varphi(\gamma) = E\Big[g_1(Y_1, x, \gamma) - g_0(Y_0, x, \gamma) \Big| w_2 = 1, D_1 > D_0\Big] = 0.$$

In addition, an optimally weighted GMM estimator based on the following nonparametric estimated moment condition:

$$\frac{\frac{1}{N}\sum_{k=1}^{N} \mathcal{Q}_{\hat{\alpha}}(x_k)\Big(\widehat{E}\left[g_1(w_{1k}, x_k, \gamma) \mid z_k = 1, x_k\right] - \widehat{E}\left[g_0(w_{1k}, x_k, \gamma) \mid z_k = 0, x_k\right]\Big)}{\frac{1}{N}\sum_{k=1}^{N} \hat{\mathcal{Q}}_{\hat{\alpha}}(x_k)\Big(\widehat{\mathcal{P}}_1(x_k) - \widehat{\mathcal{P}}_0(x_k)\Big)} = 0,$$

where $\hat{Q}_\alpha(x_k)$ can be nonparametrically estimated, parametrically estimated, or the known propensity, can easily be shown to achieve the required corresponding semiparametric efficiency bound when the propensity score is unknown, parametrically specified, or known.

5 Empirical application

The analyzed dataset contains the observations from the Family Transition Program (FTP) which was conducted in Escambia County in the state of Florida from the year 1994 to the year 1999. The subsample under consideration contains the data for 2,815 individuals applying for welfare in the year 1994 and early 1995.

The FTP program has been launched to analyze it as an alternative to the welfare program existing at the time, the Aid to Families with Dependent Children (AFDC). The main differences between the two programs are, first, that FTP had a rigid time limit when a family can receive cash assistance (up to 24 months within any 72 - months period). Second, under FTP much more intensive training was offered to the participants aiming at improvement of job skills as well as job search skills.

The individuals applying for welfare were randomly assigned to either AFDC or FTP which allows one to compare the relative effect of the rules of the two welfare programs. In addition to the immediate effect of the program, the collected dataset tracks the individuals for the next 4 years after the program allowing to compare long-term impacts of welfare programs on individuals.

The main sample contains the data for 2,815 heads of single parent households who applied for welfare and were randomly assigned to one of the welfare options between May 20, 1994 and February 31, 1996. In this sample 1,405 individuals were assigned to FTP and 1,410 individuals were assigned to AFDC. The data contains three main blocks. The administrative record data contain the data for individual incomes from three sources in the state administration. First, the earnings from work from the state's Unemployment Insurance system. The second source of incomes are the payments from AFDC. The third source are Food Stamp payments. In addition, this dataset contains the information about the background characteristics of individuals and the data from the private opinion survey. The adult survey data contain the information obtained by MRDC (Manpower Demonstration Research Corporation). This information was collected in 45-minute interviews with 1,730 individuals from the main data sample which were administered between October 1998. This additional set contains information about characteristics of individual (including education, job experience, family and dependents, housing, food security, and living conditions). The child survey data are based on 1,100 additional interviews with adult survey participants, who have at least one child between 5 and 12 years old. This survey is inquiring about the school outcomes, kid's interaction with other children. The information contained in the survey includes parenting, father's involvement. The administrative record data contain 1132 variables and the survey data contain 849 variables in the adult survey and 679 variables in the child survey.

One of the surprising outcomes of the program is the relative deterioration of the school performance of children in the least disadvantaged families. Specifically in the group of families with the largest earning impacts the school performance of children including grades and suspension is worse in the FTP sample than in AFDC sample. One of the hypothesis to explain this is that in this group the parents worked the longest hours and were not able to monitor their children closely. However, we cannot directly use the data to test this hypothesis because of selection on unobservable ability: the unobserved ability of parents should be correlated with both the school performance of children and with the impact of training on parents. In this case if we use the assignment to a specific program as an instrument, then we will be able to identify the impact of parents' training on the children's

school performance on the set of complier who will only be employed because of training.

In particular, we study the influence of the work status of parents on the count indicator of a child's achievement which grades the school achievement from 1 to 5. The main problem in these circumstances is that a simple linear relationship between the indicator of achievement and the fact that the parent is working is contaminated by the influence of the unobserved ability. In fact, the parent's ability, indicating his or her capability of finding a job should be correlated with the child's ability, which influences the achievement grade. For this reason, to obtain the correct measure of dependence of child's achievement on the parent's employment we can use the instruments to correct for the biased caused by the endogeneity of the employment dummy. One of such instrument can be the participation of the parent in FTP as compared to AFDC, because the former participation has increased the probability of employment.

In Table 1 we present the results of such modeling for a subset of individuals who ever took a job (dropping those who never worked for pay). We regress the child's achievement variable on the the dummy indicating that the parent is currently employed, child's gender, age dummies for the parent, and the dummy indicating that the parent doe not have a high school diploma. In columns (1) we report the results of a simple linear regression. Column (2) reports the results of Poisson regression, and column (3) reports the results of the negative binomial regression where the achievement of a child is considered as a count outcome. The coefficient of parent's employment is quite small in all the models. In column (4) we report the results of the IV regression where we use instruments for the employment dummy including the FTP/AFDC dummy and hourly wage of the last taken job. As one can see, the coefficient indicates now that the parent's employment leads to the decline in child's achievement by almost two points. This suggests that if we do not take into account the endogeneity of employment dummy, we will understate the influence of parent's employment on the child's achievement at school.

To apply our estimation method we adopt the conditional moment condition implied by the count structure of the outcome variable (child's achievement grade). The moment condition corresponds to the score of the Poisson regression mode with

$$g\left(w,x,\beta\right)=\left[w_1-\exp\left(\beta_0 w_2+x'\beta_1\right)\right]\begin{pmatrix}w_2\\ x\end{pmatrix},$$

Table 1: Regression outcomes

	Model number					
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Employment dummy	-0.196	-0.049	-0.049	-1.882	-0.1550	-0.1550
	(2.56)*	(2.54)**	(2.54)**	(3.88)**	(3.02)**	(3.02)**
Male dummy	-0.235	-0.058	-0.058	-0.31	-0.0126	-0.0126
	(3.36)**	(3.37)**	(3.37)**	(3.34)**	(4.01)**	(4.01)**
Age 25-34	-0.309	-0.076	-0.076	-0.394	-0.0585	-0.0585
	(4.13)**	(4.14)**	(4.14)**	(3.97)**	(3.11)**	(3.11)**
Age 35-44	-0.127	-0.031	-0.031	-0.16	-0.0247	-0.0247
	-1.09	-1.08	-1.08	-1.11	-0.94	-0.94
No high school degree	-0.139	-0.035	-0.035	0.086	-0.1308	-0.1308
	-1.86	(1.86)*	(1.86)*	-0.72	-0.88	-0.88
Constant	4.87	1.601	1.601	7.202	1.5119	1.5119
	(32.83)**	(44.21)**	(44.21)**	(10.58)**	(20.21)**	(20.21)**
N. obs	918	918	918	887	918	918
$\mathbf{R}^2$	0.04					

Robust t-statistics in parentheses * significant at 5% level; ** significant at 1% level

and the moment condition:

$$\varphi\left(w_2, x, \beta\right) = E\left[g\left(w, x, \beta\right) \mid x, w_2, d_1 > d_0\right] = 0.$$

We apply the efficient estimator developed in the previous section to estimate β.

For the negative binomial model the moment condition was formed by:

$$g\left(w, x, \beta\right) = \left[w_1 - \frac{\delta^{-2} + w_1}{\delta^{-2} + \exp\left(\beta_0 w_2 + x'\beta_1\right)} \exp\left(\beta_0 w_2 + x'\beta_1\right)\right] \begin{pmatrix} w_2 \\ x \end{pmatrix},$$

and the moment conditions were written similarly to the case of the Poisson model. The results of efficient estimation adopted to the moment equations implied by the scores of Poisson and negative binomial regression models are presented in Column (5) and (6).

Similarly to the 2SLS case, the coefficients in the count data models with endogeneity taken into account are significantly larger in the absolute value than in the models which do not take endogeneity into account. This implies that the endogeneity of job participation causes and upward bias in the estimate of the influence of parent's employment on child's achievement. One can also see that the values of other coefficients remain the same in the order of magnitude which indicates robustness of our results.

The marginal effects of variables in the Poisson model are consistent with the estimates from linear models. Specifically the treatment effect in the model (2) is -.196 with a standard error of .0773 with is almost identical to the corresponding estimate from the linear model. In the case with moment condition, taking endogeneity into account, the marginal effect estimate is is -.538 with a standard error of .177. This is smaller than the 2SLS estimate but almost 4 times larger than the marginal effect in the model not taking endogeneity of job participation into account.

