Given a binary treatment $D$, a response $Y$ and covariates $X$, often $X$ interacts with $D$ in an unknown way to make the treatment effect heterogeneous. If $D$ is exogenous, there are various semi-parametric approaches (matching, inverse probability weighting, etc.) to estimate a weighted average of the heterogeneous effects without specifying the $Y$ model. If $D$ is endogenous, however, then there is hardly any practical semi-parametric approach available.

With an instrument $\delta$ for endogenous $D$ available, this paper proposes a simple instrumental variable estimator (IVE) without specifying the $Y$ model. The IVE is consistent for the ‘$\text{Cov}(\delta,D|X)$-weighted average’ of the heterogeneous effects. Going further, an weighted IVE removing $\text{Cov}(\delta,D|X)$ is consistent for the average of the heterogeneous effects. The IVE is easy to implement with hardly any decision needed by the user. Also it has an asymptotic variance estimator that works well in small samples, and can be extended for multiple treatments. A simulation study and an empirical analysis are provided.

Conflicts of Interest: None.

Running Head: Simple instrumental variable estimator.

MSC 2000 Subject Classifications: 62J02 (primary), 62G05 (secondary).

Key Words: Propensity score, treatment endogeneity, sub-propensity score, instrumental variable estimator.
1 Introduction

Given a binary treatment $D$, a response variable $Y$ and covariates $X$, consider

$$Y = \mu_0(X) + \mu_1(X)D + U, \quad E(U|X) = 0 \quad (1.1)$$

where $\mu_0(X)$ and $\mu_1(X)$ are unknown functions of $X$, and $U$ is a mean-zero error term. Here, $X$ affects $Y$ directly through $\mu_0(X)$, and interacts with $D$ through $\mu_1(X)$. The effect of $D$ on $Y$ is $\mu_1(X)$, which is heterogeneous as long as $\mu_1(X)$ is not a constant. The linearity in $D$ of (1.1) is not a restriction, because an unknown nonlinear function of $(X,D)$, say $f(X,D)$, can be always rewritten as $\mu(X,0) + \{\mu(X,1) - \mu(X,0)\}D$ that is equivalent to (1.1).

In the nonparametric model (1.1) with the heterogeneous effect $\mu_1(X)$, suppose $D$ is endogenous and an instrumental variable (IV) $\delta$ for $D$ is available. Hoping somehow to end up with a nice weighted average of $\mu_1(X)$, we may apply the instrumental variable estimator (IVE) of $Y = E(Y)$ on $D = E(D)$, but this does not work because the omitted $\mu_0(X)$ correlated with $\mu_1(X)$ makes the IVE inconsistent in general.

Differently from this IVE, the IVE of $Y - E(Y)$ on $D - E(D)$ purges $\mu_0(X)$ out of (1.1) to be consistent for $E\{\omega_{sp}(X)\mu_1(X)\}$, where the weighting function $\omega_{sp}(X)$ is proportional to $Cov(\delta,D|X)$ with $Cov$ standing for covariance. Going further, weighted IVE with $\omega_{sp}(X)^{-1}$ as the weight can be employed to estimate $E\{\mu_1(X)\}$, instead of $E\{\omega_{sp}(X)\mu_1(X)\}$; the probability limit $E\{\mu_1(X)\}$ of this weighted IVE is the same regardless of $\delta$, differently from the ‘local average treatment effect (LATE)’ in Imbens and Angrist (1994), Angrist et al. (1996) and Angrist et al. (2000) which varies as $\delta$ does. Essentially, the IVE consistent for $E\{\omega_{sp}(X)\mu_1(X)\}$ and its weighted version consistent for $E\{\mu_1(X)\}$ are our proposals.

What is remarkable in the scenario just outlined is the contrast between the complexity allowed by the model (1.1) and the simplicity of the proposed IVE. The IVE can be easily implemented with $Y - \bar{Y}$, $D - \hat{E}(D|X)$ and $\delta$, where $\bar{Y}$ is the sample average of $Y$ and $\hat{E}(D|X)$ is an estimator for $E(D|X)$ such as the probit/logit of $D$ on $X$; a nonparametric estimator for $E(D|X)$ can be used as well, if desired. Because the IV $\delta$ as well as $X$ affects $D$, the genuine propensity score (PS) would be $E(D|Z)$, not $E(D|X)$, where

$$Z \equiv (\delta, X').$$

In view of this, our specific proposal is the simple IVE with the ‘sub-propensity score (Sub-PS)’ residual $D - \hat{E}(D|X)$. 


In the treatment effect literature (see, e.g., Rosenbaum 2002; Lee 2005, 2016; Pearl 2009; and Imbens and Rubin 2015), the potential responses \( Y^d \) for \( D = d \) appear so that the observed \( Y \) equals \( (1 - D)Y^0 + DY^1 \). If \( (Y^0, Y^1) \perp D | X \) (i.e., if \( D \) is exogenous) where ‘\( A \perp B | C \)’ denotes the independence between \( A \) and \( B \) given \( C \), then

\[
E(Y|D = 1, X) - E(Y|D = 0, X) = E(Y^1|D = 1, X) - E(Y^0|D = 0, X) = E(Y^1 - Y^0|X).
\]

Because nonparametrically controlling \( X \) is difficult, PS \( \pi(X) \equiv E(D|X) \) is widely used for exogenous \( D \), which is based on (Rosenbaum and Rubin 1983)

\[
Y^d \perp D | X \implies Y^d \perp D | \pi(X) \quad \forall d.
\]

Despite the popularity of propensity score matching (PSM), however, PSM is not straightforward to implement, as the user has to make several arbitrary decisions in matching a subject in one group with close subjects in the opposite group. Furthermore, inference with PSM is difficult, despite the advance made by Abadie and Imbens (2016).

To overcome these shortcomings of PSM, Lee (2018) proposed

\[
\text{Ordinary Least Squares estimator (OLS) of } Y - E\{Y|\pi(X)\} \text{ on } D - \pi(X) \quad (1.2)
\]

for (1.1) with exogenous \( D \). Instead of going fully nonparametric, Lee (2018) used probit specifying \( \pi(X) = \Phi(X'\alpha) \) for the \( N(0, 1) \) distribution function \( \Phi \) and a parameter \( \alpha \). \( E\{Y|\pi(X)\} \) in (1.2) can be replaced by the predicted \( Y \) in the OLS of \( Y \) on a polynomial function of \( X'\hat{\alpha} \), say \( \{1, X'\hat{\alpha}, ..., (X'\hat{\alpha})^q\} \), where \( \hat{\alpha} \) is the probit estimator and \( q \) can take any non-negative integer. The simulation study in Lee (2018) showed that the OLS is numerically stable with reliable inference, and outperforms the existing alternatives for exogenous \( D \), such as PSM, regression imputation, inverse probability weighting, etc. by a big margin.

Despite that the OLS in (1.2) has advantages as just mentioned, the OLS as well as PSM and the other exogenous-treatment-based approaches does not allow endogenous treatment. The goal of this paper is to propose an estimator that allows possibly endogenous \( D \), while remaining as simple as the OLS in (1.2). This is done by the IVE with Sub-PS (residual), which includes the OLS in (1.2) as a special case when \( D \) is exogenous and thus \( \delta = D \).

In contrast to the “classical” linear model IVE applied to a linearly specified ‘structural-form’ \( Y \) model such as \( Y = \beta_x'X + \beta_{xd}'XD + U \) for parameters \( \beta_x \) and \( \beta_{xd} \) without specifying the \( D \)-model, IVE with Sub-PS specifies the ‘reduced-form’ model for \( D \) without specifying...
the Y-model; ‘reduced form’ in the sense that only the predicted $E(D|X)$ is used, not the individual coefficient estimates in $E(D|X)$. IVE with Sub-PS differs from the classical linear-model IVE which specifies $\beta'_{sX}X$ as the heterogeneous treatment effect.

There are many sources for IV, and the “canonical” case is randomized trial with non-compliance where the randomization dummy $\delta$ becomes an IV for $D$ that is ridden with non-compliance problems; this is used as a primary example throughout this paper. In general, an IV should meet three conditions: (i) included in the $D$ equation (i.e., $\delta$ affects $D$), (ii) excluded from the $Y$ equation (i.e., $\delta$ can affect $Y$ only indirectly through $D$), and (iii) unrelated to $U$. We also require these conditions for $\delta$.

As for studies allowing endogenous $D$ for (1.1), Das (2005) took $E(D|Z)$ on (1.1):

$$E(Y|Z) = \mu_0(X) + \mu_1(X) \cdot E(D|Z) \quad \text{under } E(U|Z) = 0.$$  

Das (2005) estimated $\mu_0(X)$ and $\mu_1(X)$ in two nonparametric stages using series-approximation: obtain an estimator $\hat{E}(D|Z)$ for $E(D|Z)$ first, and then estimate $\{\mu_0(X), \mu_1(X)\}$ using $Y$ and $\hat{E}(D|Z)$. Although the Das’ (2005) approach is general, it is hardly practical with two bandwidths to choose. Also, Das (2005) focused on estimating $\mu_1(X)$ and its interpretation as ‘conditional LATE’, whereas our focus is on the ‘marginal treatment effects’ $E[\omega_{sp}(X)\mu_1(X)]$ and $E\{\mu_1(X)\}$.

Cai et al. (2006) generalized Das (2005) by allowing $D$ to be continuously distributed. They proposed an estimator with two nonparametric stages, which make Cai et al. (2006) far from practical. Also, as Das (2005) did, the focus was on conditional LATE, not marginal effects. Differently from Das (2005) and (1.1), however, the linearity in $D$ no long holds by construction in Cai et al. (2006), but rather by a local linear approximation.

Okui et al.’s (2012) doubly robust (DR) IVE replaced the part $\mu_1(X)D$ with a parametric function, say $\mu(X, D; \beta)$, with a parameter $\beta$ whereas $\mu_0(X)$ is an unknown function. For estimation, they specify $\mu_0(X)$ as a parametric function, say $\mu_0(X; \alpha)$, with a parameter $\alpha$, and use $\delta - G(X; \gamma)$ as an instrument where $G(X; \gamma)$ is an assumed parametric function with $\gamma$ such that $G(X; \gamma) = E(\delta|X)$. Their IVE is DR because the main moment condition is $E\{[Y - \mu_0(X; \alpha) - \mu(X, D; \beta)]\{\delta - G(X; \gamma)\}\} = 0$, which holds if either $G(X; \gamma)$ or $\mu_0(X; \alpha)$ is a correct specification, not necessarily both. Okui et al. (2012) differs from our proposal because we do not assume a parametric $\mu_1(X)$ to allow for an unknown form of heterogeneity. This difference is not trivial, because allowing an unknown form of heterogeneity is the main
motivation for our proposal; specifying \( \mu_1(X) \) can be rather involved in practice.

