Supplement to "Semiparametric efficiency in nonlinear LATE models"

(*Quantitative Economics*, Vol. 1, No. 2, November 2010, 279–304)

Han Hong
Departments of Economics, Stanford University

Denis Nekipelov
Departments of Economics, University of California at Berkeley

Appendix

A. *Proof of Theorem 1*

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

$$f^1(y|d,z,x) = f(y_1 = y|d,z,x)$$

and

$$f^0(y|d,z,x) = f(y_0 = y|d,z,x).$$

If $\phi_\theta(x)$ is the Radon–Nikodym density of x with support on $\mathcal{X}$, the likelihood function for the data can be written as

$$\begin{aligned} f_\theta(w,z,x) = {} & [f_\theta^1(y|d,z,x)]^d [f_\theta^0(y|d,z,x)]^{(1-d)} \mathcal{F}_\theta^d(x,z)(1-\mathcal{F}_\theta(x,z))^{(1-d)} \\ & \times \mathcal{Q}_\theta^z(x)(1-\mathcal{Q}_\theta(x))^{(1-z)} \phi_\theta(x). \end{aligned}$$

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

$$\begin{aligned} S_\theta(w,z,x) = {} & (1-d)s_\theta^0(y|d,z,x) + d s_\theta^1(y|d,z,x) \\ & + \frac{(1-z)\dot{\mathcal{P}}_{0\theta}(x)}{\mathcal{F}(z,x)(1-\mathcal{F}(z,x))}[d-\mathcal{F}(z,x)] \\ & + \frac{z\dot{\mathcal{P}}_{1\theta}(x)}{\mathcal{F}(z,x)(1-\mathcal{F}(z,x))}[d-\mathcal{F}(z,x)] \\ & + \frac{\dot{\mathcal{Q}}_\theta(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}[z-\mathcal{Q}(x)] + s_\theta(x), \end{aligned}$$

Han Hong: hanhong@stanford.edu
Denis Nekipelov: nekipelov@econ.berkeley.edu

DOI: 10.3982/QE43

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

$$\mathcal{T} = \big\{(1-d)s^0_\theta(y|d,z,x) + ds^1_\theta(y|d,z,x) + z\xi(x,z)[d-\mathcal{F}(z,x)] \\ + (1-z)\zeta(x,z)[d-\mathcal{F}(z,x)] + a(x)[z-\mathcal{Q}(x)] + t(x)\big\},$$

where $E_\theta[s^i_\theta(y|d,z,x)|d,z,x] = 0$ 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 $\varphi(x,d,\beta)$. We assume that the support of x—the set $\mathcal{X}$—is nondegenerate. 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 $\mathcal{A}(d,x)$ is an arbitrary continuous vector function such that $\mathcal{A}\colon\{0,1\}\times\mathcal{X}\mapsto\mathbb{R}^k$. Without loss of generality, we can define this vector function by a vector

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

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

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

and

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

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

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

This can be rewritten as $E_\theta[\mathcal{A}_\theta(d,x)E_\theta[g(y,d,x,\beta(\theta))|d,x,d_1>d_0]]=0$.

Define the Jacobi matrix

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

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

$$J\frac{\partial\beta(\theta)}{\partial\theta} = -\frac{\partial}{\partial\theta}E_\theta\big[\mathcal{A}_\theta(d,x)E_\theta[g(y,d,x,\beta)|d,x,d_1>d_0]\big]. \tag{12}$$

The right-hand side component of equation (12) can be written as

$$
\begin{aligned}
&E\Big[\mathcal{A}(d,x)\int g(y,d,x,\beta)s_{**}(y|d,x)f_{**}(y|d,x)\,dy\Big] \\
&\quad + E\Big[\mathcal{A}(d,x)s_\theta(d,x)\int g(y,d,x,\beta)f_{**}(y|d,x)\,dy\Big] \\
&\quad + E\Big[\frac{\partial\mathcal{A}_\theta(d,x)}{\partial\theta}\int g(w,x,\beta)f_{**}(y|d,x)\,dy\Big].
\end{aligned} \tag{13}
$$

In the above expression, we have used the definition that

$$s_{**}(y|d,x) = \frac{\partial}{\partial\theta}\log f_{**}^{\theta}(y|d,x),$$

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

$$\int g(w,x)f_{**}(y|d,x)\,dy = 0,$$

and the last two terms of equation (13) can be removed. Using this result, the system for the directional derivative of the parameter vector can be rewritten as

$$\frac{\partial\beta(\theta)}{\partial\theta} = -J^{-1}E\Big[\mathcal{A}(d,x)\int g(y,d,x,\beta)s_{**}(y|d,x)f_{**}(y|d,x)\,dy\Big].$$

Next we introduce the notations

$$
\begin{aligned}
\tilde{g}(w,x,\beta) &= -J^{-1}A(d,x)g(w,x,\beta), \\
\hat{g}(w,x,\beta) &= (\mathcal{P}_1(x)-\mathcal{P}_0(x))\left(\frac{\mathcal{Q}(x)d}{\mathbf{P}(x)} + \frac{(1-\mathcal{Q}(x))(1-d)}{1-\mathbf{P}(x)}\right)\tilde{g}(w,x,\beta),
\end{aligned}
$$

and

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

Recall the definition that

$$\Delta(x,d,\beta) = \big\{E[g(y,d,x,\beta)|d,z=d,x] - E[g(y,d,x,\beta)|d,z=1-d,x]\big\}.$$