6 Conclusion

In this paper we derive the semiparametric efficiency bound for the estimation of a finite-dimensional parameter defined by generalized moment conditions under the local instrumental variable assumptions of Imbens and Angrist (1994) and Abadie, Angrist, and Imbens (2002). These parameters identify the treatment effect on the set of compliers under the monotonicity assumption. The moment equation characterizes the parametrized moment of the outcome distribution given a set of covariates and the treatment dummy. The distributions of covariates, treatment dummy and the binary instrument are not specified in a parametric form, making the model semiparametric. We also develop multi-step semiparametric efficient estimators that achieve the semiparametric efficiency bound.

We apply the suggested estimation procedure to analyze the effect of the parent's employment on children's achievement for single-parent households who applied for state welfare support. We use the data from the Florida Transition Program which was offered as an alternative to the existing state welfare system. FTP gives the participants more extensive training aimed at job-search skills for participants. A general concern is that both the achievement of the child and the probability of a parent to find a job are correlated with

their natural abilities. The abilities of children and parents, on the other hand, can be significantly correlated. For this reason, without accounting for the endogeneity, the estimated treatment effect will be biased toward zero due to the attenuation bias. We use the dummy for program participation as a natural instrument for selection based on individual abilities. We find that the linear regression estimate of the treatment effect has a substantial attenuation bias as compared to instrument-based methods. In general, parent's employment adversely affects child's achievement. This result suggests that ignoring the selection effect indeed leads to substantial bias in the estimate of effect of parent's employment and thus if we ignore selection, we will have significantly underestimated the negative effect.

In addition, we also construct a similar procedure to compute the semiparametric efficiency bound when a conditional moment equation is specified for the subset of noncompliers, for whom selection into the program does not change the treatment status. Both the structure of the semiparametric efficiency bound and the optimal transformation of the conditional moment turn out to be significantly simpler than those for the subset of compliers. In the extension, appendix D also derives the semiparametric efficiency bound for a semiparametric specification of the moment equations. In work in progress, we are developing efficient estimators when the moment condition is semiparametrically specifically, which will be applied to empirical analysis.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effects of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Angrist, J. (2004): "Treatment Effect Heterogeneity in Theory and Practice," *Economic Journal*, 114(494), 83.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables.," *Journal of the American Statistical Association*, 91(434).

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag New York, Inc.

Chen, X., H. Hong, and A. Tarozzi (2005): "Semiparametric Efficiency in GMM Models Nonclassical Measurement Errors," working paper, Duke University and New York University.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

Firpo, S. (2006): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*.

Frolich, M. (2006): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139(1), 35–75.

Hahn, J. (1998a): "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–332.

——— (1998b): "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–332.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Imbens, G., and D. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 64(4), 555–574.

Murphy, S., and A. van der Vaart (1997): "Semiparametric likelihood ratio inference," *Ann. Statist*, 25(4), 1471–1509.

Newey, W. (1990a): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809–837.

——— (1990b): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–82.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

Severini, T., and G. Tripathi (2001): "A simplified approach to computing efficiency bounds in semiparametric models," *Journal of Econometrics*, 102, 23–66.

A Proof of theorem 1

Consider the parametrization θ for the model with covariates x. Define

$$f^1(w_1|w_2,z,x) = f(y_1 = w_1|w_2,z,x)$$

and $f^0(w_1|w_2,z,x) = f(y_0 = w_1|w_2,z,x)$. If $\phi_\theta(x)$ is the Radon-Nykodym density of x with the support on $\mathcal{X}$, the likelihood function for the data can be written as:

$$f_\theta(w,z,x) = \left[f_\theta^1(w_1 \mid w_2,\, z,\, x)\right]^{w_2} \left[f_\theta^0(w_1 \mid w_2,\, z,\, x)\right]^{(1-w_2)} \mathcal{F}_\theta^{w_2}(x,z)\left(1-\mathcal{F}_\theta(x,z)\right)^{(1-w_2)} \times \mathcal{Q}_\theta^z(x)\left(1-\mathcal{Q}_\theta(x)\right)^{(1-z)} \phi_\theta(x).$$

The score of the model associated with the joint density of observed data is specified as

$$S_\theta(w,z,x) = (1-w_2)s_\theta^0(w_1 \mid w_2,\, z,\, x) + w_2 s_\theta^1(w_1 \mid w_2,\, z,\, x)$$

$$+\frac{(1-z)\dot{\mathcal{P}}_{0\theta}(x)}{\mathcal{F}(z,x)\left(1-\mathcal{F}(z,x)\right)}\left[w_2 - \mathcal{F}(z,x)\right] + \frac{z\dot{\mathcal{P}}_{1\theta}(x)}{\mathcal{F}(z,x)\left(1-\mathcal{F}(z,x)\right)}\left[w_2 - \mathcal{F}(z,x)\right]$$

$$+\frac{\dot{\mathcal{Q}}_\theta(x)}{\mathcal{Q}(x)\left(1-\mathcal{Q}(x)\right)}\left[z - \mathcal{Q}(x)\right] + s_\theta(x),$$

where $s_\theta(x)$ is the score corresponding to $\phi_\theta(x)$. The expression for the tangent set of the model for conditional distribution moments is given by

$$\mathcal{T} = \left\{(1-w_2)s_\theta^0(w_1 \mid w_2,\, z,\, x) + w_2 s_\theta^1(w_1 \mid w_2,\, z,\, x) + z\,\xi(x,z)\left[w_2 - \mathcal{F}(z,x)\right]\right.$$

$$\left.+(1-z)\,\zeta(x,\, z)\left[w_2 - \mathcal{F}(z,x)\right] + a(x)\left[z - \mathcal{Q}(x)\right] + t(x)\right\},$$

where $E_\theta\left[s_\theta^i(w_1 \mid w_2,\, z,\, x) \mid w_2,\, z,\, x\right] = 0$ for $i = 0, 1$, $E\{t(x)\} = 0$ and $\zeta(\cdot)$, $\xi(\cdot)$ and $a(\cdot)$ are square - integrable functions.

Now consider the directional derivative of the parameter vector β determined by the conditional moment equation $\varphi(x,\, w_2,\, \beta)$. We assume that the support of x - the set $\mathcal{X}$ is non-degenerate. In this case we can potentially identify a parameter vector β with arbitrarily many dimensions. Our strategy now will be to define a matrix of instrument functions which will transform the conditional moment equation to an exactly identified system of unconditional moments. Suppose that $\mathcal{A}(w_2,\, x)$ is an arbitrary continuous vector - function such that $\mathcal{A} : \{0,\, 1\} \times \mathcal{X} \mapsto \mathbb{R}^k$. Without loss of generality we can define this vector-function by a vector

$$\zeta(x, w_2) = \left(\frac{w_2}{\mathbf{P}(x)},\, \frac{1-w_2}{1-\mathbf{P}(x)}\right)',$$

and a matrix $\mathcal{M}(x)$ of dimension $\dim(\beta) \times 2$ such that

$$A(w_2, x) = \mathcal{M}(x)\,\zeta(w_2, x),$$

and

$$\mathcal{A}(w_2, x) = (\mathcal{P}_1(x) - \mathcal{P}_0(x)) \left(\frac{\mathcal{Q}(x) w_2}{\mathbf{P}(x)} + \frac{(1 - \mathcal{Q}(x))(1 - w_2)}{1 - \mathbf{P}(x)} \right) A(w_2, x).$$

The instrument functions transfer the model into set of unconditional moment equations in the form:

$$E\left[\mathcal{A}(w_2,\, x)\, \varphi(x,\, w_2,\, \beta)\right] = 0.$$

This can be rewritten as:

$$E_\theta\left[\mathcal{A}_\theta(w_2,\, x)\, E_\theta\left[g(w, x,\, \beta(\theta)) \,|w_2, x, d_1 > d_0\right]\right] = 0.$$

Define the Jacobi matrix:

$$J = E\left[\mathcal{A}(w_2, x)\, \frac{\partial \varphi(x,\, w_2,\, \beta)}{\partial \beta'}\right].$$

Then we can solve for the directional derivative of the parameter β by solving a $k \times k$ system of equations:

$$J\, \frac{\partial \beta(\theta)}{\partial \theta} = -\frac{\partial}{\partial \theta} E_\theta\left[\mathcal{A}_\theta(w_2,\, x)\, E_\theta\left[g(w, x,\, \beta) \,|w_2, x, d_1 > d_0\right]\right]. \tag{10}$$