Throughout this paper, we maintain the support-overlap condition

\[
0 < \pi(X) \equiv E(D|X) < 1.
\]

As for notation, \( \phi \) denotes the \( N(0, 1) \) density function, and we use ‘\(^\circ\)’ , ‘\(^\ast\)’ or ‘\(^-\)’ to denote an estimator; e.g., \( \hat{\beta} \), \( \tilde{\beta} \) or \( \bar{\beta} \) for \( \beta \). The subscript \( i \) indexing individuals is often omitted, as has been done already.

The rest of this paper is organized as follows. Section 2 briefly reviews the OLS in (1.2). Section 3 presents IVE with Sub-PS, and Section 4 examines the weighted IVE with Sub-PS. Section 5 extends binary endogenous treatment to multiple endogenous treatments. Section 6 conducts a simulation study to verify the claimed properties of IVE and weighted IVE with Sub-PS. Section 7 provides an empirical illustration. Finally, Section 8 concludes.

2 OLS with Propensity Score (PS) Residual

Lee’s (2018) OLS in (1.2) proceeds as follows. First, do the probit estimation of \( D \) on \( X \) under \( \pi(X) = \Phi(X'\alpha) \). Second, obtain the OLS \((\hat{\gamma}_0, ..., \hat{\gamma}_q)\) of \( Y \) on \( \{1, X'\hat{\alpha}, ..., (X'\hat{\alpha})^q\} \).

Third, obtain the OLS \( \hat{\beta}_{\text{psr}}^q \) of \( Y - \sum_{j=0}^q \hat{\gamma}_j (X'\hat{\alpha})^j \) on \( D - \Phi(X'\hat{\alpha}) \), where ‘psr’ stands for PS residual and \( q \) can take any non-negative integer. The estimator \( \hat{\beta}_{\text{psr}}^q \) requires the correctly specified \( \pi(X) \) regardless of the \( q \) value, but \( \hat{\beta}_{\text{psr}}^0 \) with \( q \geq 1 \) tends to be less biased than \( \hat{\beta}_{\text{psr}}^0 \) in case \( \pi(X) \) is misspecified, and \( \hat{\beta}_{\text{psr}}^2 \) or \( \hat{\beta}_{\text{psr}}^3 \) are recommended in practice. The superscript \( q \) in \( \hat{\beta}_{\text{psr}}^q \) will be often omitted.

With \( Y^0 = \mu_0(X) + U \) and \( Y^1 = \mu_0(X) + \mu_1(X) + U \), define

\[
\beta \equiv E\{\mu_1(X)\} = E\{E(Y^1 - Y^0|X)\} = E(Y^1 - Y^0).
\]

Then, \( \hat{\beta}_{\text{psr}} \) is consistent, not for \( \beta \), but for (with ‘Var’ standing for variance)

\[
\hat{\beta}_{\text{psr}} \equiv E\{\omega_{\text{psr}}(X)E(Y^1 - Y^0|X)\}, \quad \omega_{\text{psr}}(X) \equiv \frac{\pi(X)(1 - \pi(X))}{E[\pi(X)|1 - \pi(X)]} = \frac{\text{Var}(D|X)}{E\{\text{Var}(D|X)\}};
\]

\( Y \) does not have to be continuously distributed for this (e.g., \( Y \) can be binary).

There are special cases for \( \hat{\beta}_{\text{psr}} = \beta \). One is \( \mu_1(X) = \beta \) for all \( X \) (‘constant effect’), and another is that the \( D \)-determining covariates are independent of \( \mu_1(X) \)-determining covariates so that \( E\{\omega_{\text{psr}}(X)\mu_1(X)\} = E\{\omega_{\text{psr}}(X)\}E\{\mu_1(X)\} = \beta \). The weighting by \( \omega_{\text{psr}}(X) \) may look
uncalled for, but it is in fact desirable, because those with \( \pi(X) \simeq 0.5 \) are given higher weights due to \( \omega_{\text{psr}}(X) \) reaching its maximum at \( \pi(X) = 0.5 \)—they are close to being randomized—and those with \( \pi(X) \simeq 0, 1 \) are given almost zero weight—they are outliers.

In PSM, observations with \( \pi(X) \simeq 0, 1 \) have to be downweighted because they have hard time finding matched subjects, and in inverse probability weighting, observations with \( \pi(X) \simeq 0, 1 \) are also downweighted as they result in extreme values. Differently from these approaches with arbitrary features to downweight observations with \( \pi(X) \simeq 0, 1 \), \( \beta_{\text{psr}} \) has a built-in feature to do so with \( \omega_{\text{psr}}(X) \). That is, no user intervention is needed for \( \beta_{\text{psr}} \) to deal with extreme observations. In this sense, \( \beta_{\text{psr}} \) might be an answer to the call in Athey et al. (2017) for an appropriate weighted average of \( E(Y^1 - Y^0 | X) \) as a parameter of interest when PS is extreme.

As for the asymptotic distribution of \( \beta_{\text{psr}}^2 \), it holds that

\[
\sqrt{N} (\hat{\beta}^2_{\text{psr}} - \beta_{\text{psr}}) \rightarrow^d N(0, \Omega_{\text{psr}}),
\]

\[
\hat{\Omega}_{\text{psr}} \equiv \left( \frac{1}{N} \sum_i \hat{\epsilon}_i^2 \right) - \frac{1}{N} \sum_i (\hat{V}_i^{\text{psr}} \hat{\epsilon}_i + \hat{L}_d \hat{\eta}_{ai})^2 \rightarrow^d \Omega_{\text{psr}}, \quad \hat{\epsilon}_i \equiv D_i - \Phi(X_i' \hat{\alpha}),
\]

\[
\hat{V}_i^{\text{psr}} \equiv Y_i - \{ \hat{\gamma}_0 + \hat{\gamma}_1 X_i' \hat{\alpha} + \hat{\gamma}_2 (X_i' \hat{\alpha})^2 \} - \beta_{\text{psr}}^2 \hat{\epsilon}_i, \quad \hat{L}_d \equiv -\frac{1}{N} \sum_i \hat{V}_i^{\text{psr}} \phi(X_i' \hat{\alpha}) X_i,
\]

\[
\hat{\eta}_{ai} \equiv \left( \frac{1}{N} \sum_i \hat{s}_i \hat{\epsilon}_i' \right)^{-1} \hat{s}_i, \quad \hat{s}_i \equiv \frac{\hat{\epsilon}_i \phi(X_i' \hat{\alpha})}{\Phi(X_i' \hat{\alpha}) \{ 1 - \Phi(X_i' \hat{\alpha}) \}} X_i;
\]

\( \hat{V}_{\text{psr}} \) is the OLS residual, and \( \hat{L}_d \) is the “link” through which the error \( \hat{\alpha} - \alpha \) gets channeled to the asymptotic distribution of \( \beta_{\text{psr}}^2 \). If we allow more polynomial terms of \( X' \hat{\alpha} \) than \( (X' \hat{\alpha})^2 \), the only modification needed is adding those into \{ \} in \( \hat{V}_{\text{psr}} \).

### 3  IVE with Sub-Propensity Score Residual

#### 3.1  Identification

For an IV \( \delta \) that may not be binary, consider (1.1) with \( E(U | \delta, X) = 0 \):

\[
Y = \mu_0(X) + \mu_1(X) D + U, \quad E(U | \delta, X) = 0.
\]

Noting \( \pi(X) = E\{ D | \pi(X) \} \) which holds by taking \( E\{ \cdot | \pi(X) \} \) on \( \pi(X) \equiv E(D | X) \), take \( E\{ \cdot | \pi(X) \} \) on the \( Y \) equation:

\[
E\{ Y | \pi(X) \} = E\{ \mu_0(X) | \pi(X) \} + E\{ \mu_1(X) D | \pi(X) \}.
\]
Subtract this from the $Y$ equation to get, for an error term $V$,

$$Y - E\{Y|\pi(X)\} = V$$  \hspace{1cm} (3.3)

where

$$V \equiv \mu_0(X) - E\{\mu_0(X)|\pi(X)\} + \mu_1(X)D - E\{\mu_1(X)D|\pi(X)\} + U$$

which is reminiscent of Robinson (1988).

Under the weak assumption

$$E\{Cov(\delta, D | X)\} \neq 0,$$  \hspace{1cm} (3.4)

a sufficient condition of which is $Cov(\delta, D | X) \geq 0$ (‘monotonicity’) and $Cov(\delta, D | X) \neq 0$ for some $X$, the population version of our proposed IVE is

$$\beta_{sp} = \frac{E[\{\delta - E(\delta|X)\}\{Y - E\{Y|\pi(X)\}\}]}{E[\{\delta - E(\delta|X)\}\{D - \pi(X)\}]} = \frac{E[[\delta - E(\delta|X)\} Y - E(Y)]]}{E[[\delta - E(\delta|X)\} D - \pi(X)]]}$$  \hspace{1cm} (3.5)

because $E[[\delta - E(\delta|X)\} E\{Y|\pi(X)\} = E[[\delta - E(\delta|X)\} E(Y)] = 0$. The denominator of $\beta_{sp}$ is $E\{Cov(\delta, D | X)\}$ because

$$E[\{\delta - E(\delta|X)\} \{D - \pi(X)\}] = E( E[[\delta - E(\delta|X)\} \{D - \pi(X)\}] | X \).$$

In the first ratio of (3.5), we may drop $E(\delta|X)$—more on this below.