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

$$
\begin{aligned}
&s_{**}(y|d,x)f_{**}(y|d,x) \\
&\quad = \frac{d\mathcal{P}_1(x)+(1-d)(1-\mathcal{P}_0(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}s_\theta(y|d,z=d,x)f(y|d,z=d,x) \\
&\qquad - \frac{d\mathcal{P}_0(x)+(1-d)(1-\mathcal{P}_1(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}s_\theta(y|d,z=1-d,x)f(y|d,z=1-d,x)
\end{aligned}
$$

$$+\left[\frac{[1-d+\mathcal{P}_1(x)(2d-1)]\dot{\mathcal{P}}_{0\theta}(x)-[1-d+\mathcal{P}_0(x)(2d-1)]\dot{\mathcal{P}}_{1\theta}(x)}{(\mathcal{P}_1(x)-\mathcal{P}_0(x))^2}\right]$$
$$\times (f(y|d,z=d,x)-f(y|d,z=1-d,x)).$$

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

$$\begin{aligned}
&\Psi(w,z,x)\\
&\quad=\frac{\mathbf{P}(x)dz}{\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))}\big(\hat{g}(y,d,x,\beta)-E[\hat{g}(y,d,x,\beta)|d=1,z=1,x]\big)\\
&\qquad-\frac{\mathbf{P}(x)d(1-z)}{(1-\mathcal{Q}(x))(\mathcal{P}_1(x)-\mathcal{P}_0(x))}\\
&\qquad\times\big(\hat{g}(y,d,x,\beta)-E[\hat{g}(y,d,x,\beta)|d=1,z=0,x]\big)\\
&\qquad+\frac{(1-\mathbf{P}(x))(1-d)(1-z)}{(1-\mathcal{Q}(x))(\mathcal{P}_1(x)-\mathcal{P}_0(x))}\\
&\qquad\times\big(\hat{g}(y,d,x,\beta)-E[\hat{g}(y,d,x,\beta)|d=0,z=0,x]\big)\\
&\qquad-\frac{(1-\mathbf{P}(x))(1-d)z}{\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))}\big(\hat{g}(y,d,x,\beta)-E[\hat{g}(y,d,x,\beta)|d=0,z=1,x]\big)\\
&\qquad+\frac{\hat{\Delta}(d=1,x)\mathbf{P}(x)}{(\mathcal{P}_1(x)-\mathcal{P}_0(x))^2}\left[\frac{\mathcal{P}_1(x)(1-z)}{1-\mathcal{Q}(x)}-\frac{\mathcal{P}_0(x)z}{\mathcal{Q}(x)}\right](d-\mathcal{F}(z,x))\\
&\qquad+\frac{\hat{\Delta}(d=0,x)(1-\mathbf{P}(x))}{(\mathcal{P}_1(x)-\mathcal{P}_0(x))^2}\left[\frac{(1-\mathcal{P}_1(x))(1-z)}{1-\mathcal{Q}(x)}-\frac{(1-\mathcal{P}_0(x))z}{\mathcal{Q}(x)}\right](d-\mathcal{F}(z,x)).
\end{aligned}$$

The first two lines correspond to the $ds^1_\theta(y|d,z,x)$ component of $\mathcal{T}$. The third and fourth lines correspond to the $(1-d)s^0_\theta(y|d,z,x)$ component of $\mathcal{T}$. The last two lines correspond to the

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

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

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

$$\begin{aligned}
&\mathcal{P}_1(x)E[\hat{g}(Y,D,X,\beta)|D=1,Z=1,x]\\
&\quad=\mathcal{P}_0(x)E[\hat{g}(Y,D,X,\beta)|D=1,Z=0,x],
\end{aligned}\tag{14}$$

$$(1-\mathcal{P}_0(x))E[\hat{g}(Y,D,X,\beta)|D=0,Z=0,x]$$
$$=(1-\mathcal{P}_1(x))E[\hat{g}(Y,D,X,\beta)|D=0,Z=1,x].$$

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

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

We can now express the semiparametric efficiency bound as the variance of the efficient influence function $V(\hat{\beta})=E\{\Psi\Psi'\}$. Note that the vector $\mathcal{A}(d,x)$ can be represented as

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

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

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

In this case, the Jacobi matrix can be written as

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

To facilitate the manipulations, denote

$$\omega_{d,z}(x)=V(g(y,d,x,\beta)|d,z,x)$$

and

$$\gamma_{d,z}(x)=E(g(y,d,x,\beta)|d,z,x).$$

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

$$
\begin{aligned}
&V(\Psi_1(y,d,x,z))\\
&= J^{-1} E\left\{ \mathcal{M}(x) D(x) \begin{pmatrix} \frac{\mathcal{P}_1(x)\omega_{11}(x)}{\mathcal{Q}(x)} + \frac{\mathcal{P}_0(x)\omega_{10}(x)}{(1-\mathcal{Q}(x))} & 0 \\ + \frac{\mathcal{P}_1(x)\mathbf{P}(x)}{\mathcal{P}_0(x)\mathcal{Q}(x)(1-\mathcal{Q}(x))}\gamma_{11}^2(x) & \\ 0 & \frac{(1-\mathcal{P}_0(x))\omega_{00}(x)}{(1-\mathcal{Q}(x))} + \frac{(1-\mathcal{P}_1(x))\omega_{01}(x)}{\mathcal{Q}(x)} \\ & + \frac{(1-\mathcal{P}_0(x))(1-\mathbf{P}(x))}{(1-\mathcal{P}_1(x))\mathcal{Q}(x)(1-\mathcal{Q}(x))}\gamma_{00}^2(x) \end{pmatrix} \right.\\
&\qquad \left. \times D(x)\mathcal{M}(x)' \right\} J^{-1\prime}.
\end{aligned}
$$