The right-hand side component of equation (10) can be written as:

$$\begin{aligned} &E\left[\mathcal{A}(w_2, x) \int g(w, x, \beta)\, s_{**}(w_1 \mid w_2,\, x)\, f_{**}(w_1 \mid w_2,\, x)\, dw_1\right] \\ &\quad + E\left[\mathcal{A}(w_2, x)\, s_\theta(w_2, x) \int g(w, x, \beta)\, f_{**}(w_1 \mid w_2,\, x)\, dw_1\right] \\ &\quad\quad + E\left[\tfrac{\partial \mathcal{A}_\theta(w_2, x)}{\partial \theta} \int g(w, x, \beta)\, f_{**}(w_1 \mid w_2,\, x)\, dw_1\right]. \end{aligned} \tag{11}$$

In the above we have used the definition that

$$s_{**}(w_1 \mid w_2,\, x) = \frac{\partial}{\partial \theta} \log f^\theta_{**}(w_1 \mid w_2,\, x),$$

where $f_{**}(w_1 \mid w_2, x)$ takes the form of either 1 of 2 depending on whether $w_2 = 1$ or $w_2 = 0$. Note that by definition of the function $g(w,x,\beta)$ the integral:

$$\int g(w, x)\, f_{**}(w_1 \mid w_2,\, x)\, dw_1 = 0,$$

and the last two terms of equation (11) can be removed. Using this result the system for the directional derivative of the parameter vector can be rewritten as:

$$\frac{\partial \beta(\theta)}{\partial \theta} = -J^{-1} E\left[\mathcal{A}(w_2, x) \int g(w, x, \beta)\, s_{**}(w_1 \mid w_2,\, x)\, f_{**}(w_1 \mid w_2,\, x)\, dw_1\right]$$

Next we introduce the notations:

$$\widetilde{g}(w,\, x,\, \beta) = -J^{-1} A(w_2, x) g(w,\, x,\, \beta),$$

$$\widehat{g}(w,\, x,\, \beta) = (\mathcal{P}_1(x) - \mathcal{P}_0(x)) \left(\frac{\mathcal{Q}(x) w_2}{\mathbf{P}(x)} + \frac{(1 - \mathcal{Q}(x))(1 - w_2)}{1 - \mathbf{P}(x)}\right) \widetilde{g}(w,\, x,\, \beta),$$

and

$$\widehat{\Delta}(w_2,\, x,\, \beta) = -J^{-1} \mathcal{A}(w_2, x) \Delta(w_2,\, x,\, \beta).$$

Recall the definition that

$$\Delta(x, w_2, \beta) = \{E[g(w_1, w_2, x, \beta) \mid w_2, z = w_2, x] - E[g(w_1, w_2, x, \beta) \mid w_2, z = 1 - w_2, x]\}.$$

Differentiating equations (1) and (2) and combining notations give rise to the following expression for the score:

$$\begin{aligned}
& s_{**}(w_1 \mid w_2, x)\, f_{**}(w_1 \mid w_2, x) \\
= \;& \frac{w_2 \mathcal{P}_1(x) + (1 - w_2)(1 - \mathcal{P}_0(x))}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} s_\theta(w_1 \mid w_2, z = w_2, x)\, f(w_1 \mid w_2, z = w_2, x) \\
- \;& \frac{w_2 \mathcal{P}_0(x) + (1 - w_2)(1 - \mathcal{P}_1(x))}{\mathcal{P}_1(x) - \mathcal{P}_0(x)} s_\theta(w_1 \mid w_2, z = 1 - w_2, x)\, f(w_1 \mid w_2, z = 1 - w_2, x) \\
+ \;& \left[\frac{[1 - w_2 + \mathcal{P}_1(x)(2w_2 - 1)]\, \dot{\mathcal{P}}_{0\theta}(x) - [1 - w_2 + \mathcal{P}_0(x)(2w_2 - 1)]\, \dot{\mathcal{P}}_{1\theta}(x)}{(\mathcal{P}_1(x) - \mathcal{P}_0(x))^2}\right] \\
\times \;& (f(w_1 \mid w_2, z = w_2, x) - f(w_1 \mid w_2, z = 1 - w_2, x)).
\end{aligned}$$

Following Newey (1990b) we look for a set of influence functions $\Psi(w, z, x)$ belonging to the tangent space $\mathcal{T}$ with the properties that

$$\frac{\partial \beta(\theta)}{\partial \theta} = E\Psi(w, z, x) S_\theta(w, z, x).$$

We conjecture and subsequently verify that the efficient influence function takes the form of:

$$\begin{aligned}
\Psi(w,z,x) &= \frac{\mathbf{P}(x)w_2\, z}{\mathcal{Q}(x)\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)}\left(\widehat{g}(w,x,\beta)-E\left[\widehat{g}(w,x,\beta)\mid w_2=1,z=1,x\right]\right)\\
&- \frac{\mathbf{P}(x)w_2\,(1-z)}{(1-\mathcal{Q}(x))\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)}\left(\widehat{g}(w,x,\beta)-E\left[\widehat{g}(w,x,\beta)\mid w_2=1,z=0,x\right]\right)\\
&+ \frac{(1-\mathbf{P}(x))\,(1-w_2)\;(1-z)}{(1-\mathcal{Q}(x))\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)}\left(\widehat{g}(w,x,\beta)-E\left[\widehat{g}(w,x,\beta)\mid w_2=0,z=0,x\right]\right)\\
&- \frac{(1-\mathbf{P}(x))\,(1-w_2)\;z}{\mathcal{Q}(x)\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)}\left(\widehat{g}(w,x,\beta)-E\left[\widehat{g}(w,x,\beta)\mid w_2=0,z=1,x\right]\right)\\
&+ \frac{\widehat{\Delta}(w_2=1,x)\mathbf{P}(x)}{\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)^2}\left[\frac{\mathcal{P}_1(x)(1-z)}{1-\mathcal{Q}(x)}-\frac{\mathcal{P}_0(x)z}{\mathcal{Q}(x)}\right](w_2-\mathcal{F}(z,x))\\
&+ \frac{\widehat{\Delta}(w_2=0,x)\,(1-\mathbf{P}(x))}{\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)^2}\left[\frac{(1-\mathcal{P}_1(x))\,(1-z)}{1-\mathcal{Q}(x)}-\frac{(1-\mathcal{P}_0(x))\,z}{\mathcal{Q}(x)}\right](w_2-\mathcal{F}(z,x)).
\end{aligned}$$

The first two lines correspond to the $w_2 s_\theta^1(w_1|w_2,z,x)$ component of $\mathcal{T}$. The third and fourth lines correspond to the $(1-w_2)\,s_\theta^0(w_1|w_2,z,x)$ component of $\mathcal{T}$. The last two lines correspond to the

$$z\,\xi(x,z)\left[w_2-\mathcal{F}(z,x)\right]+(1-z)\,\zeta(x,\,z)\left[w_2-\mathcal{F}(z,x)\right]$$

component of the tangent space $\mathcal{T}$. The last two components of tangent space are null components.

Next we use the identities, implied by the expressions for conditional density (1) and (2):

$$\begin{aligned}
&\mathcal{P}_1(x)\,E\left[\widehat{g}(w,x,\beta)\mid w_2=1,z=1,x\right]=\mathcal{P}_0(x)\,E\left[\widehat{g}(w,x,\beta)\mid w_2=1,z=0,x\right]\\
&(1-\mathcal{P}_0(x))\,E\left[\widehat{g}(w,x,\beta)\mid w_2=0,z=0,x\right]=(1-\mathcal{P}_1(x))\,E\left[\widehat{g}(w,x,\beta)\mid w_2=0,z=1,x\right].
\end{aligned}\tag{12}$$

In addition, we substitute the expression for the weighing matrix $\mathcal{A}(w_2,x)$ into obtained expression for the influence function. This leads us to the final result for the efficient influence function:

$$\begin{aligned}
\Psi(w,z,x) &= \left(w_2 z-\frac{\mathcal{Q}(x)}{1-\mathcal{Q}(x)}w_2(1-z)+(1-w_2)(1-z)-\frac{1-\mathcal{Q}(x)}{\mathcal{Q}(x)}(1-w_2)\,z\right)\tilde{g}(w,x,\beta)\\
&-\frac{(z-\mathcal{Q}(x))}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\left\{E\left[w_2 z\tilde{g}\mid w_2=1,z=1,x\right]-E\left[(1-w_2)(1-z)\,\tilde{g}\mid w_0=1,z=0,x\right]\right\}\\
&=\Psi_1(w,z,x)-\Psi_2(w,z,x).
\end{aligned}$$