Substituting $Y - E\{Y|\pi(X)\}$ in (3.3) into the first ratio in (3.5) and using $\pi(X) = \Phi(X'\alpha)$ that is one-to-one to $X'\alpha$, the numerator of the first ratio in (3.5) becomes

$$E(\{\delta - E(\delta|X)\} [\mu_0(X) - E\{\mu_0(X)|X'\alpha\} + \mu_1(X)D - E\{\mu_1(X)D|X'\alpha\} + U]).$$

All terms except $\mu_1(X)D$ drop out as they are uncorrelated with $\delta - E(\delta|X)$ to leave

$$E[[\delta - E(\delta|X)\} \mu_1(X)D] = E[[\delta - E(\delta|X)\} \{D - \pi(X)\} \mu_1(X)] = E\{Cov(\delta, D | X)\mu_1(X).$$

Hence, we have

$$\beta_{sp} = E\{\omega_{sp}(X) \mu_1(X)\}, \hspace{1cm} \omega_{sp}(X) \equiv \frac{Cov(\delta, D | X)}{E\{Cov(\delta, D | X)|};$$

$\beta_{sp}$ includes $\beta_{psr}$ as a special case when $\delta = D$.

Remark 1. It is important to see that $D - E(D|Z)$ cannot be the regressor because $D - E(D|Z)$ is uncorrelated with the IV $\delta$ due to $Z$ including $\delta$, which makes $\delta$ useless as an IV. This is why the Sub-PS $\pi(X)$ is used, not the true PS $\pi(Z)$.
Remark 2. The weight \( \omega_{sp}(X) \) may look strange, but IVE tends to weight this way. In the “classical” simple IVE of \( Y \) on \((1,D)\) with \( \delta \) as an IV, the probability limit of the slope estimator can be written as a weighted average of the individual slopes where the changes in \( D \) and \( Y \) are gauged relative to their means:

\[
\frac{E[\delta (Y-E(Y))]}{E[\delta (D-E(D))]} = E\{\frac{\delta - E(\delta)}{\delta - E(\delta)} \{D - E(D)\}} \frac{Y-E(Y)}{D-E(D)}
\]

\[
= E\{\omega(\delta, D) \frac{Y-E(Y)}{D-E(D)}\}, \quad \omega(\delta, D) \equiv \frac{\{\delta - E(\delta)\} \{D - E(D)\}}{E[\{\delta - E(\delta)\} \{D - E(D)\}]}. \tag{3.6}
\]

Remark 3. To appreciate \( \omega_{sp}(X) \) better, consider the canonical case of randomized binary \( \delta \) with \( E(\delta) = 0.5, \pi(X) = \Phi(X'\alpha) \) and \( \pi(Z) = \pi(\delta, X) = \Phi(\alpha_0 + X'\alpha) \) for a parameter \( \alpha_0 > 0 \), which makes \( \text{Cov}(\delta, D|X) = E(\delta D|X) - E(\delta|X)E(D|X) \) equal to

\[
E(D|\delta = 1, X) P(\delta = 1|X) - 0.5 \Phi(X'\alpha) = 0.5 \{ \Phi(\alpha_0 + X'\alpha) - \Phi(X'\alpha) \} \approx 0.5 \Phi(X'\alpha) \alpha_0.
\]

As in the OLS with PS residual, \( \omega_{sp}(X) \) is the largest when \( X'\alpha \approx 0 \) so that \( \pi(X) \approx 0.5 \) with \( X \) playing no role for \( D \). Also as in the OLS with PS residual, \( \omega_{sp}(X) \approx 0 \) when \( X'\alpha \approx \pm \infty \) (outliers) so that \( \pi(X) \approx 0, 1 \). This shows that weighting with \( \omega_{sp}(X) \) can be desirable.

Remark 4. We may drop \( E(\delta|X) \) from the first ratio of (3.5) to simplify IVE with Sub-PS. Nevertheless, we keep \( E(\delta|X) \) for three reasons. First, the IVE with \( \delta - E(\delta|X) \) as the IV includes the OLS in (1.2) as a special case when \( \delta = D \). Second, the asymptotic distribution of the IVE with \( \delta - E(\delta|X) \) is simpler than that of the IVE with \( \delta \), because the IV residual \( \delta - E(\delta|X) \) cancels several terms as can be seen in the appendix. Third, in the canonical case of randomized \( \delta \), \( E(\delta|X) \) can be simply replaced with the sample average \( \bar{\delta} \) of \( \delta \).

Remark 5. Denoting the sample version of (3.5) as \( \hat{\beta}_{sp} \), if \( \delta \perp X \) (e.g., \( \delta \) is randomized), we may then use the classical simple IVE (i.e., the sample version of (3.6)), which differs from \( \hat{\beta}_{sp} \), because the omitted \( X \)-part is \( \mu_0(X) \) for the classical simple IVE whereas it is \( \mu_0(X) - E\{\mu_0(X)|X'\alpha\} \) for \( \hat{\beta}_{sp} \). The error term variance may be smaller in \( \hat{\beta}_{sp} \) at the expense of specifying \( \pi(X) \) and \( E(\delta|X) \), compared with the classical simple IVE. If \( X \) is of low dimension, the variance reduction in \( \hat{\beta}_{sp} \) can be substantial.

### 3.2 Estimation

The second ratio in (3.5) suggests to obtain \( \pi(X) = \Phi(X'\alpha) \) and

\[
\text{IVE of } Y - E(Y) \text{ on } D - \Phi(X'\alpha) \text{ using } \delta - E(\delta|X) \text{ as IV}.
\]
Let $E(\delta|X, \hat{\psi})$ be an estimator for $E(\delta|X)$ parametrized with $\psi$: if $\delta$ is binary and $E(\delta|X) = \Phi(X'\hat{\psi})$ is adopted, then $E(\delta|X, \hat{\psi}) = \Phi(X'\hat{\psi})$ with $\hat{\psi}$ being the probit estimator of $\delta$ on $X$; if $\delta$ is continuously distributed with $E(\delta|X) = X'\psi$, then $E(\delta|X, \hat{\psi}) = X'\hat{\psi}$ with $\hat{\psi}$ being the OLS of $\delta$ on $X$. Obtain then

$$\text{IVE of } Y - \bar{Y} \text{ on } D - \Phi(X'\hat{\alpha}) \text{ with } \delta - E(\delta|X, \hat{\psi}) \text{ as IV.} \quad (3.7)$$

If $\delta$ is randomized, then there is no need to estimate $E(\delta|X)$ because of $E(\delta|X) = E(\delta)$. Instead of (3.7), we actually advocate what the first ratio in (3.5) suggests:

$$\text{IVE of } Y - E\{Y|\pi(X)\} \text{ on } D - \Phi(X'\alpha) \text{ using } \delta - E(\delta|X) \text{ as IV.}$$

That is, obtain the OLS $\hat{\gamma} \equiv (\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_q)'$ of $Y$ on $\{1, X'\hat{\alpha}, \ldots, (X'\hat{\alpha})^q\}$ and

$$\text{IVE } \hat{\beta}_sp^q \text{ of } Y - \sum_{j=0}^q \hat{\gamma}_j(X'\hat{\alpha})^j \text{ on } D - \Phi(X'\hat{\alpha}) \text{ with } \delta - E(\delta|X, \hat{\psi}) \text{ as IV.} \quad (3.8)$$

This is consistent for any of $q = 0, 1, \ldots$, but set $q = 2 \sim 3$ in practice as in Lee (2018).

The difference between (3.7) and ‘(3.8) with $q \geq 1$’ is that whereas (3.7) leaves $E\{Y|\pi(X)\}$ in the error term, (3.8) pulls $E\{Y|\pi(X)\}$ out of the error and account for it. There are reasons why (3.8) is preferred. First, (3.8) becomes $\hat{\beta}_{psr}^q$ when $\delta - E(\delta|X, \hat{\psi})$ is replaced by $D - \Phi(X'\hat{\alpha})$. Second, in case of misspecified $E(\delta|X)$, (3.8) tends to perform better than (3.7), analogously to $\hat{\beta}_{psr}^q$ with $q \geq 1$ doing better than $\hat{\beta}_{psr}^0$.

As for the asymptotic distribution, under $\pi(X) = \Phi(X'\alpha)$ and $E(\delta|X) = X'\psi$, the appendix proves that, with $\hat{\varepsilon} \equiv D_i - \Phi(X'_i\hat{\alpha})$,

$$\sqrt{N}(\hat{\beta}_{sp}^q - \beta_{sp}) \rightarrow d N(0, \Omega_{sp}^0), \quad \hat{V}_i \equiv Y_i - \{\hat{\gamma}_0 + \hat{\gamma}_1X'_i\hat{\alpha} + \hat{\gamma}_2(X'_i\hat{\alpha})^2\} - \hat{\beta}_{sp}^2\hat{\varepsilon}_i,$$

$$\hat{\Omega}_{sp}^0 \equiv \left\{\frac{1}{N} \sum_i (\delta_i - X'_i\hat{\psi})\hat{\varepsilon}_i\right\}^{-2} \frac{1}{N} \sum_i \{(\delta_i - X'_i\hat{\psi})\hat{V}_i + \hat{L}_\psi\hat{\eta}_\psi\}^2 \rightarrow \Omega_{sp}^0,$$  

$$\hat{L}_\psi \equiv -\frac{1}{N} \sum_i \hat{V}_iX'_i, \quad \hat{\eta}_\psi = \left(\frac{1}{N} \sum_i X_iX'_i\right)^{-1}X_i(\delta_i - X'_i\hat{\psi}).$$

If we want to allow $(X'\hat{\alpha})^3$ additionally for $E(Y|X'\alpha)$, the requisite modification is adding $\hat{\gamma}_3(X'_i\hat{\alpha})^3$ into $\{\}$ in $\hat{V}_i$.