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

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

The corresponding variance is

$$
\begin{aligned}
&V(\Psi_2(y,d,x,z))\\
&= J^{-1}E\left\{ \frac{\mathcal{M}(x)D(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))} \right.\\
&\quad \times \begin{pmatrix} \mathcal{P}_1^2(x)\gamma_{11}^2(x) & -\mathcal{P}_1(x)(1-\mathcal{P}_0(x))\gamma_{11}(x)\gamma_{00}(x) \\ -\mathcal{P}_1(x)(1-\mathcal{P}_0(x))\gamma_{11}(x)\gamma_{00}(x) & (1-\mathcal{P}_0(x))^2\gamma_{00}^2(x) \end{pmatrix}\\
&\quad \left. \times D(x)\mathcal{M}(x)' \right\} J^{-1\prime}.
\end{aligned}
$$

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

$$
\operatorname{Cov}(\Psi_1(y,d,x,z), \Psi_2(y,d,x,z)) = -V(\Psi_2(y,d,x,z)).
$$

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

$$
\begin{aligned}
V(\hat{\beta}) &= J^{-1}E\{\mathcal{M}(x)D(x)\bar{\Omega}(x)D(x)\mathcal{M}(x)'\}J^{-1\prime}\\
&= E\{(\mathcal{P}_1(x)-\mathcal{P}_0(x))\mathcal{M}(x)D(x)\theta(x)\}^{-1}\\
&\quad \times E\{\mathcal{M}(x)D(x)\bar{\Omega}(x)D(x)\mathcal{M}(x)'\}\\
&\quad \times E\{(\mathcal{P}_1(x)-\mathcal{P}_0(x))\mathcal{M}(x)D(x)\theta(x)\}^{-1\prime},
\end{aligned}
$$

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

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

$$\times \left[1 - \frac{\mathcal{P}_1(x)\mathcal{P}_0(x)}{\mathbf{P}(x)}\right]\Bigg),$$

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

and

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

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

$$V(\hat{\beta}) = E\{(\mathcal{P}_1(x) - \mathcal{P}_0(x))^2\theta(x)'\bar{\Omega}(x)^{-1}\theta(x)\}^{-1}.$$

B. *Proofs of Theorems 4 and 5*

The proof of Theorem 4 is self-evident. In particular, Assumption 2(ii)–(iv) combine to insure that, uniformly, the estimated $\hat{\mathcal{Q}}(x)$ and $\hat{A}(x,d)$ can be replaced by their true quantities. Assumption 2(v) insures that a law of large numbers uniform over β applies to $\frac{1}{N}\sum_{k=1}^{N}\psi_k(\beta, \mathcal{Q}_0, A_0)$.

Before verifying the regularity conditions of Theorem 5, we discuss the intuition of the asymptotic distribution. It is not difficult to see that estimating $\hat{\mathcal{M}}(x)$ has no impact on the asymptotic variance, because, for example, for all x,

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

Following the argument of Newey (1994), the asymptotic representation

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

holds. 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 $\Psi_1(y,d,x,z)$ and $\Psi_2(y,d,x,z)$, which correspond 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

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

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

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

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

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

This means, in particular, that

$$\hat{\psi}(\beta) = J[\Psi_1(y,d,x,z) - \Psi_2(y,d,x,z)] = J\Psi(y,d,x,z),$$

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

$$V(\hat{\beta}) = J^{-1}V(J\Psi(y,d,x,z))J^{-1} = V(\Psi(y,d,x,z)).$$

Thus, the estimator achieves the semiparametric efficiency bound.

Assumptions 3(i) and (ii), and 2(ii)–(iv) combine to insure stochastic equicontinuity, that is, conditions (3.2) and (3.3) in Theorem 3 of Chen, Linton, and Van Keilegom (2003) (CLK).

Assumption 3(v) is used to justify item (2.3)(ii) of Theorem 2 of CLK. Assumption 3(iii) and (iv) is used to justify items (2.3)(i) and (2.4) of Theorem 2 of CLK.

Assumption 4 basically ensures that

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

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

C. *Proofs of Theorems in Section 2.3*

The proofs of Theorems 2, 3, and 6 depend exclusively only on the following lemma, the proof of which can be found, for example, in Robins and Rotnitzky (1995) and Hahn (1998).

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

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

The rest of the proofs basically amount to rewriting the parameters of interest in collections of components that take the form $E(E(W|Z=a,X))$.

C.1. *Proof of Theorem 2*

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

$$\begin{aligned}\beta_1 &= E[E(y_1|d=1,x,D_1>D_0)|D_1>D_0] = E[E(y|d=1,x,D_1>D_0)|D_1>D_0]\\ &= E\big[(\mathcal{P}_1(x)-\mathcal{P}_0(x))E(y|d=1,x,D_1>D_0)\big]\frac{1}{P(D_1>D_0)},\end{aligned}$$

so that β_1 is defined by the moment condition

$$E\big[(\mathcal{P}_1(x)-\mathcal{P}_0(x))E(y|d=1,x,D_1>D_0)\big]-E\big[(\mathcal{P}_1(x)-\mathcal{P}_0(x))\big]\beta_1=0.$$

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

$$\begin{aligned}E(y|d=1,d_1>d_0,x) = {}&\frac{\mathcal{P}_1(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E(y|d=1,z=1,x)-\frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)}\\ &-\mathcal{P}_0(x)E(y|d=1,z=0,x).\end{aligned}$$