We can now express the semiparametric efficiency bound as the variance of the efficient influence function

$$V\left(\widehat{\beta}\right) = E\left\{\Psi\,\Psi'\right\}.$$

Note that the vector $\mathcal{A}(w_2, x)$ can be represented as:

$$\mathcal{A}(w_2, x) = (\mathcal{P}_1(x) - \mathcal{P}_0(x))\left(\frac{\mathcal{Q}(x)w_2}{\mathbf{P}(x)} + \frac{(1-\mathcal{Q}(x))(1-w_2)}{1-\mathbf{P}(x)}\right)\mathcal{M}(x)\begin{pmatrix} \frac{w_2}{\mathbf{P}(x)} \\ \frac{1-w_2}{1-\mathbf{P}(x)} \end{pmatrix},$$

where $\mathcal{M}(x)$ is a $k \times 2$ matrix (k is the size of the Euclidean parameter β). Denote

$$D(x) = \operatorname{diag}\left\{\frac{\mathcal{Q}(x)}{\mathbf{P}(x)},\ \frac{1-\mathcal{Q}(x)}{1-\mathbf{P}(x)}\right\}.$$

In this case the Jacobi matrix can be written as

$$J = E\left\{(\mathcal{P}_1(x) - \mathcal{P}_0(x))\,\mathcal{M}(x)\begin{pmatrix} \frac{\mathcal{Q}(x)w_2}{\mathbf{P}^2(x)} \\ \frac{(1-\mathcal{Q}(x))(1-w_2)}{(1-\mathbf{P}(x))^2} \end{pmatrix}\frac{\partial\varphi(w_2,x,\beta)}{\partial\beta'}\right\} = E\left\{(\mathcal{P}_1(x) - \mathcal{P}_0(x))\,\mathcal{M}(x)D(x)\theta(x)\right\}.$$

To facilitate the manipulations denote $\omega_{w_2,z}(x) = V\left(g\left(w, x, \beta\right) \mid w_2, z, x\right)$ and

$$\gamma_{w_2,z}(x) = E\left(g\left(w, x, \beta\right) \mid w_2, z, x\right).$$

Note that the expression for the variance has three components. The first component corresponds to the variance of the first component $\Psi_1(w, x, z)$:

$$V\left(\Psi_1(w, x, z)\right) =$$

$$J^{-1}\,E\left\{\mathcal{M}(x)D(x)\begin{pmatrix} \frac{\mathcal{P}_1(x)\omega_{11}(x)}{\mathcal{Q}(x)} + \frac{\mathcal{P}_0(x)\omega_{10}(x)}{(1-\mathcal{Q}(x))} + \frac{\mathcal{P}_1(x)\mathbf{P}(x)}{\mathcal{P}_0(x)\mathcal{Q}(x)(1-\mathcal{Q}(x))}\gamma_{11}^2(x) & 0 \\ 0 & \frac{(1-\mathcal{P}_0(x))\omega_{00}(x)}{(1-\mathcal{Q}(x))} + \frac{(1-\mathcal{P}_1(x))\omega_{01}(x)}{\mathcal{Q}(x)} + \frac{(1-\mathcal{P}_0(x))(1-\mathbf{P}(x))}{(1-\mathcal{P}_1(x))\mathcal{Q}(x)(1-\mathcal{Q}(x))}\gamma_{00}^2(x) \end{pmatrix}D(x)\mathcal{M}(x)'\right\}J^{-1\prime}.$$

The second component can be re-arranged using the Jacobi matrix and instrument matrix $\mathcal{M}(x)$:

$$\Psi_2\left(w, x, z\right) = J^{-1}\mathcal{M}(x)\begin{bmatrix} \frac{\mathcal{P}_1(x)\mathcal{Q}(x)}{\mathbf{P}(x)}\gamma_{11}(x) \\ -\frac{(1-\mathcal{P}_0(x))(1-\mathcal{Q}(x))}{1-\mathbf{P}(x)}\gamma_{00}(x) \end{bmatrix}\frac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}.$$

The corresponding variance is:

$$V\left(\Psi_2\left(w, x, z\right)\right)$$

$$= J^{-1}E\left\{\frac{\mathcal{M}(x)D(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\begin{pmatrix} \mathcal{P}_1^2(x)\gamma_{11}^2(x) & -\mathcal{P}_1(x)(1-\mathcal{P}_0(x))\gamma_{11}(x)\gamma_{00}(x) \\ -\mathcal{P}_1(x)(1-\mathcal{P}_0(x))\gamma_{11}(x)\gamma_{00}(x) & (1-\mathcal{P}_0(x))^2\gamma_{00}^2(x) \end{pmatrix}D(x)\mathcal{M}(x)'\right\}J^{-1\prime}.$$

The third component is the covariance between the first two elements:

$$cov\left(\Psi_1\left(w,x,z\right),\ \Psi_2\left(w,x,z\right)\right)=-V\left(\Psi_2\left(w,x,z\right)\right).$$

The variance of the efficient influence function can then be written as:

$$V\left(\widehat{\beta}\right)=J^{-1}E\left\{\mathcal{M}(x)D(x)\overline{\Omega}(x)D(x)\mathcal{M}(x)'\right\}J^{-1\prime}=E\left\{\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)\mathcal{M}(x)D(x)\theta(x)\right\}^{-1}$$
$$\times E\left\{\mathcal{M}(x)D(x)\overline{\Omega}(x)D(x)\mathcal{M}(x)'\right\}E\left\{\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)\mathcal{M}(x)D(x)\theta(x)\right\}^{-1\prime},$$

where $\overline{\Omega}(x)$ is a 2×2 matrix constructed from conditional variances and conditional expectations of the moment function. The components of $\overline{\Omega}(x)$ can be expressed in the following way:

$$\overline{\Omega}_{11}(x)=\left(\tfrac{\mathcal{P}_1(x)\omega_{11}(x)}{\mathcal{Q}(x)}+\tfrac{\mathcal{P}_0(x)\omega_{10(x)}}{1-\mathcal{Q}(x)}+\tfrac{\gamma_{11}^2(x)\mathcal{P}_1(x)\mathbf{P}(x)}{\mathcal{P}_0(x)\mathcal{Q}(x)(1-\mathcal{Q}(x))}\left[1-\tfrac{\mathcal{P}_1(x)\mathcal{P}_0(x)}{\mathbf{P}(x)}\right]\right),$$

$$\overline{\Omega}_{22}(x)=\left(\tfrac{(1-\mathcal{P}_1(x))\omega_{01}(x)}{\mathcal{Q}(x)}+\tfrac{(1-\mathcal{P}_0(x))\omega_{00}(x)}{1-\mathcal{Q}(x)}+\tfrac{\gamma_{00}^2(x)(1-\mathcal{P}_0(x))(1-\mathbf{P}(x))}{\mathcal{Q}(x)(1-\mathcal{Q}(x))(1-\mathcal{P}_1(x))}\left[1-\tfrac{(1-\mathcal{P}_0(x))(1-\mathcal{P}_1(x))}{1-\mathbf{P}(x)}\right]\right),$$

and

$$\overline{\Omega}_{21}(x)=\overline{\Omega}_{12}(x)=\left(\tfrac{\mathcal{P}_1(x)(1-\mathcal{P}_0(x))}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\gamma_{11}(x)\gamma_{00}(x)\right).$$

By standard GMM-type arguments we find that the minimum variance is achieved when $\mathcal{M}(x)=\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)\theta(x)'\overline{\Omega}(x)^{-1}D(x)^{-1}$ and the semiparametric efficiency bound is

$$V\left(\widehat{\beta}\right)=E\left\{\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)^2\theta(x)'\overline{\Omega}(x)^{-1}\theta(x)\right\}^{-1}.$$

B Proof of theorems 2 and 3

The proof of theorem 2 is self-evident. In particular, assumption 2.2, 2.3, 2.4 combined to insure that uniformly, the estimated $\hat{\mathcal{Q}}\left(x\right)$ and $\hat{A}\left(x,w_2\right)$ can be replaced by their true quantities. Assumption 2.5 insures that a law of large numbers uniform over β applies to $\frac{1}{N}\sum_{k=1}^{N}\psi_k\left(\beta,\mathcal{Q}_0,A_0\right)$.