The specification $E(\delta|X) = X'\psi$ is likely to work fine even for binary $\delta$, but if one wants to use $E(\delta|X) = \Phi(X'\psi)$ for binary $\delta$, then denoting the estimator with $E(\delta|X) = \Phi(X'\psi)$
as \( \beta^\delta \), the appendix proves that
\[
\sqrt{N}(\beta_{sp}^2 - \beta_{sp}) \xrightarrow{d} N(0, \Omega_{sp}^1), \quad \bar{V}_i \equiv Y_i - \{\hat{\gamma}_0 + \hat{\gamma}_1 X_i' \hat{\alpha} + \hat{\gamma}_2 (X_i' \hat{\alpha})^2\} - \tilde{\beta}_{sp}^2 \tilde{\varepsilon}_i,
\]
\[
\bar{\Omega}_{sp}^1 \equiv \left[ \frac{1}{N} \sum_i \{\delta_i - \Phi(X_i' \hat{\psi})\} \tilde{\varepsilon}_i \right]^{-2} \frac{1}{N} \sum_i \{\delta_i - \Phi(X_i' \hat{\psi})\} \bar{V}_i + \bar{L}_\psi \tilde{\eta}_\psi ,
\]
\[
\bar{L}_\psi = - \frac{1}{N} \sum_i \phi(X_i' \hat{\psi}) \bar{V}_i X_i' \tilde{\eta}_\psi = (\frac{1}{N} \sum_i \tilde{s}_i \tilde{s}_i')^{-1} \tilde{s}_i, \quad \tilde{s}_i = \frac{\{\delta_i - \Phi(X_i' \hat{\psi})\} \phi(X_i' \hat{\psi}) X_i}{\Phi(X_i' \hat{\psi}) (1 - \Phi(X_i' \hat{\psi}))}.
\]

## 4 Weighted IVE (WIV)

With weighted IVE (WIV), we can estimate \( \beta = E\{\mu_1(X)\} \) as was already noted. The idea is dividing the moments defining \( \beta_{sp} \) by \( Cov(\delta, D|X)^{-1} \) so that \( Cov(\delta, D|X) \) in \( \beta_{sp} \) is removed. Implementing WIV in practice requires selecting the observations with \( |Cov(\delta, D|X)| > \tau \) for a chosen small positive constant \( \tau > 0 \) (‘\( \tau \)’ for trimming). Hence, strictly speaking, the identified parameter is not \( \beta \), but
\[
E\{\omega_\tau(X) \mu_1(X)\} \quad \text{where} \quad \omega_\tau(X) = \frac{1[|Cov(\delta, D|X)| > \tau]}{E[1[|Cov(\delta, D|X)| > \tau]}.
\]

and \( 1[A] = 1 \) if \( A \) holds and 0 otherwise. For the sake of simplicity, however, we assume \( Cov(\delta, D|X) \neq 0 \) for all \( X \) so that trimming can be ignored for identification.

As an example for \( Cov(\delta, D|X) \neq 0 \) for all \( X \), suppose that \( \delta \) is a randomization dummy with \( E(\delta) = 0.5 \), and \( D = 1[\alpha_\delta \delta + X' \alpha_x + \varepsilon > 0] \) for parameters \( \alpha_\delta \neq 0, \alpha_x \) with \( \varepsilon \sim N(0, 1) \) \( \Pi(\delta, X) \). Then
\[
Cov(\delta, D|X) = E(\delta D|X) - 0.5E(D|X)
\]
\[
= E(D|X, \delta = 1)P(\delta = 1|X) - 0.5E\{\Phi(\alpha_\delta \delta + X' \alpha_x)|X\}
\]
\[
= \Phi(\alpha_\delta + X' \alpha_x)0.5 - 0.5\{0.5\Phi(X' \alpha_x) + 0.5\Phi(\alpha_\delta + X' \alpha_x)\}
\]
\[
= 0.25\{\Phi(\alpha_\delta + X' \alpha_x) - \Phi(X' \alpha_x)\} \neq 0 \quad \text{for all} \ X.
\]

Define
\[
\omega(X) \equiv Cov(\delta, D|X) = E(\delta D|X) - E(\delta|X)\pi(X).
\]

Divide the moments in the last ratio in (3.5) by \( \omega(X) \):
\[
\frac{E[\{\delta - E(\delta|X)\} \{Y - E(Y)\}]/\omega(X)]}{E[\{\delta - E(\delta|X)\} \{D - \pi(X)\}]/\omega(X)]} = E[\{\delta - E(\delta|X)\} \{Y - E(Y)\}]/\omega(X)] \quad (4.1)
\]
because the denominator of the left side of (4.1) is one. Doing analogously to the steps from (3.5) to $E\{\omega_{sp}(X)\mu_1(X)\}$, we have

$$
E[\{\delta - E(\delta|X)\} \{Y - E(Y)\}/\omega(X)] = E[\{\delta - E(\delta|X)\} \mu_1(X)D/\omega(X)]
$$

$$
= E[\{\delta - E(\delta|X)\} \mu_1(X) \{D - \pi(X)\}/\omega(X)] = E\{\mu_1(X)\}.
$$

Consider WIV for (3.8): with an estimator $\hat{\omega}(X)$ for $\omega(X)$,

$$
\hat{\beta}_{\omega_{sp}}^q \equiv \frac{\sum_i (\delta_i - E(\delta|X_i, \hat{\psi})) \{Y_i - \sum_j^q \hat{\gamma}_j X'_j(\hat{\alpha})\} / \hat{\omega}(X_i)}{\sum_i (\delta_i - E(\delta|X_i, \hat{\psi})) \{D_i - \Phi(X'_i\hat{\alpha})\} / \hat{\omega}(X_i)}.
$$

In practice, as was mentioned above, $\hat{\beta}_{\omega_{sp}}^q$ without trimming can take extreme values. Hence, although not explicit, for this estimator and the discussion below, only the observations with $|\hat{\omega}(X_i)| > \tau$ should be used, e.g., with $|\hat{\omega}(X_i)| > 0.01$.

To implement WIV, we need to find $E(\delta D|X)$ extra. When $\delta$ is a randomization dummy and $D = 1[\alpha_\delta \delta + X'\alpha_x + \varepsilon > 0]$ with $\varepsilon \sim N(0, \sigma^2)$ II $\delta, X$ for a constant $\sigma$, we have $E(\delta|X) \approx \bar{\delta}$, and

$$
E(\delta D|X) = E(\delta 1[\alpha_\delta \delta + X'\alpha_x + \varepsilon > 0]|X)
$$

$$
= E(1[\alpha_\delta + X'\alpha_x + \varepsilon > 0]|\delta = 1, X)P(\delta = 1|X) = \Phi(\frac{\alpha_\delta + X'\alpha_x}{\sigma})E(\delta).
$$

Denoting the probit estimator of $D$ on $(\delta, X)$ as $(\hat{\alpha}_\delta, \hat{\alpha}'_x)'$, we have $\hat{\omega}(X) = \Phi(\hat{\alpha}_\delta + X'\hat{\alpha}_x)\bar{\delta} - \Phi(X'\hat{\alpha})\bar{\delta}$. Since multiplying $\hat{\omega}(X)$ by $\bar{\delta}$ does not matter, we can use instead

$$
\hat{\omega}^0(X) \equiv \Phi(\hat{\alpha}_\delta + X'\hat{\alpha}_x) - \Phi(X'\hat{\alpha}).
$$

When $\delta$ is not necessarily binary, still under $\varepsilon \sim N(0, \sigma^2)$ II $\delta, X$, we may use another weighting function $\hat{\omega}^1(X)$ below because

$$
E(\delta D|X) = E[\delta E\{1[\alpha_\delta \delta + X'_i\alpha_x + \varepsilon > 0]|\delta, X\} |X]
$$

$$
= E\{\delta \Phi(\frac{\alpha_\delta + X'_x\alpha_x}{\sigma}) |X\} \approx \frac{1}{N} \sum_{j=1}^N \delta_j \Phi(\hat{\alpha}_\delta \delta_j + X'\hat{\alpha}_x)
$$

$$
\Rightarrow \hat{\omega}^1(X) \equiv \frac{1}{N} \sum_{j=1}^N \delta_j \Phi(\hat{\alpha}_\delta \delta_j + X'\hat{\alpha}_x) - E(\delta|X, \hat{\psi})\Phi(X'\hat{\alpha}).
$$

In practice, the simplest approach might be specifying a reduced-form linear model $E(\delta D|X) = X'\lambda$ for a parameter $\lambda$, as we need only the predicted value of $\delta D$, not the parameters in $E(\delta D|X)$. This leads to, with the OLS $\hat{\lambda}$ of $\delta D$ on $X$,

$$
\hat{\omega}^2(X) \equiv X'\hat{\lambda} - E(\delta|X, \hat{\psi})\Phi(X'\hat{\alpha}).
$$
Since estimating weights does not affect the asymptotic distribution (i.e., estimated weights are as good as the true weights), denoting the weighed IVE with $q$ as $\tilde{\beta}_{\omega \text{sp}}^q$,

$$
\sqrt{N}(\beta_{\omega \text{sp}}^2 - \beta) \rightarrow^d N(0, \Omega), \quad \hat{\delta}_i = \delta - E(\delta|X_i, \hat{\psi}),
$$

$$
\hat{V}_i^\omega \equiv Y_i - \{\gamma_0 + \gamma_1 X_i^\prime \hat{\alpha} + \gamma_2 (X_i^\prime \hat{\alpha})^2\} - \tilde{\beta}_{\omega \text{sp}}^2 \hat{\varepsilon}_i
$$

$$
\hat{\Omega}_\omega \equiv \left\{ \frac{1}{N} \sum_i \hat{\delta}_i \hat{\varepsilon}_i \right\} - \frac{1}{N} \sum_i \left( \hat{\delta}_i \hat{V}_i^\omega \right)^2 + \hat{\varepsilon}_i \hat{\varepsilon}_i^\prime 
\rightarrow^p \Omega, 
$$

$$
\hat{L}_\psi^q \equiv \frac{-1}{N} \sum_i \hat{V}_i^\omega X_i^\prime \hat{\alpha} \quad \text{or} \quad \frac{-1}{N} \sum_i \phi(X_i^\prime \hat{\psi}) \hat{V}_i^\omega X_i^\prime \hat{\alpha} \quad \text{if } E(\delta|X_i, \hat{\psi}) = X_i^\prime \hat{\psi} \text{ or } \Phi(X_i^\prime \hat{\psi}); 
$$

$\hat{\omega}(X)$ is $\hat{\omega}^0(X), \hat{\omega}^1(X)$ or $\hat{\omega}^2(X)$, and $\hat{\eta}_{\psi i}$ is the $\hat{\eta}_{\psi i}$ in (3.9) if $E(\delta|X_i, \hat{\psi}) = X_i^\prime \hat{\psi}$ or $\hat{\eta}_{\psi i}$ in (3.10) if $E(\delta|X_i, \hat{\psi}) = \Phi(X_i^\prime \hat{\psi})$. The notation $\tilde{\beta}_{\omega \text{sp}}^q$ is used here to encompass both $\tilde{\beta}_{\omega \text{sp}}^q$ corresponding to $E(\delta|X, \hat{\psi}) = X^\prime \hat{\psi}$ and $\tilde{\beta}_{\omega \text{sp}}^q$ corresponding to $E(\delta|X, \hat{\psi}) = \Phi(X^\prime \hat{\psi})$. In our simulation section below, we use both $\tilde{\beta}_{\omega \text{sp}}^q$ and $\tilde{\beta}_{\omega \text{sp}}^q$.