The moment condition that defines β_1 is then rewritten as

$$\begin{aligned}&E\mathcal{P}_1(x)E(y|d=1,z=1,x)-E\mathcal{P}_0(x)E(y|d=1,z=0,x)\\ &\quad-E\big[(\mathcal{P}_1(x)-\mathcal{P}_0(x))\big]\beta_1=0\end{aligned} \tag{15}$$

or, equivalently,

$$\begin{aligned}&EE(dy|z=1,x)-EE(dy|z=0,x)\\ &\quad-(EE(d|z=1,x)-EE(d|z=0,x))\beta_1=0.\end{aligned} \tag{16}$$

Similar calculations can be applied to β_0. Consider

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

This translates into the moment condition

$$E\big[(\mathcal{P}_1(x)-\mathcal{P}_0(x))E(y|d=0,x,D_1>D_0)\big]-E\big[(\mathcal{P}_1(x)-\mathcal{P}_0(x))\big]\beta_0=0.$$

Recall the related identification condition

$$\begin{aligned}E(y|d=0,d_1>d_0,x) = {}&\frac{(1-\mathcal{P}_0(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E(y|d=0,z=0,x)\\ &-\frac{(1-\mathcal{P}_1(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E(y|d=0,z=1,x).\end{aligned}$$

The relevant moment condition for β_0 is then rewritten as

$$\begin{aligned} &E(1-\mathcal{P}_0(x))E(y|d=0,z=0,x)-(1-\mathcal{P}_1(x))E(y|d=0,z=1,x) \\ &\quad -E\big[(\mathcal{P}_1(x)-\mathcal{P}_0(x))\big]\beta_0=0 \end{aligned} \tag{17}$$

or, equivalently,

$$\begin{aligned} &EE((1-d)y|z=0,x)-EE((1-d)y|z=1,x) \\ &\quad -(EE(d|z=1,x)-EE(d|z=0,x))\beta_0=0. \end{aligned} \tag{18}$$

Combining (16) and (18), $\beta=\beta_1-\beta_0$ is defined through

$$\begin{aligned} &EE(y|z=1,x)-EE(y|z=0,x) \\ &\quad -(EE(d|z=1,x)-EE(d|z=0,x))\beta_0=0. \end{aligned} \tag{19}$$

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

$$\begin{aligned} &\frac{1}{P(D_1>D_0)}\bigg\{\frac{z}{\mathcal{Q}(x)}(y-E(y|z=1,x))+E(y|z=1,x) \\ &\quad -\frac{1-z}{1-\mathcal{Q}(x)}(y-E(y|z=0,x))-E(y|z=0,x) \\ &\quad -\bigg(\frac{z}{\mathcal{Q}(x)}(d-E(d|z=1,x))+E(d|z=1,x) \\ &\quad -\frac{1-z}{1-\mathcal{Q}(x)}(d-E(d|z=0,x))-E(d|z=0,x)\bigg)(\beta_1-\beta_0)\bigg\}. \end{aligned}$$

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

$$\begin{aligned} \gamma_1 &= E(y_1|d=1,D_1>D_0)=E(y|d=1,D_1>D_0) \\ &= \int E(y|d=1,D_1>D_0,x)f(x|d=1,D_1>D_0)\,dx. \end{aligned}$$

Note that the above conditional density can be written as

$$f(x|d=1,D_1>D_0)\,dx=\frac{f(x,d=1,D_1>D_0)}{P(d=1,D_1>D_0)}=\frac{f(x)Q(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))}{EQ(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))},$$

so that γ_1 is defined by the moment condition

$$EE(y|d=1,D_1>D_0,x)\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))-\big(E\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))\big)\gamma_1=0.$$

Using the identification result that

$$\begin{aligned} E(y|d=1,d_1>d_0,x) &= \frac{\mathcal{P}_1(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E(y|d=1,z=1,x) \\ &\quad -\frac{\mathcal{P}_0(x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E(y|d=1,z=0,x), \end{aligned}$$

the moment condition that defines γ_1 can be rewritten as

$$\begin{aligned} &E\mathcal{Q}(x)\mathcal{P}_1(x)E(y|d=1,z=1,x) - E\mathcal{Q}(x)\mathcal{P}_0(x)E(y|d=1,z=0,x) \\ &\quad - E\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))\gamma_1 = 0. \end{aligned}$$

This can be equivalently written as

$$EE(yd|x) - EE(yd|z=0,x) - (EE(d|x) - EE(d|z=0,x))\gamma_1 = 0.$$

Now consider the analogous derivation for γ_0,

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

so that γ_0 is defined by the moment condition

$$E\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))E[y|d=0,D_1>D_0,x] - E\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))\gamma_0 = 0.$$

Again using the identification condition that

$$\begin{aligned} E(y|d=0,d_1>d_0,x) = &\frac{(1-\mathcal{P}_0(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E(y|d=0,z=0,x) \\ &- \frac{(1-\mathcal{P}_1(x))}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}E(y|d=0,z=1,x), \end{aligned}$$

the moment condition for γ_0 can be manipulated to be

$$EE(y(1-d)|z=0,x) - EE(y(1-d)|x) - (EE(d|x) - EE(d|z=0,x))\gamma_0 = 0.$$

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

$$EE(y|x) - EE(y|z=0,x) - (EE(d|x) - EE(d|z=0,x))\gamma_1 = 0.$$

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

$$\begin{aligned} &\frac{1}{P(d=1,D_1>D_0)}\Big\{y - \frac{1-z}{1-\mathcal{Q}(x)}(y-E(y|z=0,x)) - E(y|z=0,x) \\ &\quad - \left(d - \frac{1-z}{1-\mathcal{Q}(x)}(d-E(d|z=0,x)) - E(d|z=0,x)\right)(\gamma_1-\gamma_0)\Big\}. \end{aligned}$$