Before verifying the regularity conditions of theorem 3 we discuss the intuition of the asymptotic distribution. It is not difficult to see that estimating $\hat{\mathcal{M}}\left(x\right)$ has no impact on the asymptotic variance, because for example, for all x,

$$E\left[\frac{\partial\psi_k\left(\beta,\mathcal{Q}\left(x\right),\mathcal{M}\left(x\right)\right)}{\partial\mathcal{M}\left(x\right)}\middle|x\right]=0.$$

Following the argument of Newey (1994)), the following asymptotic representation holds:

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\widehat{\psi}_k\left(\beta\right)=\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\mathcal{M}(x_k)\left\{\chi_k\left(\beta\right)+E\left[\frac{\partial\chi_k\left(\beta\right)}{\partial Q}\middle|x_k\right]\left(z_k-\mathcal{Q}(x_k)\right)\right\}+o_p\left(1\right).$$

Next we will show that the asymptotic variance of this moment function when $\mathcal{M}(x)$ is chosen optimally equals the semiparametric efficiency bound. Let us use the notations $\Psi_1(w,x,z)$ and $\Psi_2(w,x,z)$ corresponding to the two components of the efficient influence function in the proof of Theorem 1. In this case the first component of this expression corresponds to:

$$\mathcal{M}(x)\chi(\beta) = J\,\Psi_1(w,x,z),$$

for (in the above we omitted the subscript k in $\mathcal{M}(x_k)$ and $\chi_k(\beta)$)

$$J = E\left[\mathcal{A}(w_2,x)\frac{\partial\varphi(x,\,w_2,\,\beta)}{\partial\beta'}\right].$$

The second component corresponds to the sampling uncertainty due to the error in estimation of probability $\mathcal{Q}(x)$. To compute it note that

$$\mathcal{M}(x)E\left[\tfrac{\partial\chi(\beta)}{\partial Q}\big|x\right](z-\mathcal{Q}(x)) = \tfrac{\mathcal{M}(x)(z-\mathcal{Q}(x))}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\begin{bmatrix} -\frac{\mathcal{P}_1(x)\mathcal{Q}(x)}{\mathbf{P}(x)}\gamma_{11}(x) \\ \frac{(1-\mathcal{P}_0(x))(1-\mathcal{Q}(x))}{1-\mathbf{P}(x)}\gamma_{00}(x)\end{bmatrix} = -J\,\Psi_2(w,x,z).$$

This means in particular that:

$$\widehat{\psi}(\beta) = J\left[\Psi_1(w,x,z) - \Psi_2(w,x,z)\right] = J\,\Psi(w,x,z),$$

so that the influence function is a scaled efficient influence function. Therefore, the variance of this estimator can be represented as

$$V\left(\widehat{\beta}\right) = J^{-1}V\left(J\,\Psi(w,x,z)\right)J^{-1} = V\left(\Psi(w,x,z)\right).$$

Thus, the estimator achieves the semiparametric efficiency bound.

Assumptions 3.1, 3.2 and 2.2, 2.3 and 2.4 combined to insure stochastic equicontinuity, i.e, conditions (3.2) and (3.3) in Theorem 3 of Chen, Linton, and Van Keilegom (2003) (CLK).

Assumption 3.5 is used to justify (2.3).ii of Theorem 2 of CLK. Assumptions 3.3, 3.4 are used to justify (2.3).i and (2.4) of Theorem 2 of CLK.

Assumption 4 basically ensures that

$$E\left[\delta_0(X)\left(\hat{\mathcal{Q}}(X) - \mathcal{Q}_0(X)\right)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_0(X_i)\left(Z_i - \mathcal{Q}_0(X_i)\right) + o_p(1).$$

It summarizes the key elements in (A.7) to (A.10) in proving Theorem 6.1 of Newey (1994).

C Proof of theorems in section 4

The proofs of theorems 4, 5, 6 depend exclusively only on the following lemma, whose proof can be found in, for example, Robins and Rotnitzky (1995) and Hahn (1998b).

Lemma 1 *For a categorial variable Z and a constant a on the support of Z, the semiparametric efficiency variance for estimating $E\left(E\left(W|Z=a,X\right)\right)$ is given by the variance of the following influence function, for $\mathcal{Q}_a\left(X\right)=P\left(Z=a|X\right)$*

$$\frac{1\left(Z=a\right)}{\mathcal{Q}_a\left(X\right)}\left(W-E\left(W|Z=a,X\right)\right)+E\left(W|Z=a,X\right).$$

The rest of the proofs basically amounts to rewrite the parameters of interets in collections of components that take the form of $E\left(E\left(W|Z=a,X\right)\right)$.

C.1 Proof of theorem 4

Part 1:(ATE on compliers) We discuss β_1 and β_0 in turn. Note first that due to independence between y_1 and w_2 conditional on x and $D_1 > D_0$,

$$\begin{aligned}\beta_1 =&E\left[E\left(y_1|w_2=1,x,D_1>D_0\right)|D_1>D_0\right]=E\left[E\left(w_1|w_2=1,x,D_1>D_0\right)|D_1>D_0\right]\\ =&E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)E\left(w_1|w_2=1,x,D_1>D_0\right)\right]\frac{1}{P\left(D_1>D_0\right)}.\end{aligned}$$

So that β_1 is defined by the moment condition:

$$E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)E\left(w_1|w_2=1,x,D_1>D_0\right)\right]-E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\right]\beta_1=0.$$

It suffices to project this equation onto the tangent set. Recall the identification condition:

$$\begin{aligned}&E\left(w_1\mid w_2=1,d_1>d_0,\,x\right)\\ &=\frac{\mathcal{P}_1(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1\mid w_2=1,z=1,x\right)-\frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1\mid w_2=1,z=0,x\right).\end{aligned}$$

The moment condition that defines β_1 is then rewritten as

$$\begin{aligned}E\mathcal{P}_1(x)E\left(w_1\mid w_2=1,z=1,x\right)-E\mathcal{P}_0(x)E\left(w_1\mid w_2=1,z=0,x\right)\\ -E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\right]\beta_1=0.\end{aligned}\tag{13}$$

Or equivalently,

$$\begin{aligned}EE\left(w_2w_1\mid z=1,x\right)-EE\left(w_2w_1\mid z=0,x\right)\\ -\left(EE\left(w_2\mid z=1,x\right)-EE\left(w_2\mid z=0,x\right)\right)\beta_1=0.\end{aligned}\tag{14}$$

Similarly calculations can be applied to β_0, consider

$$\begin{aligned}\beta_0 =&E\left[E\left(y_0|w_2=0,x,D_1>D_0\right)|D_1>D_0\right]=E\left[E\left(w_1|w_2=0,x,D_1>D_0\right)|D_1>D_0\right]\\ =&E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)E\left(w_1|w_2=0,x,D_1>D_0\right)\right]\frac{1}{P\left(D_1>D_0\right)}.\end{aligned}$$

This translates into the moment condition:

$$E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)E\left(w_1|w_2=0,x,D_1>D_0\right)\right]-E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\right]\beta_0=0.$$

Recall the related identification condition:

$$\begin{aligned}&E\left(w_1\mid w_2=0,d_1>d_0,\,x\right)\\&=\frac{\left(1-\mathcal{P}_0(x)\right)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1\mid w_2=0,z=0,x\right)-\frac{\left(1-\mathcal{P}_1(x)\right)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1\mid w_2=0,z=1,x\right).\end{aligned}$$

The relevant moment condition for β_0 is then rewritten as

$$\begin{aligned}E\left(1-\mathcal{P}_0(x)\right)E\left(w_1\mid w_2=0,z=0,x\right)-\left(1-\mathcal{P}_1(x)\right)E\left(w_1\mid w_2=0,z=1,x\right)\\-E\left[\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\right]\beta_0=0.\end{aligned}\tag{15}$$

Or equivalently,

$$\begin{aligned}EE\left(\left(1-w_2\right)w_1\mid z=0,x\right)-EE\left(\left(1-w_2\right)w_1\mid z=1,x\right)\\-\left(EE\left(w_2\mid z=1,x\right)-EE\left(w_2\mid z=0,x\right)\right)\beta_0=0.\end{aligned}\tag{16}$$

Combining (14) and (16), $\beta=\beta_1-\beta_0$ is defined through

$$\begin{aligned}EE\left(w_1\mid z=1,x\right)-EE\left(w_1\mid z=0,x\right)\\-\left(EE\left(w_2\mid z=1,x\right)-EE\left(w_2\mid z=0,x\right)\right)\beta_0=0.\end{aligned}\tag{17}$$

Invoking Lemma 1 immediately produces the efficient influence function for $\beta_1-\beta_0$:

$$\begin{aligned}\frac{1}{P\left(D_1>D_0\right)}\Bigg\{&\frac{z}{\mathcal{Q}\left(x\right)}\left(w_1-E\left(w_1|z=1,x\right)\right)+E\left(w_1|z=1,x\right)\\&-\frac{1-z}{1-\mathcal{Q}\left(x\right)}\left(w_1-E\left(w_1|z=0,x\right)\right)-E\left(w_1|z=0,x\right)\\&-\Bigg(\frac{z}{\mathcal{Q}\left(x\right)}\left(w_2-E\left(w_2|z=1,x\right)\right)+E\left(w_2|z=1,x\right)\\&\qquad-\frac{1-z}{1-\mathcal{Q}\left(x\right)}\left(w_2-E\left(w_2|z=0,x\right)\right)-E\left(w_2|z=0,x\right)\Bigg)\left(\beta_1-\beta_0\right)\Bigg\}.\end{aligned}$$