## 5 IVE for Multiple Treatments

Suppose $D$ takes on $0, 1, \ldots, J$, which represent ordered, partly ordered, or non-ordered treatments. Define ‘generalized propensity score’ (Imbens 2000, and Imai and van Dyk 2004):

$$
\pi(X) \equiv \{\pi_1(X), \ldots, \pi_J(X)\}' \quad \text{where } \pi_d(X) \equiv E(D_d|X), \ D_d \equiv 1[D = d].
$$

Consider a constant effect model with an IV vector $\delta \equiv (\delta_1, \ldots, \delta_J)'$:

$$
Y = \mu_0(X) + \sum_{d=1}^J \beta_d D_d + U \quad (\text{with } E(U|\delta, X) = 0)
$$

$$
= \mu_0(X) + (D^\prime \beta + U \quad \text{where } D^\prime \equiv (D_1, \ldots, D_J)', \ \beta \equiv (\beta_1, \ldots, \beta_J)'.
$$

As in Lee (2018), it is difficult to find a formula analogous to $\beta_{\text{sp}}$ for multiple treatments, which is why we stick to the constant-effect or ‘parallel shift’ model.

Taking $E\{\cdot | \pi_d(X)\}$ on $\pi_d(X) \equiv E(D_d|X)$ gives $\pi_d(X) = E\{D_d|\pi_d(X)\}$; also taking $E\{\cdot | \pi(X)\}$ on $\pi_d(X) \equiv E(D_d|X)$ gives $\pi_d(X) = E\{D_d|\pi(X)\}$. Using these, take $E\{\cdot | \pi(X)\}$ on the $Y$ equation to obtain

$$
E\{Y|\pi(X)\} = E\{\mu_0(X)|\pi(X)\} + \pi(X)'\beta.
$$

Subtract this from the $Y$ equation to have

$$
Y - E\{Y|\pi(X)\} = \{D^\prime - \pi(X)'\beta + V, \ V \equiv \mu_0(X) - E\{\mu_0(X)|\pi(X)\} + U. \quad (5.2)
$$
Assuming the $J \times J$ matrix $E[\{\delta - E(\delta|X)\}{D^u - \pi(X)}'] = E[Cov(\delta, D^u|X)]$ to be invertible, the population version of IVE is, with $E^{-1}(\cdot)$ denoting $\{E(\cdot)\}^{-1}$,

$$E^{-1}[\{\delta - E(\delta|X)\}{D^u - \pi(X)}'] \times E(\{\delta - E(\delta|X)\}[Y - E\{Y|\pi(X)\}]).$$

Substituting (5.2) into (5.3), the second matrix in (5.3) becomes $E[\{\delta - E(\delta|X)\}{D^u - \pi(X)}']\beta$ to make the population version of the IVE equal to $\beta$.

To implement (5.3), we need to estimate $\pi(X)$, which depends on the nature of $D$. For instance, if $D$ is ordered, we can use ordered probit to estimate $\pi(X)$, all components of which depend on the single underlying regression function, say $X'\alpha$, for the ordered probit. Then the IVE is a sample version of

$$E^{-1}[\{\delta - E(\delta|X)\}{D^u - \pi(X)}'] \times E[\{\delta - E(\delta|X)\}{Y - E(Y|X'|\alpha)}].$$

The $d$th component of the vector $E(\delta|X)$ can be found with the OLS of $\delta_d$ on $X$, $\pi(X)$ with the predicted ordered probit probabilities, and $E(Y|X'|\alpha)$ with the OLS predicted value of $Y$ on $\{1, X'\hat{\alpha}, \ldots, (X'\hat{\alpha})^q\}$ where $\hat{\alpha}$ is the estimator for $\alpha$ in the ordered probit. The details of this procedure with exogenous $D$ can be seen in Lee (2018).

The preceding case is the simplest, and the opposite is unordered (i.e., multinomial) $D$, where $\pi(X)$ may be estimated by multinomial (nested) logit, which entails $J$ linear index functions, say $X'\alpha_1, \ldots, X'\alpha_J$. Then $E(Y|X'|\alpha_1, \ldots, X'\alpha_J)$ can be found with the OLS of $Y$ on a polynomial function of $(X'\hat{\alpha}_1, \ldots, X'\hat{\alpha}_J)$, where $\hat{\alpha}_d$ is the estimator for $\alpha_d$ in the multinomial (nested) logit. An intermediate case with two linear index functions appeared in Ju and Lee (2017) who used a nearly parametric approach. We omit details on these cases.

As for inference, we can derive the asymptotic variance of the aforementioned IVE’s, which is, however, cumbersome. Instead, we propose to use nonparametric bootstrap. Because all procedures involved are smooth and because well-converging estimators such as ordered probit and multinomial (nested) logit are employed, nonparametric bootstrap sampling from the original sample with replacement should work fine in practice.
6 Simulation Study

Our base simulation design is as follows:

\[ D = 1[0 < \alpha_s \delta + \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon], \quad \delta \text{ is binary with } P(\delta = 1) = 0.5, \]

\[ X_2, X_3 \sim N(0,1) \text{ with } Cor(X_2, X_3) = 0.5^{1/2} \approx 0.7, \quad \varepsilon \sim N(0,1) \Pi (\delta, X_2, X_3), \]

\[ Y = \beta_d D + \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U, \quad U \sim N(0,1) \Pi (X_2, X_3), \quad Cor(U, \varepsilon) \approx 0.72 \]

\[ \alpha_3 = 3, \alpha_1 = -2, \alpha_2 = 1, \alpha_3 = -1, \beta_3 = 1, \beta_1 = 0, \beta_2 = \beta_3 = 1, \]

where \( E(D) \approx 0.43 \) and \( Cor(\delta, D) \approx 0.7; \) we set \( N = 500,1500 \) to repeat the simulation 5000 times. When \( \alpha_3 = -1, (X_2, X_3) \) overlap well across the two groups with the averages \((-0.1,0.1)\) for the control group and \((0.1,-0.1)\) for the treatment group. When \( \alpha_3 = 1, \) however, \((X_2, X_3)\) overlap poorly with the averages \((-0.4,-0.4)\) and \((0.6,0.6)\). The degree of covariate overlap matters much in our simulation results.

We try variations of the base design, and in all designs, regardless of how \( D \) and \( Y \) are generated, \((1, \delta, X_2, X_3)\) is the regressor in the probit for \( D \) in the PS-based OLS, and \((1, X_2, X_3)\) is the regressor in the probit for \( D \) in the Sub-PS-based IVE. For each estimator, we present \(|bias|, \text{Sd}, \) and the average of the asymptotic Sd estimates across 5000 repetitions using the asymptotic variance estimator, which is to see how good the asymptotic variance estimator is relative to the actual simulation Sd.

In each table, we compare 11 estimators: the linear-model OLS correctly specifying the \( Y \) model as above, \( \hat{\beta}_{par}^2 \) for the PS-based OLS, the linear-model IVE correctly specifying the \( Y \) model as above using \( \delta \) as an IV for \( D, (\hat{\beta}_{sp}, \hat{\beta}_{sp}) \) for the Sub-PS-based IVE with IV \( \delta - X'\hat{\psi}, (\hat{\beta}_{sp}, \hat{\beta}_{sp}) \) for the Sub-PS-based IVE with IV \( \delta - \Phi(X'\hat{\psi}). \) Although we try \( q = 2,3 \) for the Sub-PS-based IVE, the difference turns out to be negligible. Hence for WIV, we set \( q = 2 \) to present four WIV’s: \( \hat{\beta}_{wsp}^2 \) with weight \( \hat{\omega}^1(X) \) and \( \hat{\omega}^2(X) \), and \( \hat{\beta}_{wsp}^2 \) with weight \( \hat{\omega}^1(X) \) and \( \hat{\omega}^2(X) \); the difference between \( \hat{\beta}_{wsp}^2 \) and \( \hat{\beta}_{wsp}^2 \) is in using \( E(\delta|X) = X'\psi \) or \( E(\delta|X) = \Phi(X'\psi), \) although both were denoted collectively just as \( \hat{\beta}_{wsp}^2 \) in (4.6) for simplicity. These four WIV’s are denoted as \( \hat{\beta}_{wsp}^{21}, \hat{\beta}_{wsp}^{22}, \hat{\beta}_{wsp}^{21}, \) and \( \hat{\beta}_{wsp}^{22}. \) For WIV, we do trimming with \(|\hat{\omega}(X)| > 0.01. \)

In the right panel of Table 1, we try the above basic design, where OLS is highly biased as expected and \( \hat{\beta}_{psr}^2 \) is even more so. The performance ranking is, with ‘<’ standing for ‘worse than’,

\[ \hat{\beta}_{psr} < \text{ols} < \hat{\beta}_{wsp}^{21} = \hat{\beta}_{wsp}^{21} < \hat{\beta}_{wsp}^{22} = \hat{\beta}_{wsp}^{22} < \text{ive} = \hat{\beta}_{sp} = \hat{\beta}_{sp}. \]
Because the four Sub-PS-based IVE’s look the same in all designs, one may suspect that they are exactly the same, but this apparent equality is due to presenting only two digits: they do differ in the third or fourth digit.