C.2. *Proof of Theorem 3*

Recall that the moment condition that defines γ is given by

$$\begin{aligned} &E\mathcal{Q}(x)\mathcal{P}_1(x)E(y|d=1,z=1,x) - E\mathcal{Q}(x)\mathcal{P}_0(x)E(y|d=1,z=0,x) \\ &\quad - E\mathcal{Q}(x)(1-\mathcal{P}_0(x))E(y|d=0,z=0,x) \\ &\quad + E\mathcal{Q}(x)(1-\mathcal{P}_1(x))E(y|d=0,z=1,x) \\ &\quad - E\mathcal{Q}(x)(\mathcal{P}_1(x)-\mathcal{P}_0(x))\gamma = 0, \end{aligned}$$

which can be rewritten as

$$h(\mathcal{Q}(x)) \equiv E\mathcal{Q}(x)E(y|z=1,x) - E\mathcal{Q}(x)E(y|z=0,x) \\ - E\mathcal{Q}(x)(E(d=1|z=1,x) - E(d=1|z=0,x))\gamma.$$

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

$$\begin{aligned} &z(y - E(y|z_1=1,x)) + \mathcal{Q}(x)E(y|z_1=1,x) \\ &\quad - \frac{1-z}{1-\mathcal{Q}(x)}\mathcal{Q}(x)[y - E(y|z=0,x)] - \mathcal{Q}(x)E(y|z=0,x) \\ &\quad - \Big\{z(d - E(d|z_1=1,x)) + \mathcal{Q}(x)E(d|z_1=1,x) \\ &\quad - \frac{1-z}{1-\mathcal{Q}(x)}\mathcal{Q}(x)[d - E(d|z=0,x)] - \mathcal{Q}(x)E(d|z=0,x)\Big\}\gamma. \end{aligned}$$

In addition, when $\mathcal{Q}(x)$ is known up to a finite-dimensional parameter α, 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 $h(\mathcal{Q}_{\hat{\alpha}}(x)) - h(\mathcal{Q}(x))$. It is easy to see, using the parametric Delta method, that this additional term is given by

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

C.3. *Proof of Theorem 6*

Denote $g(w, x\beta) = dg_1(y,x,\beta) - (1-d)g_0(y,x,\beta)$, and introduce the weighting matrix

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

and a moment function $\tilde{g} = Ag(y,d,x,\beta)$. Also define

$$\overline{g}(y,d,x,\beta) = g(y,d,x,\beta) - E[g(y,d,x,\beta)|d,x,D_1 > D_0].$$

Given that the conditional equation (11) is valid, we have

$$E\big\{\mathcal{A}(d,x)E\{g(y,d,x,\beta)|d,x,D_1 > D_0\}\big\} = 0.$$

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

$$J\frac{\partial\beta(\theta)}{\partial\theta} = -\frac{\partial}{\partial\theta}E_\theta\big[\mathcal{A}_\theta(d,x)E_\theta[g(y,d,x,\beta)|d,x,D_1 > D_0]\big].$$

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

$$E\left[\mathcal{A}(d,x)\int \overline{g}(y,d,x,\beta)s_{**}(y|d,x)f_{**}(y|d,x)\,dy\right]$$
$$+E\left[\mathcal{A}(d,x)s_\theta(d,x)\int g(w,x,\beta)f_{**}(y|d,x)\,dy\right]$$
$$+E\left[\frac{\partial \mathcal{A}_\theta(d,x)}{\partial\theta}\int g(w,x,\beta)f_{**}(y|d,x)\,dy\right].$$

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

$$\begin{aligned}
&-J^{-1}\Phi_1(y,d,x,z)\\
&\quad=\frac{(z-\mathcal{Q}(x))(d+\mathcal{Q}(x)-1)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}A(d)g(y,d,x,\beta)-\frac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\\
&\qquad\times A(d)\big[E[dzg|d=1,z=1,x]-E[(1-d)(1-z)g|d=0,z=0,x]\big]\\
&\qquad-\frac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}A(d)\\
&\qquad\times\big[(d-\mathcal{P}_1(x))E[dg|d=1,x,D_1>D_0]\\
&\qquad+(d-\mathcal{P}_0(x))E[(1-d)g|d=0,x,D_1>D_0]\big].
\end{aligned}$$

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

$$\begin{aligned}
E[\Phi_2(y,d,x,z)S_\theta(y,d,x,z)]&=E\left[\mathcal{A}(d,x)s_\theta(d,x)\int g(y,d,x,\beta)f_{**}(y|d,x)\,dy\right]\\
&\quad+E\left[\frac{\partial\mathcal{A}_\theta(d,x)}{\partial\theta}\int g(w,x,\beta)f_{**}(y|d,x)\,dy\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&E[\mathcal{A}(d,x)s_\theta(x,d)E[g|d,x,D_1>D_0]]\\
&\quad=E\big[s_\theta(x)E[\mathcal{A}(d,x)g(y,d,x,\beta)|d,x,D_1>D_0]\big]\\
&\qquad+E\big[\big(E[\mathcal{A}(d,x)g|d=1,x,D_1>D_0]\\
&\qquad-E[\mathcal{A}(d,x)g|d=0,x,D_1>D_0]\big)\dot{\mathbf{P}}_\theta(x)\big]\\
&\quad=E\big[A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\\
&\qquad\times\big(\mathcal{Q}(x)E[g|d=1,x,D_1>D_0]+(1-\mathcal{Q}(x))E[g|d=0,x,D_1>D_0]\big)s_\theta(x)\big]\\
&\qquad+E\left[A(d)(\mathcal{P}_1(x)-\mathcal{P}_0(x))\left\{\frac{\mathcal{Q}(x)E[g|d=1,x,D_1>D_0]}{\mathbf{P}(x)}\right.\right.\\
&\qquad\left.\left.-\frac{(1-\mathcal{Q}(x))E[g|d=0,x,D_1>D_0]}{1-\mathbf{P}(x)}\right\}\dot{\mathbf{P}}_\theta(x)\right].
\end{aligned}$$