Part 2:(ATT on compliers) We also discuss γ_1 and γ_0 in turn. Consider first

$$\begin{aligned}\gamma_1=&E\left(y_1|w_2=1,D_1>D_0\right)=E\left(w_1|w_2=1,D_1>D_0\right)\\=&\int E\left(w_1|w_2=1,D_1>D_0,x\right)f\left(x|w_2=1,D_1>D_0\right)dx\end{aligned}$$

Note that the above conditional density can be written as:

$$\begin{aligned} f\left(x|w_2=1,D_1>D_0\right)dx =& \frac{f\left(x,w_2=1,D_1>D_0\right)}{P\left(w_2=1,D_1>D_0\right)} \\ =& \frac{f\left(x\right)Q\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)}{EQ\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)}. \end{aligned}$$

So that γ_1 is defined by the moment condition:

$$\begin{aligned} EE\left(w_1|w_2=1,D_1>D_0,x\right)\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right) \\ -\left(E\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\right)\gamma_1=0. \end{aligned}$$

Using the identification result that

$$\begin{aligned} &E\left(w_1 \mid w_2=1,d_1>d_0,\,x\right) \\ &=\frac{\mathcal{P}_1(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1 \mid w_2=1,z=1,x\right)-\frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1 \mid w_2=1,z=0,x\right). \end{aligned}$$

The moment condition that defines γ_1 can be rewritten as

$$\begin{aligned} E\mathcal{Q}\left(x\right)\mathcal{P}_1\left(x\right)E\left(w_1|w_2=1,z=1,x\right)-E\mathcal{Q}\left(x\right)\mathcal{P}_0\left(x\right)E\left(w_1|w_2=1,z=0,x\right) \\ -E\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\gamma_1=0. \end{aligned}$$

This can be equivalently written as:

$$\begin{aligned} EE\left(w_1w_2|x\right)-EE\left(w_1w_2|z=0,x\right) \\ -\left(EE\left(w_2|x\right)-EE\left(w_2|z=0,x\right)\right)\gamma_1=0. \end{aligned}$$

Now consider the analogous derivation for γ_0.

$$\begin{aligned} \gamma_0 =& E\left[y_0|w_2=1,D_1>D_0\right]=\int E\left[y_0|w_2=1,D_1>D_0,x\right]f\left(x|w_2=1,D_1>D_0\right)dx \\ =& \frac{E\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)E\left[w_1|w_2=0,D_1>D_0,x\right]}{E\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)} \end{aligned}$$

So that γ_0 is defined by the moment condition:

$$\begin{aligned} E\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)E\left[w_1|w_2=0,D_1>D_0,x\right] \\ -E\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\gamma_0=0. \end{aligned}$$

Using again the identification condition that

$$\begin{aligned} &E\left(w_1 \mid w_2=0,d_1>d_0,\,x\right) \\ &=\frac{(1-\mathcal{P}_0(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1 \mid w_2=0,z=0,x\right)-\frac{(1-\mathcal{P}_1(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E\left(w_1 \mid w_2=0,z=1,x\right). \end{aligned}$$

The moment condition for γ_0 can be manipulated to be:

$$\begin{aligned} &EE\left(w_1\left(1-w_2\right)|z=0,x\right)-EE\left(w_1\left(1-w_2\right)|x\right) \\ &\qquad -\left(EE\left(w_2|x\right)-EE\left(w_2|z=0,x\right)\right)\gamma_0=0. \end{aligned}$$

The moment condition for $\gamma=\gamma_1-\gamma_0$ therefore combines those of γ_1 and γ_0:

$$\begin{aligned} &EE\left(w_1|x\right)-EE\left(w_1|z=0,x\right) \\ &\qquad -\left(EE\left(w_2|x\right)-EE\left(w_2|z=0,x\right)\right)\gamma_1=0. \end{aligned}$$

Hence the efficient influence function for γ is given through lemma 1 by

$$\begin{aligned} &\frac{1}{P\left(w_2=1,D_1>D_0\right)}\Bigg\{w_1-\frac{1-z}{1-\mathcal{Q}\left(x\right)}\left(w_1-E\left(w_1|z=0,x\right)\right)-E\left(w_1|z=0,x\right) \\ &\qquad -\left(w_2-\frac{1-z}{1-\mathcal{Q}\left(x\right)}\left(w_2-E\left(w_2|z=0,x\right)\right)-E\left(w_2|z=0,x\right)\right)\left(\gamma_1-\gamma_0\right)\Bigg\}. \end{aligned}$$

C.2 Proof of theorem 5

Recall that the moment condition that defines γ is given by

$$\begin{aligned} &E\mathcal{Q}\left(x\right)\mathcal{P}_1\left(x\right)E\left(w_1|w_2=1,z=1,x\right) \\ &-E\mathcal{Q}\left(x\right)\mathcal{P}_0\left(x\right)E\left(w_1|w_2=1,z=0,x\right) \\ &-E\mathcal{Q}\left(x\right)\left(1-\mathcal{P}_0\left(x\right)\right)E\left(w_1\mid w_2=0,z=0,x\right) \\ &\qquad +E\mathcal{Q}\left(x\right)\left(1-\mathcal{P}_1\left(x\right)\right)E\left(w_1\mid w_2=0,z=1,x\right) \\ &\qquad -E\mathcal{Q}\left(x\right)\left(\mathcal{P}_1\left(x\right)-\mathcal{P}_0\left(x\right)\right)\gamma=0. \end{aligned}$$

which can be rewritten as

$$\begin{aligned} h\left(\mathcal{Q}\left(x\right)\right)\equiv&E\mathcal{Q}\left(x\right)E\left(w_1|z=1,x\right)-E\mathcal{Q}\left(x\right)E\left(w_1|z=0,x\right) \\ &-E\mathcal{Q}\left(x\right)\left(E\left(w_2=1|z=1,x\right)-E\left(w_2=1|z=0,x\right)\right)\gamma. \end{aligned}$$

When $\mathcal{Q}\left(x\right)$ is known, the efficient projection into the tangent space obviously follows immediately from Lemma 1:

$$\begin{aligned} &z\left(w_1-E\left(w_1|z_1=1,x\right)\right)+\mathcal{Q}\left(x\right)E\left(w_1|z_1=1,x\right) \\ &\quad -\frac{1-z}{1-\mathcal{Q}\left(x\right)}Q\left(x\right)\left[w_1-E\left(w_1|z=0,x\right)\right]-\mathcal{Q}\left(x\right)E\left(w_1|z=0,x\right) \\ &\quad -\Bigg\{z\left(w_2-E\left(w_2|z_1=1,x\right)\right)+\mathcal{Q}\left(x\right)E\left(w_2|z_1=1,x\right) \\ &\quad -\frac{1-z}{1-\mathcal{Q}\left(x\right)}Q\left(x\right)\left[w_2-E\left(w_2|z=0,x\right)\right]-\mathcal{Q}\left(x\right)E\left(w_2|z=0,x\right)\Bigg\}\gamma. \end{aligned}$$

When $\mathcal{Q}(x)$ is known up to a finite dimensional parameter α, in addition to above, there is an extra term for the efficient influence function due to the parametric Cramer-Rao lower bound for the estimation of $\hat{\alpha} - \alpha$, as in

$$h\left(\mathcal{Q}_{\hat{\alpha}}(x)\right) - h\left(\mathcal{Q}(x)\right).$$

It is easy to see using the parametric Delta method that this additional term is given by

$$\begin{aligned}
&\text{Proj}\left[\left(z - \mathcal{Q}(x)\right)\kappa(x) \mid S_\alpha(z;x)\right] \\
&= E\left(\kappa(x)\frac{\partial \mathcal{Q}}{\partial \alpha}(x,\alpha)\right)\left[ES_\alpha(z;x)\,S_\alpha(z;x)'\right]^{-1} S_\alpha(z;x).
\end{aligned}$$

C.3 Proof of theorem 6

Denote $g(w, x\beta) = w_2 g_1(w_1, x, \beta) - (1 - w_2)\, g_0(w_1, x, \beta)$ and introduce the weighting matrix

$$\mathcal{A}(w_2, x) = \left(\mathcal{P}_1(x) - \mathcal{P}_0(x)\right)\left(\frac{\mathcal{Q}(x)w_2}{\mathbf{P}(x)} + \frac{\left(1 - \mathcal{Q}(x)\right)(1 - w_2)}{1 - \mathbf{P}(x)}\right) A,$$

and a moment function $\tilde{g} = A\, g(w, x, \beta)$. Also define

$$\overline{g}(w, x, \beta) = g(w, x, \beta) - E\left[g(w, x, \beta) \mid w_2, x, D_1 > D_0\right].$$