<table>
<thead>
<tr>
<th></th>
<th>Exogenous $D$: $\text{Cor}(\varepsilon, U) = 0$</th>
<th>Endogenous $D$: $\text{Cor}(\varepsilon, U) \simeq 0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N = 500$</td>
<td>$N = 1500$</td>
</tr>
<tr>
<td></td>
<td>$N = 500$</td>
<td>$N = 1500$</td>
</tr>
<tr>
<td>ols</td>
<td>0.07 0.09 0.09</td>
<td>0.04 0.05 0.05</td>
</tr>
<tr>
<td></td>
<td>0.50 0.08 0.09</td>
<td>1.27 0.12 0.14</td>
</tr>
<tr>
<td>$\hat{\beta}^2_{psr}$</td>
<td>0.12 0.16 0.16</td>
<td>0.07 0.09 0.09</td>
</tr>
<tr>
<td></td>
<td>1.27 0.12 0.14</td>
<td>0.10 0.12 0.12</td>
</tr>
<tr>
<td>$\hat{\beta}^2_{ive}$</td>
<td>0.10 0.12 0.12</td>
<td>0.06 0.07 0.07</td>
</tr>
<tr>
<td></td>
<td>0.10 0.12 0.12</td>
<td>0.10 0.12 0.12</td>
</tr>
<tr>
<td>$\hat{\beta}^3_{sp}$</td>
<td>0.10 0.12 0.12</td>
<td>0.06 0.07 0.07</td>
</tr>
<tr>
<td></td>
<td>0.10 0.12 0.12</td>
<td>0.10 0.12 0.12</td>
</tr>
<tr>
<td>$\hat{\beta}^2_{sp}$</td>
<td>0.10 0.12 0.12</td>
<td>0.06 0.07 0.07</td>
</tr>
<tr>
<td></td>
<td>0.10 0.12 0.12</td>
<td>0.10 0.12 0.12</td>
</tr>
<tr>
<td>$\hat{\beta}^3_{sp}$</td>
<td>0.10 0.12 0.12</td>
<td>0.06 0.07 0.07</td>
</tr>
<tr>
<td></td>
<td>0.10 0.12 0.12</td>
<td>0.10 0.12 0.12</td>
</tr>
<tr>
<td>$\hat{\beta}^2_{wsp}$</td>
<td>0.26 0.34 0.32</td>
<td>0.14 0.18 0.17</td>
</tr>
<tr>
<td></td>
<td>0.27 0.36 0.33</td>
<td>0.14 0.18 0.17</td>
</tr>
<tr>
<td>$\hat{\beta}^2_{wsp}$</td>
<td>0.12 0.15 0.15</td>
<td>0.07 0.08 0.08</td>
</tr>
<tr>
<td></td>
<td>0.12 0.16 0.15</td>
<td>0.07 0.08 0.08</td>
</tr>
<tr>
<td>$\hat{\beta}^2_{wsp}$</td>
<td>0.26 0.34 0.36</td>
<td>0.14 0.18 0.19</td>
</tr>
<tr>
<td></td>
<td>0.27 0.36 0.36</td>
<td>0.14 0.18 0.19</td>
</tr>
<tr>
<td>$\hat{\beta}^2_{wsp}$</td>
<td>0.12 0.15 0.21</td>
<td>0.07 0.08 0.12</td>
</tr>
<tr>
<td></td>
<td>0.12 0.16 0.21</td>
<td>0.07 0.08 0.12</td>
</tr>
</tbody>
</table>

good overlap with $\alpha_3 = -1$; asy. Sd is avg. of 5,000 asy. Sd estimates;

ols,ive: OLS, IVE specifying $Y$ model; $\hat{\beta}^2_{psr}$ is order-2 OLS with PS residual;

$\hat{\beta}^2_{sp}$, $\hat{\beta}^3_{sp}$: IVE with #–order polynomial for $E(Y|X\alpha) \& E(\delta|X) = X\psi, \Phi(X\psi)$;

$\hat{\beta}^2_{wsp}$, $\hat{\beta}^3_{wsp}$: WIV for $\hat{\beta}^2_{wsp}$ & $\hat{\beta}^3_{wsp}$ with weight $\omega(X)$.

In the left panel of Table 1, we try exogenous $D$ to show that the linear-model OLS and $\hat{\beta}^2_{psr}$ are consistent when $D$ is exogenous, which is indeed the case. A number of findings in this panel hold for all designs in this section. First, although $\hat{\beta}^2_{psr}$ does worse than the linear-model OLS when $D$ is exogenous, the four Sub-PS-based IVE’s ($\hat{\beta}^2_{sp}$, $\hat{\beta}^3_{sp}$, $\hat{\beta}^2_{sp}$ and $\hat{\beta}^3_{sp}$) do almost as well as the linear-model IVE despite that the four IVE’s do not specify the $Y$ model; the four IVE’s perform almost the same. Second, the two WIV’s $\hat{\beta}^{21}_{wsp}$ and $\hat{\beta}^{21}_{wsp}$ with $\omega^1(X)$ do almost the same, which is also the case for the other two WIV’s with $\omega^2(X)$. Third, the more elaborate weighting scheme $\omega^1(X)$ returns a worse performance than the simpler OLS-based $\omega^2(X)$. Fourth, WIV with $\omega^2(X)$ performs a little worse than the four Sub-PS-based IVE’s. Fifth, comparing the second and third numbers in each three-number entry, the
asymptotic variance estimators for the Sub-PS-based IVE’s and those for the WIV’s work well in small samples, although the former does a little better than the latter.

In the left panel of Table 2 still with good X-overlap, heterogenous effect is allowed with $X_2D$ added into the Y model, where $X_2 \sim U[0, 2]$ II $X_3$ differently from the base design; we let $N = 3000$ instead of 1500, as the improvement from $N = 500$ is not much visible with $N = 1500$. In this heterogeneous effect design, $\beta = E(1 + X_2) = 2$ but $\beta_{sp} \neq 2$, and thus we present $|\text{bias}|/2$, not $|\text{bias}|$. Somewhat surprisingly, all Sub-PS-based IVE’s perform still fine; it is shown shortly, however, that all Sub-PS-based IVE’s perform much worse in the poor X-overlap design. The above performance ranking still holds for the left panel of Table 2. In the right panel of Table 2, the Y model is misspecified by omitting $X_2X_3$ that enters the Y model with slope 1. Although this misspecification worsens all estimators’ performance, the above performance ranking stays the same.

| Table 2. | $|\text{Bias}|$, Sd, Asymptotic Sd for Good Overlap Design |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                  | Misspecified Y model (X_2X_3 omitted) |
|                  | N = 500          | N = 3000         | N = 500          | N = 1500         |
| ols              | 0.25 0.09 0.09   | 0.25 0.04 0.04   | 0.50 0.14 0.14   | 0.50 0.08 0.08   |
| $\tilde{\beta}_{psr}^2$ | 0.63 0.12 0.14   | 0.63 0.05 0.06   | 1.27 0.22 0.23   | 1.27 0.13 0.13   |
| ive              | 0.05 0.13 0.13   | 0.02 0.05 0.05   | 0.15 0.19 0.19   | 0.09 0.11 0.11   |
| $\beta_{sp}^2$   | 0.05 0.13 0.13   | 0.02 0.05 0.05   | 0.15 0.19 0.19   | 0.09 0.11 0.11   |
| $\beta_{sp}^3$   | 0.05 0.13 0.13   | 0.02 0.05 0.05   | 0.15 0.19 0.19   | 0.09 0.11 0.11   |
| $\tilde{\beta}_{sp}^2$ | 0.05 0.13 0.13   | 0.02 0.05 0.05   | 0.15 0.19 0.19   | 0.09 0.11 0.11   |
| $\tilde{\beta}_{sp}^3$ | 0.05 0.13 0.13   | 0.02 0.05 0.05   | 0.15 0.19 0.19   | 0.09 0.11 0.11   |
| $\tilde{\beta}_{wsp}^{21}$ | 0.12 0.30 0.28   | 0.04 0.11 0.11   | 0.33 0.44 0.40   | 0.17 0.22 0.21   |
| $\beta_{wsp}^{22}$   | 0.06 0.14 0.14   | 0.02 0.06 0.06   | 0.18 0.23 0.22   | 0.10 0.13 0.12   |
| $\tilde{\beta}_{wsp}^{21}$ | 0.12 0.30 0.30   | 0.04 0.11 0.12   | 0.33 0.44 0.43   | 0.17 0.22 0.23   |
| $\beta_{wsp}^{22}$   | 0.06 0.14 0.18   | 0.02 0.06 0.07   | 0.18 0.23 0.26   | 0.10 0.13 0.15   |

|-----------------|-----------------|-----------------|-----------------|-----------------|

In Tables 3 and 4, we try the poor overlap design with $\alpha_3 = 1$, which makes big differences.
from the good overlap design with \( \alpha_3 = -1 \). In the right panel of Table 3, all estimators do worse compared with Tables 1 and 2, but still the above performance ranking holds.

<p>| Table 3. [Bias], Sd, Asymptotic Sd for Poor Overlap Design |
|-------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Exogenous ( D: \ Cor(\varepsilon, U) = 0 ) | Endogenous ( D: \ Cor(\varepsilon, U) \simeq 0.7 ) |</p>
<table>
<thead>
<tr>
<th>( N = 500 )</th>
<th>( N = 1500 )</th>
<th>( N = 500 )</th>
<th>( N = 1500 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ols</td>
<td>0.09 0.11 0.11</td>
<td>0.05 0.06 0.06</td>
<td>0.59 0.10 0.10</td>
</tr>
<tr>
<td>( \hat{\beta}_{psr}^2 )</td>
<td>0.13 0.16 0.16</td>
<td>0.07 0.09 0.09</td>
<td>1.24 0.12 0.13</td>
</tr>
<tr>
<td>ive</td>
<td>0.14 0.18 0.18</td>
<td>0.08 0.10 0.10</td>
<td>0.14 0.18 0.18</td>
</tr>
<tr>
<td>( \hat{\beta}_{sp}^2 )</td>
<td>0.14 0.18 0.18</td>
<td>0.08 0.10 0.10</td>
<td>0.14 0.18 0.18</td>
</tr>
<tr>
<td>( \hat{\beta}_{sp}^3 )</td>
<td>0.14 0.18 0.18</td>
<td>0.08 0.10 0.10</td>
<td>0.14 0.18 0.18</td>
</tr>
<tr>
<td>( \tilde{\beta}_{sp}^2 )</td>
<td>0.14 0.18 0.18</td>
<td>0.08 0.10 0.10</td>
<td>0.14 0.18 0.18</td>
</tr>
<tr>
<td>( \tilde{\beta}_{sp}^3 )</td>
<td>0.14 0.18 0.18</td>
<td>0.08 0.10 0.10</td>
<td>0.14 0.18 0.18</td>
</tr>
<tr>
<td>( \hat{\beta}_{wsp}^{21} )</td>
<td>0.39 0.50 0.48</td>
<td>0.20 0.25 0.25</td>
<td>0.39 0.50 0.48</td>
</tr>
<tr>
<td>( \hat{\beta}_{wsp}^{22} )</td>
<td>0.20 0.25 0.25</td>
<td>0.11 0.14 0.14</td>
<td>0.20 0.25 0.25</td>
</tr>
<tr>
<td>( \tilde{\beta}_{wsp}^{21} )</td>
<td>0.39 0.50 0.48</td>
<td>0.20 0.25 0.26</td>
<td>0.39 0.50 0.48</td>
</tr>
<tr>
<td>( \tilde{\beta}_{wsp}^{22} )</td>
<td>0.20 0.25 0.25</td>
<td>0.11 0.14 0.14</td>
<td>0.20 0.25 0.25</td>
</tr>
</tbody>
</table>

poor overlap with \( \alpha_3 = 1 \); asy. Sd is avg. of 5,000 asy. Sd estimates;

ols,ive: OLS, IVE specifying \( Y \) model; \( \hat{\beta}_{psr}^2 \) is order-2 OLS with PS residual;

\( \hat{\beta}_{sp}^2, \hat{\beta}_{sp}^3 \): IVE with \#-order polynomial for \( E(Y\mid X') = \delta'X, \Phi(X') \);

\( \hat{\beta}_{wsp}^{21}, \hat{\beta}_{wsp}^{22}, \tilde{\beta}_{wsp}^{21}, \tilde{\beta}_{wsp}^{22} \): WIV for \( \hat{\beta}_{wsp}^2, \hat{\beta}_{wsp}^2 \) with weight \( \omega^#(X) \).