Next we can express the directional derivative

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

Consider the expression

$$\begin{aligned}
&E\bigg[\frac{\partial \mathcal{A}_\theta(d,x)}{\partial\theta}\int g(w,x,\beta)f_{**}(y|d,x)\,dy\bigg]\\
&\quad = E\big[A\big(\mathcal{Q}(x)E[g|d=1,x,D_1>D_0]+(1-\mathcal{Q}(x))E[g|d=0,x,D_1>D_0]\big)\\
&\qquad \times(\dot{\mathcal{P}}_{1\theta}(x)-\dot{\mathcal{P}}_{0\theta}(x))\big]\\
&\qquad + E\big[A(\mathcal{P}_1(x)-\mathcal{P}_0(x))(E[g|d=1,x,D_1>D_0]\\
&\qquad - E[g|d=0,x,D_1>D_0])\dot{\mathcal{Q}}(x)\big]\\
&\qquad + E\bigg[A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\bigg\{-\frac{\mathcal{Q}(x)E[g|d=1,x,D_1>D_0]}{\mathbf{P}(x)}\\
&\qquad + \frac{(1-\mathcal{Q}(x))E[g|d=0,x,D_1>D_0]}{1-\mathbf{P}(x)}\bigg\}\dot{\mathbf{P}}_\theta(x)\bigg].
\end{aligned}$$

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

$$\begin{aligned}
&E[\Phi_2(y,d,x,z)S_\theta(y,d,x,z)]\\
&\quad = -J^{-1}E\big[A\big(\mathcal{Q}(x)E[g|d=1,x,D_1>D_0]\\
&\qquad + (1-\mathcal{Q}(x))E[g|d=0,x,D_1>D_0]\big)\\
&\qquad \times\big((\mathcal{P}_1(x)-\mathcal{P}_0(x))s_\theta(x)+\dot{\mathcal{P}}_{1\theta}(x)-\dot{\mathcal{P}}_{0\theta}(x)\big)\big]\\
&\qquad + E\big[A(\mathcal{P}_1(x)-\mathcal{P}_0(x))(E[g|d=1,x,D_1>D_0]\\
&\qquad - E[g|d=0,x,D_1>D_0])\dot{\mathcal{Q}}(x)\big].
\end{aligned}$$

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

$$\begin{aligned}
&-J\Phi_2(y,d,x,z,)\\
&\quad = A(\mathcal{P}_1(x)-\mathcal{P}_0(x))\big(\mathcal{Q}(x)E[g|d=1,x,D_1>D_0]\\
&\qquad + (1-\mathcal{Q}(x))E[g|d=0,x,D_1>D_0]\big)\bigg[1+\frac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\frac{d-\mathcal{F}(z,x)}{\mathcal{P}_1(x)-\mathcal{P}_0(x)}\bigg]\\
&\qquad + A(\mathcal{P}_1(x)-\mathcal{P}_0(x))(E[g|d=1,x,D_1>D_0]-E[g|d=0,x,D_1>D_0])\\
&\qquad \times(z-\mathcal{Q}(x)).
\end{aligned}$$

Combining the two components of the efficiency bound, we obtain, for $\tilde{g} = -J^{-1}Ag(y, d, x, \beta)$,

$$\begin{aligned}
&\Phi(y, d, x, z) \\
&= \left\{dz - \frac{\mathcal{Q}(x)d(1-z)}{1-\mathcal{Q}(x)} + (1-d)(1-z) - \frac{(1-\mathcal{Q}(x))(1-d)z}{\mathcal{Q}(x)}\right\}\tilde{g}(y, d, x, \beta) \\
&\quad + \frac{z-\mathcal{Q}(x)}{\mathcal{Q}(x)(1-\mathcal{Q}(x))}\big[(1-\mathcal{P}_1(x))(1-\mathcal{Q}(x))E[\tilde{g}|d=0, z=1, x] \\
&\quad - \mathcal{P}_0(x)\mathcal{Q}(x)E[\tilde{g}|d=1, z=0, x]\big].
\end{aligned}$$

D. *Technical assumptions*

To state the regularity conditions, we make use of the definitions of the weighted sup norm metric $\|h\|_{\infty,\omega}$ from Chen, Hong, and Torozzi (2008). Let $\mathcal{Q}_0(x)$ denote the true $\mathcal{Q}(x)$, and similarly for other estimated quantities. The first set of assumptions concerns the consistency 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*:

(i) $E[E[\mathcal{A}(x, d)g(y, d, x, \beta)|D_1 > D_0, x, d]] = 0$ *if and only if* $\beta = \beta_0$.

(ii) $\mathcal{Q}_0(\cdot) \in \mathcal{H} = \{\mathcal{Q}(\cdot) : 0 < \underline{q} \le \mathcal{Q}(x) \le \overline{q} < 1\}$ *for some* $\gamma > 0$.