Given that the conditional equation (9) is valid, we have:

$$E\left\{\mathcal{A}(w_2, x)\, E\left\{g(w, x, \beta)\,\middle|\, w_2, x, D_1 > D_0\right\}\right\} = 0.$$

As a result we can express the directional derivative of the Euclidean parameter in the same way as before

$$J\frac{\partial \beta(\theta)}{\partial \theta} = -\frac{\partial}{\partial \theta} E_\theta\left[\mathcal{A}_\theta(w_2,\, x)\, E_\theta\left[g(w, x,\, \beta)\,|w_2, x, D_1 > D_0\right]\right].$$

Note that the difference between this formula and the formula for the conditional moment equation is that the weighting function $\mathcal{A}(w_2, x)$ depends on the parametrization path. The derivative of the right-hand side will contain three components:

$$\begin{aligned}
&E\left[\mathcal{A}(w_2, x)\int \overline{g}(w, x, \beta)\, s_{**}(w_1 \mid w_2,\, x)\, f_{**}(w_1 \mid w_2,\, x)\; dw_1\right] \\
&+E\left[\mathcal{A}(w_2, x)\, s_\theta(w_2, x)\int g(w, x, \beta)\; f_{**}(w_1 \mid w_2,\, x)\; dw_1\right] + E\left[\tfrac{\partial \mathcal{A}_\theta(w_2, x)}{\partial \theta}\int g(w, x, \beta)\; f_{**}(w_1 \mid w_2,\, x)\; dw_1\right].
\end{aligned}$$

The first and the second components will have the same structure as in the conditional moment equation case. The first component multiplied by the Jacobi matrix can be written as:

$$-J^{-1}\Phi_1(w,x,z) = \tfrac{(z-\mathcal{Q}(x))(w_2+\mathcal{Q}(x)-1)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))} A(w_2)\, g(w,x,\beta) - \tfrac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}$$
$$\times A(w_2)\left[E[w_2 z g \mid w_2=1, z=1, x] - E[(1-w_2)(1-z)\, g \mid w_2=0, z=0, x]\right] - \tfrac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))} A(w_2)$$
$$\times\Big[(w_2-\mathcal{P}_1(x))\, E[w_2 g \mid w_2=1, x, D_1>D_0] + (w_2-\mathcal{P}_0(x))\, E[(1-w_2)\, g \mid w_2=0, x, D_1>D_0]\Big].$$

To derive the second component of the influence function, note that it should solve:

$$E[\Phi_2(w,x,z)\, S_\theta(w,x,z)] = E\left[\mathcal{A}(w_2,x)\, s_\theta(w_2,x) \int g(w,x,\beta)\; f_{**}(w_1 \mid w_2,\, x)\; dw_1\right]$$
$$+E\left[\tfrac{\partial \mathcal{A}_\theta(w_2,x)}{\partial \theta} \int g(w,x,\beta)\; f_{**}(w_1 \mid w_2,\, x)\; dw_1\right].$$

Note that:

$$E[\mathcal{A}(w_2,x)\, s_\theta(x,w_2) E[g \mid w_2, x, D_1>D_0]] = E[s_\theta(x) E[\mathcal{A}(w_2,x)\, g(w,x,\beta) \mid w_2, x, D_1>D_0]]$$
$$+E\left[\left(E[\mathcal{A}(w_2,x)\, g \mid w_2=1, x, D_1>D_0] - E[\mathcal{A}(w_2,x)\, g \mid w_2=0, x, D_1>D_0]\right)\dot{\mathbf{P}}_\theta(x)\right]$$
$$= E\Bigg[A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\Bigg(\mathcal{Q}(x) E[g \mid w_2=1, x, D_1>D_0]$$
$$+(1-\mathcal{Q}(x))\, E[g \mid w_2=0, x, D_1>D_0]\Bigg) s_\theta(x)\Bigg]$$
$$+E\left[A(w_2)\,(\mathcal{P}_1(x)-\mathcal{P}_0(x))\left\{\tfrac{\mathcal{Q}(x)E[g|w_2=1,x,D_1>D_0]}{\mathbf{P}(x)} - \tfrac{(1-\mathcal{Q}(x))E[g|w_2=0,x,D_1>D_0]}{1-\mathbf{P}(x)}\right\}\dot{\mathbf{P}}_\theta(x)\right].$$

Next we can express the directional derivative:

$$\tfrac{\partial \mathcal{A}_\theta(w_2,x)}{\partial \theta} = \mathcal{A}(w_2,x)\, \tfrac{\dot{\mathcal{P}}_{1\theta}(x)-\dot{\mathcal{P}}_{0\theta}(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)} + A \tfrac{(\mathcal{P}_1(x)-\mathcal{P}_0(x))(w_2-\mathbf{P}(x))\dot{\mathcal{Q}}(x)}{\mathbf{P}(x)(1-\mathbf{P}(x))}$$
$$+A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\,\dot{\mathbf{P}}(x)\left[-\tfrac{\mathcal{Q}(x)w_2}{\mathbf{P}^2(x)} + \tfrac{(1-\mathcal{Q}(x))(1-w_2)}{(1-\mathbf{P}(x))^2}\right].$$

Consider expression:

$$E\left[\tfrac{\partial \mathcal{A}_\theta(w_2,x)}{\partial \theta} \int g(w,x,\beta)\; f_{**}(w_1 \mid w_2,\, x)\; dw_1\right]$$
$$= E\Bigg[A\Bigg(\mathcal{Q}(x) E[g \mid w_2=1, x, D_1>D_0] + (1-\mathcal{Q}(x))\, E[g \mid w_2=0, x, D_1>D_0]\Bigg)$$
$$\times\left(\dot{\mathcal{P}}_{1\theta}(x) - \dot{\mathcal{P}}_{0\theta}(x)\right)\Bigg]$$
$$+E\left[A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\left(E[g \mid w_2=1, x, D_1>D_0] - E[g \mid w_2=0, x, D_1>D_0]\right)\dot{\mathcal{Q}}(x)\right]$$
$$+E\left[A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\left\{-\tfrac{\mathcal{Q}(x)E[g|w_2=1,x,D_1>D_0]}{\mathbf{P}(x)} + \tfrac{(1-\mathcal{Q}(x))E[g|w_2=0,x,D_1>D_0]}{1-\mathbf{P}(x)}\right\}\dot{\mathbf{P}}_\theta(x)\right]$$

Finally, we can write the expression for the second component of the directional derivative of the parameter vector as:

$$
\begin{aligned}
&E\left[\Phi_2\left(w,x,z\right)S_\theta\left(w,x,z\right)\right]\\
&=-J^{-1}E\Bigg[A\Bigg(\mathcal{Q}(x)E\left[g\mid w_2=1,x,D_1>D_0\right]+\left(1-\mathcal{Q}(x)\right)E\left[g\mid w_2=0,x,D_1>D_0\right]\Bigg)\\
&\qquad\qquad\times\Big(\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)s_\theta(x)+\dot{\mathcal{P}}_{1\theta}(x)-\dot{\mathcal{P}}_{0\theta}(x)\Big)\Bigg]\\
&+E\Big[A\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)\left(E\left[g\mid w_2=1,x,D_1>D_0\right]-E\left[g\mid w_2=0,x,D_1>D_0\right]\right)\dot{\mathcal{Q}}(x)\Big].
\end{aligned}
$$

This means that the second component of the efficient influence function takes the form:

$$
\begin{aligned}
-J\,\Phi_2\left(w,x,z,\right)&=A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\Bigg(\mathcal{Q}(x)E\left[g\mid w_2=1,x,D_1>D_0\right]\\
&\qquad+\left(1-\mathcal{Q}(x)\right)E\left[g\mid w_2=0,x,D_1>D_0\right]\Bigg)\left[1+\tfrac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\,\tfrac{w_2-\mathcal{F}(z,x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}\right]\\
&\quad+A\left(\mathcal{P}_1(x)-\mathcal{P}_0(x)\right)\left(E\left[g\mid w_2=1,x,D_1>D_0\right]-E\left[g\mid w_2=0,x,D_1>D_0\right]\right)\left(z-\mathcal{Q}(x)\right).
\end{aligned}
$$