In the left panel of Table 4 with heterogenous effect, differently from the left panel of Table 2, the two WIV’s \( \hat{\beta}_{wsp}^{22} \) and \( \tilde{\beta}_{wsp}^{22} \) with \( \omega^2(X) \) are less biased than the four Sub-PS-based IVE’s, although their Sd’s are higher; again, \|bias\|/2 is shown, instead of \|bias\|. With \( N = 3000 \), the two WIV’s bias becomes more than halved, whereas the bias of the four Sub-PS-based IVE’s stays about the same. In the right panel with misspecified \( Y \) model, the linear-model IVE, which has been doing almost the same as the four Sub-PS-based IVE’s, is doing much worse than the four Sub-PS-based IVE’s.
Table 4. |Bias|, Sd, Asymptotic Sd for Poor Overlap Design

<table>
<thead>
<tr>
<th></th>
<th>Misspecified Y model (X_2X_3 omitted)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 500</td>
</tr>
<tr>
<td>ols</td>
<td>0.19 0.11 0.12</td>
</tr>
<tr>
<td>$\beta^2_{psr}$</td>
<td>0.57 0.13 0.15</td>
</tr>
<tr>
<td>ive</td>
<td>0.14 0.19 0.19</td>
</tr>
<tr>
<td>$\beta^2_{sp}$</td>
<td>0.14 0.19 0.19</td>
</tr>
<tr>
<td>$\beta^3_{sp}$</td>
<td>0.14 0.19 0.19</td>
</tr>
<tr>
<td>$\tilde{\beta}^2_{sp}$</td>
<td>0.14 0.19 0.19</td>
</tr>
<tr>
<td>$\tilde{\beta}^3_{sp}$</td>
<td>0.14 0.19 0.19</td>
</tr>
<tr>
<td>$\beta^2_{wsp}$</td>
<td>0.22 0.56 0.53</td>
</tr>
<tr>
<td>$\beta^2_{wsp}$</td>
<td>0.13 0.33 0.33</td>
</tr>
<tr>
<td>$\beta^3_{wsp}$</td>
<td>0.22 0.56 0.53</td>
</tr>
<tr>
<td>$\tilde{\beta}^2_{wsp}$</td>
<td>0.13 0.33 0.34</td>
</tr>
</tbody>
</table>

poor overlap with $\alpha_3 = 1$; asy. Sd is avg. of 5,000 asy. Sd estimates;
ols,ive: OLS, IVE specifying $Y$ model; $\beta^2_{psr}$ is order-2 OLS with PS residual;
$\beta^2_{sp}, \beta^3_{sp}$: IVE with $\#$-order polynomial for $E(Y | X, \alpha) & E(\delta | X) = X' \psi; \Phi(X' \psi)$;
$\tilde{\beta}^2, \tilde{\beta}^3$: WIV for $\tilde{\beta}^2_{wsp}$ & $\tilde{\beta}^2_{wsp}$ with weight $\omega^\#(X)$.

Summarizing this simulation section, first, Sub-PS-based IVE performs at least as well as the linear-model IVE. Second, between the two versions of Sub-PS-based IVE, the version with $E(\delta | X) = X' \psi$ is preferred to the version with $E(\delta | X) = \Phi(X' \psi)$, because the former is simpler yet performs as well as the latter despite binary $\delta$. Third, WIV with the simpler weight $\omega^2(X)$ is better than WIV with the more elaborate weight $\omega^1(X)$. Fourth, WIV does worse than Sub-PS-based IVE, but if one wants to avoid negative weights in $\beta_{sp}$ that occur when $\text{Cov}(\delta, D | X)$ is negative for some $X$, then WIV should be used. Fifth, $q = 2$ seems adequate for Sub-PS-based IVE and WIV.

7 Empirical Analysis

This section provides an empirical example using the National Longitudinal Survey data in Card (1995), which was also used in Tan (2010) and Okui et al. (2012); the data are downloadable from ‘http://davidcard.berkeley.edu/data_sets.html’. In our analysis with
\(N = 3010\), \(Y\) is the logarithm of wage in 1976, \(D\) is \(1[\text{schooling years} \geq 13]\) for some college education or higher, \(\delta\) for whether one grew up near a four year college or not, and \(X\) consists of age, the dummy for black, the dummies for 9 residence regions in 1966, the dummy for living in a standard metropolitan statistical area (SMSA) in 1966, the dummy for living in SMSA in 1976 (“SMSA76”), and the dummy for living in South in 1976 (“south”). Although the dummy for living in South in 1966 is also available in the data, it is not used because it is perfectly collinear with the 9 residence region dummies in 1966.

Our setup differs somewhat from that in Okui et al. (2012) because we need \(D\) to be binary, which is why we defined \(D\) as above, not schooling years as in Okui et al. (2012). Compared with Okui et al. (2012), we include age in \(X\) while excluding experience, because it is not clear how experience was obtained in Okui et al. (2012): if it is ‘age-schooling-6’ as typically done in applied works, then experience is likely to be endogenous due to schooling included in its definition.

With schooling years as treatment, some of the findings for the effect (t-value) of one additional schooling year in Okui et al. (2012) is

\[
\text{OLS} : 0.075 (25) \quad \text{IVE} : 0.13 (2.1); \quad (7.1)
\]

the DR estimates of Okui et al. (2012) range over \(0.13 \sim 0.17\) with t-values \(0.95 \sim 1.87\).

Before we proceed further, we note a dilemma in using SMSA76 and south as covariates, because they might have been affected by \(D\), and if so, we would be controlling covariates affected by the treatment. On the other hand, if we do not control them, they may cause another problem because they are possibly related to \(D\) (and \(\delta\) as well). Hence, we present two results with and without SMSA76 and south. If anything, controlling them is better than not, because at least the ‘direct effect’ of \(D\) on \(Y\) is identified although the ‘indirect effect’ of \(D\) on \(Y\) through SMSA76 and south is missed.

Table 5 presents our main results, where the left panel is for using SMSA76 and south, and the right panel is for not using them. In the left panel, the OLS estimate (0.13) is virtually the same as \(\hat{\beta}_{psr}^2\). Since the average schooling years among those with \(D = 1\) is 15.4 years, which is 3.4 years beyond high school completion, we can compare 0.13 to 0.075 \times 3.4 = 0.26 from (7.1): the OLS effect magnitude 0.13 is half of what (7.1) indicates. The linear model IVE effect (0.43) in the left panel is much greater than the OLS, and almost statistically significant. Multiplying the IVE estimate in (7.1) by 3.4 to get 0.13 \times 3.4 = 0.44, this
magnitude is almost the same as the linear model IVE effect 0.43. The four Sub-PS-based IVE’s are similar to one another (0.32 ~ 0.36), falling a little short of statistical significance; the effects are smaller than the linear model IVE effect 0.43, and larger than the OLS effect 0.13.

When SMSA76 and south are not used in the right panel, the linear-model OLS and $\hat{\beta}_{psr}^2$ remain almost the same as in the left panel, but the linear model IVE and the Sub-PS-based IVE’s increase much to become very similar to one another, falling in 0.51 ~ 0.55; they are all statistically significant, differently from the left panel.

Table 5. Estimates (T-value) for College Education Effect ($N = 3010$)

<table>
<thead>
<tr>
<th></th>
<th>SMSA76 &amp; South included in X</th>
<th>SMSA76 &amp; South Excluded from X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau = 0.01$     $\tau = 0.02$ $\tau = 0.03$</td>
<td>$\tau = 0.01$ $\tau = 0.02$ $\tau = 0.03$</td>
</tr>
<tr>
<td>ols</td>
<td>0.13 (8.8)</td>
<td>0.14 (9.5)</td>
</tr>
<tr>
<td>$\hat{\beta}_{psr}^2$</td>
<td>0.13 (8.7)</td>
<td>0.13 (9.2)</td>
</tr>
<tr>
<td>ive</td>
<td>0.43 (1.8)</td>
<td>0.55 (2.5)</td>
</tr>
<tr>
<td>$\hat{\beta}_{sp}^2$</td>
<td>0.34 (1.5)</td>
<td>0.52 (2.4)</td>
</tr>
<tr>
<td>$\hat{\beta}_{sp}^3$</td>
<td>0.36 (1.6)</td>
<td>0.55 (2.5)</td>
</tr>
<tr>
<td>$\tilde{\beta}_{sp}^2$</td>
<td>0.32 (1.4)</td>
<td>0.51 (2.3)</td>
</tr>
<tr>
<td>$\tilde{\beta}_{sp}^3$</td>
<td>0.34 (1.4)</td>
<td>0.53 (2.4)</td>
</tr>
<tr>
<td>$\tilde{\beta}_{omega}^{21}$</td>
<td>0.25 (0.54)</td>
<td>0.39 (1.12) 0.53 (1.61) 0.21 (-0.37) 0.16 (0.18) 0.35 (0.56)</td>
</tr>
<tr>
<td>$\tilde{\beta}_{omega}^{22}$</td>
<td>2.79 (0.34)</td>
<td>0.01 (0.03) 0.29 (0.80) 1.53 (0.72) 0.11 (0.39) 0.05 (0.15)</td>
</tr>
<tr>
<td>$\tilde{\beta}_{omega}^{21}$</td>
<td>0.38 (0.81)</td>
<td>0.47 (1.29) 0.60 (1.71) -0.09 (-0.17) 0.36 (0.41) 0.47 (0.75)</td>
</tr>
<tr>
<td>$\tilde{\beta}_{omega}^{22}$</td>
<td>3.83 (0.24)</td>
<td>-0.03 (-0.08) 0.23 (0.59) 1.57 (0.69) 0.09 (0.30) -0.02 (-0.06)</td>
</tr>
</tbody>
</table>

ols,ive: OLS, IVE specifying Y model; $\hat{\beta}_{psr}^2$ is order-2 OLS with PS residual;
$\hat{\beta}_{sp}^#$, $\tilde{\beta}_{sp}^#$: IVE with #-order polynomial for $E(Y|X'\alpha) & E(\delta|X) = X'\psi; \Phi(X'\psi);$
$\tilde{\beta}_{omega}^{2#}$, $\tilde{\beta}_{omega}^{2#}$: WIV for $\hat{\beta}_{omega}^2$ & $\tilde{\beta}_{omega}^2$ with weight $\omega^#(X)$.