(iii) $\|\hat{\mathcal{Q}}(\cdot) - \mathcal{Q}_0(\cdot)\|_{\infty,\omega} \xrightarrow{p} 0$, $\|\hat{A}(\cdot, k) - A_0(\cdot, k)\|_{\infty,\omega} \xrightarrow{p} 0$ *for* $k = 0, 1$.

(iv) $E[\sup_{\beta\in B} \|g(y_i, d_i, X_i; \beta)\|^2(1 + \|X_i\|^2)^{\omega}] < \infty$.

(v) *There is a nonincreasing function* $b(\cdot)$ *such that* $b(\delta) \to 0$ *as* $\delta \to 0$ *and*

$$E\Big[\sup_{\|\beta-\tilde{\beta}\|<\delta} \|g(y_i, d_j, X_i; \beta) - g(y_i, d_i, X_i, \tilde{\beta})\|^2\Big] \le b(\delta)$$

for all small positive value δ.

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

Assumption 3. *The following conditions hold*:

(i) *There exist a constant* $\varepsilon \in (0, 1]$ *and a small* $\delta_0 > 0$ *such that*

$$E\Big[\sup_{\|\beta-\tilde{\beta}\|<\delta} \|m(Z_i; \beta) - m(Z_i, \tilde{\beta})\|^2\Big] \le const.\ \delta^{\varepsilon}$$

for any small positive value $\delta \le \delta_0$.

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

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

(iv) $E[\sup_{\beta\in B}\|g(y_i, d_i, X_i; \beta)\|(1+\|X_i\|^2)^{2\omega}] < \infty$.

(v) $E[\sup_{|\beta-\beta_0|\leq\delta}\|\partial E[g(y_i, d_i, X_i; \beta)|d_i, x_i]/\partial\beta\|^2(1+\|X_i\|^2)^{\omega}] < \infty$.

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

$$\delta_0(x) = -\frac{1}{1-\mathcal{Q}_0(x)}E(dA(x,d)g(y,d,x,\beta_0)|z=0,x) + \frac{1}{\mathcal{Q}_0(x)}E((1-d)A(x,d)g(w,x,\beta_0)|z=1,x).$$

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,

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

Assumption 4. *The following conditions hold*:

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

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

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

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

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

(i) $0 < \underline{p} \leq \mathcal{P}_0(x) \leq \overline{p} < 1$, $0 < \underline{p} \leq \mathcal{P}_1(x) \leq \overline{p} < 1$.

(ii) $\|\hat{\mathcal{P}}_k(\cdot) - \mathcal{P}_k^0(\cdot)\|_{\infty,\omega} = o_p(n^{-1/4})$ *for* $k=0,1$.

(iii) $\sup_{|\beta-\beta_0|\leq\delta_n}\|\hat{\omega}_{jk}(\cdot,\beta)-\omega_{jk}^0(\cdot,\beta)\|_{\infty,\omega} = o_p(n^{-1/4})$ *for* $j,k=0,1$.

(iv) $\sup_{|\beta-\beta_0|\leq\delta_n}\|\hat{\gamma}_{jk}(\cdot,\beta)-\gamma_{jk}^0(\cdot,\beta)\|_{\infty,\omega} = o_p(n^{-1/4})$ *for* $j,k=0,1$.

(v) $\varphi(\beta, x, d)$ *is twice continuously differentiable in* β *uniformly over* x *and* d.

(vi) $\frac{1}{\sqrt{nh}} + h^2 = o(n^{-1/4})$.

The last condition, in particular, requires that $h \to 0$ at a rate that is slower than $n^{-1/4}$ but faster than $n^{-1/8}$. The notation h^2 characterizes the bias of the two sided numerical derivative under condition (v).

PROPOSITION 1. *Assumption 5 implies condition* (iii) *in Assumption 3.*

TECHNICAL ADDENDUM

E. *Efficiency bound under semiparametric restrictions*

Consider the semiparametric conditional moment equation

$$\varphi(x, d, \mu(x), \beta) = E\big[g(y, d, x, \mu(x), \beta)|d, x, D_1 > D_0\big] = 0, \tag{20}$$

where $g(\cdot)$ 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 an 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 or 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 (20).

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

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

In this theorem, we use notations

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

$$\tilde{\Omega}(x) = E[\Psi(y,d,x,z)\Psi(y,d,x,z)'],$$

and

$$\begin{aligned}\Psi(w,z,x) = &\left(dz - \frac{Q(x)}{1-Q(x)}d(1-z) + (1-d)(1-z) - \frac{1-Q(x)}{Q(x)}(1-d)z\right)\\ &\times \tilde{\zeta}(d,x)g(y,d,x,\beta)\\ &- \frac{(z-Q(x))}{Q(x)(1-Q(x))}\big\{E[dz\tilde{\zeta}(d,x)g|d=1,z=1,x]\\ &- E[(1-d)(1-z)\tilde{\zeta}(d,x)g|w_0=1,z=0,x]\big\}.\end{aligned}$$

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

$$\frac{\partial\varphi}{\partial\beta'}\dot{\beta}+\frac{\partial\varphi}{\partial\mu}\dot{\mu}(x)+\int g(y, d, x, \mu(x), \beta)s_{**}(y|d, x)f_{**}(y|d, x)\,dy=0.$$