Combining two components of the efficiency bound, we obtain for $\tilde{g}=-J^{-1}A\,g\left(w,x,\beta\right)$:

$$
\begin{aligned}
\Phi\left(w,x,z\right)&=\left\{w_2z-\tfrac{\mathcal{Q}(x)w_2(1-z)}{1-\mathcal{Q}(x)}+\left(1-w_2\right)\left(1-z\right)-\tfrac{(1-\mathcal{Q}(x))(1-w_2)z}{\mathcal{Q}(x)}\right\}\tilde{g}\left(w,x,\beta\right)\\
&+\tfrac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\Big[\left(1-\mathcal{P}_1(x)\right)\left(1-\mathcal{Q}(x)\right)E\left[\tilde{g}\mid w_2=0,z=1,x\right]-\mathcal{P}_0(x)\mathcal{Q}(x)E\left[\tilde{g}\mid w_2=1,z=0,x\right]\Big].
\end{aligned}
$$

D Efficiency bound under semiparametric restrictions

Consider the semiparametric conditional moment equation:

$$
\varphi\left(x,w_2,\mu(x),\beta\right)=E\left[g\left(w,x,\mu(x),\beta\right)\mid w_2,x,d_1>d_0\right]=0, \tag{18}
$$

where $g\left(\cdot\right)$ is a known function $g:\mathbb{R}\times\{0,1\}\times\mathcal{X}\times\mathbb{R}\times\mathcal{B}\mapsto\mathbb{R}$, and $\mu(\cdot)$ is some unknown function of x (which needs to be estimated along with β). The presence of additional semiparametric component in the model expectedly increases the efficiency bound. However, the presence of this component does not change the general structure of the efficiency bound and the structure of the optimal instrument. The intuition for this result is that the efficiency bound is, in general, determined by the projection of the parametric part of the score of the model on the tangent set of the model. Thus an extra semiparametric component of the moment equation will not change the parametric score, but it will change its projection. In the following theorem we establish the structure of the semiparametric efficiency bound and the optimal instrument for model (18).

Theorem 7 *Under Assumption 1 the semiparametric efficiency bound for a finite-dimensional parameter β characterizing the treatment effect for the subsample of compliers with $P(D_1 > D_0) = 1$ can be expressed as:*

$$V(\beta) = E\left(E\left[\frac{\partial \varphi(w_2, x, \beta)}{\partial \beta} \widetilde{\zeta}(x, w_2)' \middle| x \right] \widetilde{\Omega}(x)^{-1} E\left[\widetilde{\zeta}(x, w_2) \frac{\partial \varphi(w_2, x, \beta')}{\partial \beta} \middle| x \right] \right)^{-1}.$$

In this theorem we use notations:

$$\widetilde{\zeta}(w_2, x) = \zeta(w_2, x) - \left\{ \frac{\partial \varphi}{\partial \mu} \right\}^{-1} E\left[\zeta(w_2, x) \frac{\partial \varphi}{\partial \mu} \middle| x \right],$$

$$\widetilde{\Omega}(x) = E\left[\Psi(w, x, z) \Psi(w, x, z)' \right],$$

and

$$\begin{aligned} \Psi(w, z, x) = \Big(w_2 z - \tfrac{\mathcal{Q}(x)}{1-\mathcal{Q}(x)} w_2 (1 - z) + (1 - w_2)(1 - z) - \tfrac{1-\mathcal{Q}(x)}{\mathcal{Q}(x)} (1 - w_2) z \Big) \widetilde{\zeta}(w_2, x) g(w, x, \beta) \\ - \frac{(z - \mathcal{Q}(x))}{\mathcal{Q}(x)(1 - \mathcal{Q}(x))} \Big\{ E\left[w_2 z \widetilde{\zeta}(w_2, x) g \mid w_2 = 1, z = 1, x \right] \\ - E\left[(1 - w_2)(1 - z) \widetilde{\zeta}(w_2, x) g \mid w_0 = 1, z = 0, x \right] \Big\}. \end{aligned}$$

Proof:

In the following we will use the result for the efficiency bound obtained for the case $\varphi(x, w_2, \beta)$. Let us consider a specific parametric path θ for the model and differentiate the conditional moment equation with respect to this parametric path.

$$\frac{\partial \varphi}{\partial \beta'} \dot{\beta} + \frac{\partial \varphi}{\partial \mu} \dot{\mu}(x) + \int g(w, x, \mu(x), \beta) s_{**}(w_1 \mid w_2, x) f_{**}(w_1 \mid w_2, x) \, dw_1 = 0.$$

This allows us to express the (scalar) directional derivative $\dot{\mu}$ in terms of the directional derivative of the finite-dimensional parameter of interest and the integral over $g(\cdot)$. Then consider a transformation of the conditional moment equation into the system of unconditional moments. This transformation will have the same structure as in case of only one finite-dimensional parameter. In fact, if we impose the restriction that $\mu(x) \in \mathbf{L}^2\left(\mathbb{R}^k\right)$, then the function $g(\cdot)$ will still belong to $\mathbf{L}^2$ as a function of x and w_2 for each β. Differentiating the set of unconditional moments along the parametrization path we evaluate the relevant gradients at $\theta = 0$ which leads to:

$$\begin{aligned} J\dot{\beta} + E\left[\mathcal{A}(w_2, x) \tfrac{\partial \varphi(w_2, x, \mu(x), \beta)}{\partial \mu} \dot{\mu}(x) \right] \\ + E\left[\mathcal{A}(w_2, x) \textstyle\int g(w, x, \mu(x), \beta) s_{**}(w_1 \mid w_2, x) f_{**}(w_1 \mid w_2, x) \, dw_1 \right] = 0. \end{aligned} \tag{19}$$

Now denote $\lambda(x) = E\left[\mathcal{A}(w_2, x)\,\frac{\partial\varphi}{\partial\mu} \mid x\right]$ then (19) can be rewritten as:

$$J\dot{\beta} + E\left[\lambda(x)\dot{\mu}(x)\right] + E\left[\mathcal{A}(w_2, x)\int g(w, x, \mu(x), \beta)\, s_{**}(w_1 \mid w_2,\, x)\, f_{**}(w_1 \mid w_2,\, x)\, dw_1\right] = 0.$$

Then we can substitute the expression for $\dot{\mu}(x)$ obtained from the directional derivative of the conditional moment which transforms the expression (19) into:

$$E\left[\left(\mathcal{A}(w_2, x) - \left\{\tfrac{\partial\varphi}{\partial\mu}\right\}^{-1}\lambda(x)\right)\tfrac{\partial\varphi}{\partial\beta'}\right]\dot{\beta}$$
$$+E\left[\left(\mathcal{A}(w_2, x) - \left\{\tfrac{\partial\varphi}{\partial\mu}\right\}^{-1}\lambda(x)\right)\int g(w, x, \mu(x), \beta)\, s_{**}(w_1 \mid w_2,\, x)\, f_{**}(w_1 \mid w_2,\, x)\, dw_1\right] = 0.$$

All further manipulations are identical to the completely parametric case and are based on finding the efficient influence function, which by Newey's argument can be found from the representation:

$$\dot{\beta} = E\left[\Psi\, S_\theta\right].$$

The structure of the efficient influence function will be the same as for the fully parametric case. Semiparametric efficiency bound and the optimal instrument, however, will be different. In particular denote

$$\widetilde{\zeta}(w_2, x) = \zeta(w_2, x) - \left\{\frac{\partial\varphi}{\partial\mu}\right\}^{-1} E\left[\zeta(w_2, x)\,\frac{\partial\varphi}{\partial\mu}\middle| x\right].$$

Then the semiparametric efficiency bound is obtained from the semiparametric efficiency bound for the model only parametrized by β by substituting $\widetilde{\zeta}(w_2, x)$ for $\zeta(w_2, x)$. It will have the form:

$$V(\beta) = E\left(E\left[\frac{\partial\varphi(w_2, x, \beta)}{\partial\beta}\widetilde{\zeta}(x, w_2)'\middle| x\right]\widetilde{\Omega}(x)^{-1}E\left[\widetilde{\zeta}(x, w_2)\frac{\partial\varphi(w_2, x, \beta')}{\partial\beta}\middle| x\right]\right)^{-1}.$$

The matrix $\Omega(x)$ can be obtained from the similar matrix for the parametric moment condition by substituting $\zeta(w_2, x)$ with $\widetilde{\zeta}(w_2, x)$. It can then be expressed as:

$$\widetilde{\Omega}(x) = E\left[\Psi(w, x, z)\,\Psi(w, x, z)'\right],$$

with efficient influence function $\Psi(\cdot)$ which structure does not change. Note that the semiparametric efficiency bound for the parameter in the conditional moment equation $\varphi(w_2, x, \mu(x), \beta) = 0$ will be above that for the conditional moment equation $\varphi(w_2, x, 0, \beta) = 0$.