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 the case of only one finite-dimensional parameter. In fact, if we impose the restriction that $\mu(x)\in\mathbf{L}^2(\mathbb{R}^k)$, then the function $g(\cdot)$ will still belong to $\mathbf{L}^2$ as a function of x and d for each β. Differentiating the set of unconditional moments along the parametrization path, we evaluate the relevant gradients at $\theta=0$, which leads to

$$\begin{aligned}&J\dot{\beta}+E\biggl[\mathcal{A}(d, x)\frac{\partial\varphi(d, x, \mu(x), \beta)}{\partial\mu}\dot{\mu}(x)\biggr]\\&\quad+E\biggl[\mathcal{A}(d, x)\int g(y, d, x, \mu(x), \beta)s_{**}(y|d, x)f_{**}(y|d, x)\,dy\biggr]=0.\end{aligned}\tag{21}$$

Now denote $\lambda(x)=E[\mathcal{A}(d, x)\frac{\partial\varphi}{\partial\mu}|x]$. Then (21) can be rewritten as

$$\begin{aligned}&J\dot{\beta}+E[\lambda(x)\dot{\mu}(x)]\\&\quad+E\biggl[\mathcal{A}(d, x)\int g(y, d, x, \mu(x), \beta)s_{**}(y|d, x)f_{**}(y|d, x)\,dy\biggr]=0.\end{aligned}$$

Then we can substitute the expression for $\dot{\mu}(x)$ obtained from the directional derivative of the conditional moment and transform the expression (21) into

$$\begin{aligned}&E\biggl[\biggl(\mathcal{A}(d, x)-\biggl\{\frac{\partial\varphi}{\partial\mu}\biggr\}^{-1}\lambda(x)\biggr)\frac{\partial\varphi}{\partial\beta'}\biggr]\dot{\beta}+E\biggl[\biggl(\mathcal{A}(d, x)-\biggl\{\frac{\partial\varphi}{\partial\mu}\biggr\}^{-1}\lambda(x)\biggr)\\&\quad\times\int g(y, d, x, \mu(x), \beta)s_{**}(y|d, x)f_{**}(y|d, x)\,dy\biggr]=0.\end{aligned}$$

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

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

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

$$\tilde{\zeta}(d, x)=\zeta(d, x)-\biggl\{\frac{\partial\varphi}{\partial\mu}\biggr\}^{-1}E\biggl[\zeta(d, x)\frac{\partial\varphi}{\partial\mu}\bigg|x\biggr].$$

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

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

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

$$\tilde{\Omega}(x) = E[\Psi(y, d, x, z)\Psi(y, d, x, z)']$$

with efficient influence function $\Psi(\cdot)$ the structure of which does not change. Note that the semiparametric efficiency bound for the parameter in the conditional moment equation $\varphi(d, x, \mu(x), \beta) = 0$ will be above that for the conditional moment equation $\varphi(d, x, 0, \beta) = 0$. The efficient estimation procedure may be implemented in two stages, where, in the first stage, one estimates β and $\mu(\cdot)$ using an inefficient system of weights (e.g., if one uses a sieve approximation for $\mu(\cdot)$, this will be equivalent to running a nonlinear instrumental variable (IV) regression on the covariates and the sieve terms). Then using these estimates, one can construct the efficient estimator. □

F. *Empirical application*

The analyzed data set contains 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 data for 2815 individuals who applied for welfare in the year 1994 and early 1995.

The FTP program was 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-month 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 data set tracks the individuals for the next 4 years after the program, thus allowing one to compare long-term impacts of welfare programs on individuals.

The main sample contains the data for 2815 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, 1405 individuals were assigned to FTP and 1410 individuals were assigned to AFDC. The data contain 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 is the payments from

AFDC. The third source is Food Stamp payments. In addition, this data set contains information about the background characteristics of the individuals and 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 1730 individuals from the main data sample which were administered in October 1998. This additional set contains information about characteristics of individuals (including education, job experience, family and dependents, housing, food security, and living conditions). The child survey data are based on 1100 additional interviews with adult survey participants who have at least one child between 5 and 12 years old. This survey inquires about school outcomes and kids' interactions with other children. The information contained in the survey includes parenting and fathers' 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 hypotheses 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 compliers 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 to find 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 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 V 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 dummy that indicates the parent is currently employed, child's gender, age dummies for the parent, and the dummy that indicates the parent do not have a high school diploma. In column 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

TABLE V. Regression Outcomes.[a]

Variable	Model number 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
R^2	0.04					

[a] Robust t-statistics are given in parentheses. *significant at 5% level; **significant at 1% level.

the models. In column 4, we report the results of the IV regression, where we use instruments for the employment dummy that include the FTP/AFDC dummy and hourly wage of the last job taken. As one can see, the coefficient indicates now that the parent's employment leads to the decline in the child's achievement by almost 2 points. This suggests that if we do not take into account the endogeneity of the 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

$$g(y,d,x,\beta) = [y - \exp(\beta_0 d + x'\beta_1)]\begin{pmatrix} d \\ x \end{pmatrix}$$

and the moment condition

$$\varphi(d,x,\beta) = E[g(y,d,x,\beta)|x,d,D_1 > D_0] = 0.$$

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

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

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

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

Similarly to the two-stage least squares (2SLS) case, the coefficients in the count data models with endogeneity taken into account are significantly larger in absolute value than in which do not take endogeneity into account. This implies that the endogeneity of job participation causes an 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 on the same 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 model 2 is -0.196 with a standard error of 0.0773, which 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 -0.538 with a standard error of 0.177. This is smaller than the 2SLS estimate, but almost four times larger than the marginal effect in the model that does not take endogeneity of job participation into account.