One may wonder if the treatment effect is heterogeneous or not. After some searching, we found that $D$ interacts possibly with age or south: with $D \times age$ and $D \times south$ added, in the linear model IVE using ($\delta$, $\delta \times age$, $\delta \times south$) as instruments for ($D$, $D \times age$, $D \times south$), the two $D$-interacting terms have the following estimates (t-values):

(SMSA76, south) in : 0.070 (1.67) for $D \times age$, 0.96 (2.4) for $D \times south$;
(SMSA76, south) out : 0.072 (1.95) for $D \times age$, -0.11 (-1.4) for $D \times south$. 

20
Using Sub-PS-based IVE saves this kind of efforts needed to search for treatment effect heterogeneity source.

Table 5 reveals that the WIV’s are highly sensitive to the trimming constant $\tau$ as $\tau$ changes over $0.01 \sim 0.03$. Also, the WIV’s are statistically insignificant. This is disappointing, given the fairly good performance exhibited by the WIV’s in the simulation section. Searching for the reason why the WIV’s do poorly, we examined descriptive statistics of $\hat{\omega}^1(X)$ and $\hat{\omega}^2(X)$ in Table 6, which shows that many weights are negative, despite that the slope of $\delta$ is significantly positive in the probit estimation of $D$ on $(\delta, X)$. Since $\text{Cor}(\delta, D|X)$ is likely to be more positive than negative, $\hat{\omega}^2(X)$ seems preferred to $\hat{\omega}^1(X)$ in this regard with the proportion of positive weights ($P(\cdot > 0)$ in Table 6) being higher. Essentially, the problem is that the instrument is “weak” for many values of $X$; otherwise, $\text{Cor}(\delta, D|X)$ would be far from 0 to make the WIV’s insensitive to $\hat{\omega}^1(X)$ and $\hat{\omega}^2(X)$.

Table 6. Distribution of Weight Estimates: $P(\cdot > 0)$ for positive weight proportion

<table>
<thead>
<tr>
<th></th>
<th>SMSA76 &amp; South included in $X$</th>
<th>SMSA76 &amp; South Excluded from $X$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (Sd)</td>
<td>Min, Max</td>
</tr>
<tr>
<td>$\hat{\omega}^1(X)$</td>
<td>0.002 (0.12)</td>
<td>-0.15, 0.31</td>
</tr>
<tr>
<td>$\hat{\omega}^2(X)$</td>
<td>0.012 (0.034)</td>
<td>-0.082, 0.11</td>
</tr>
</tbody>
</table>

8 Conclusions

When a binary treatment $D$ is endogenous, i.e., when there are unobserved confounders, popular procedures controlling covariates $X$ such as matching, stratification, inverse probability weighting and regression imputation are no more applicable, because they require treatment exogeneity for response $Y$. If there is an instrument (IV) $\delta$ for $D$, then instrumental variable estimator (IVE) can be applied, which, however, fully specifies the structural-form $Y$ model as a function of $(D, X)$.

In this paper, we proposed an IVE using ‘sub-propensity score (Sub-PS) residual’ as the regressor, instead of $D$; the sub-propensity score is $E(D|X)$ whereas the usual propensity score (PS) is $E(D|\delta, X)$. The IVE is semiparametric because it specifies $E(D|X)$ and $E(\delta|X)$ parametrically with probit/logit to make the IVE practical, but the IVE is nonparametric otherwise because it does not specify how $Y$ depends on $(D, X)$. The IVE allows $X$ to affect directly $Y$ through an unknown function, and allows an unknown form of heterogeneous
treatment effect, say $\mu_1(X) \equiv E(Y^1 - Y^0|X)$, where the $Y^d$'s are the potential responses. For a randomized dummy $\delta$, the IVE does not specify $E(\delta|X)$ because $E(\delta|X) = E(\delta)$.

As it turns out, Sub-PS-based IVE is consistent for $E\{\omega_{sp}(X)\mu_1(X)\}$, not for the usual effect on the population $E\{\mu_1(X)\} = E(E(Y^1 - Y^0|X)) = E(Y^1 - Y^0)$, where the weighting function $\omega_{sp}(X)$ is proportional to $Cov(\delta, D|X)$. The $\omega_{sp}(X)$ weighting can be good when $\delta$ is a randomization dummy, because $\omega_{sp}(X)$ gives high weights to those with $E(D|X) \approx 0.5$ whose $X$ plays no role for $D$, and low weights to extreme observations with $E(D|X) \approx 0, 1$. That is, $E\{\omega_{sp}(X)\mu_1(X)\}$ is a legitimate parameter of interest; in fact, IVE in general tends to weight this way. Despite this, if desired, we can do a weighted IVE using an estimated $Cov(\delta, D|X)^{-1}$ as the weight to estimate $E\{\mu_1(X)\}$.

A simulation study demonstrated that Sub-PS-based IVE and its weighted version work fine as they are supposed to. In our empirical example, Sub-PS-based IVE gave effect estimates that either almost equal the linear model IVE, or fall between the linear model IVE and the linear model OLS. The weighted version, however, turned out to be too sensitive to the trimming used in weighting due to the instrument being “weak”.

*Given that there is no practical estimator to find treatment effects when the treatment is endogenous and the effect is heterogeneous in an unknown way, the main contribution of this paper is proposing a simple and highly practical IVE along its weighted version for such cases, which also have asymptotic variance estimators that work well in small samples. We expect the estimators to be particularly useful for randomized experiments with non-compliance problems because the randomization dummy becomes an automatic IV. Even if there is no randomization, still natural- or quasi-experimental situations arise in observations data, for which Sub-PS-based IVE and its weighted version may be fruitfully applied.*

**Appendix for (3.9) and (3.10)**

The IVE in (3.9) satisfies

$$\frac{1}{\sqrt{N}} \sum_i m(\hat{\alpha}, \hat{\gamma}_{sp}, \hat{\psi}, \hat{\gamma}) = 0$$

where

$$m(a, b, c, g) \equiv (\delta - X') [Y - g_0 - g_1 X'a - g_2 (X'a)^2 - b(D - \Phi(X'a))].$$

Expanding the moment condition (times $\sqrt{N}$) around $\beta_{sp}$ gives, for some $\hat{\gamma}_{sp}^2 \in (\beta_{sp}^2, \beta_{sp})$,

$$0 = \frac{1}{\sqrt{N}} \sum_i m(\hat{\alpha}, \hat{\beta}_{sp}, \hat{\psi}, \hat{\gamma}) + \frac{1}{N} \sum_i \frac{\partial m(\hat{\alpha}, \hat{\beta}_{sp}, \hat{\psi}, \hat{\gamma})}{\partial \gamma} \sqrt{N}(\hat{\beta}_{sp} - \beta_{sp}).$$
From the expansion, we get

\[ \sqrt{N}(\hat{\beta}_p^2 - \beta_p) = -E^{-1}\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \psi} \} \left[ \frac{1}{\sqrt{N}} \sum_i m(\alpha, \beta_{sp}, \psi, \gamma) \right] \]

+ \left[ E\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \alpha'} \} \sqrt{N}(\hat{\alpha} - \alpha) \right] + \left[ E\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \beta'} \} \sqrt{N}(\hat{\beta}_p - \beta) \right] + o_p(1) \sim N(0, \Omega^0_{sp})

where \( \Omega^0_{sp} = E^{-1}\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \psi} \} E(\eta') E^{-1}\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \psi} \}, \)

\[ \eta \equiv m(\alpha, \beta_{sp}, \psi, \gamma) \]

+ \left[ E\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \alpha'} \} \eta_\alpha \right] + \left[ E\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \beta'} \} \eta_\beta \right] + \left[ E\{ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \gamma'} \} \eta_\gamma \right]

where \( \eta_\alpha, \eta_\beta \) and \( \eta_\gamma \) are influence functions for \( \hat{\alpha}, \hat{\psi} \) and \( \hat{\gamma} \).

As for the derivatives, we have

\[ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \psi} = -(\delta - X'c)\varepsilon, \quad V \equiv Y - \gamma_0 - \gamma_1(X'\alpha) - \gamma_2(X'\alpha)^2 - \beta_{sp}\varepsilon, \]

\[ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \alpha'} = (\delta - X'c)\{-\gamma_1 - \gamma_22X'\alpha + \beta_{sp}\phi(X'\alpha)\}X', \]

\[ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \beta'} = -VX', \quad \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \gamma'} = -(\delta - X'c)W', \quad W \equiv \{1, (X'\alpha), (X'\alpha)^2\}'. \]

Since \( E(\delta - X'\psi|X) = 0 \) whereas \( E(V|X'=0) = 0 \), not necessarily \( E(V|X) = 0 \), the two “link matrices” in front of \( \eta_\alpha, \eta_\beta \) and \( \eta_\gamma \) are zero to render

\[ \eta = m(\alpha, \beta_{sp}, \psi, \gamma) + E(-VX')\eta_\psi = (\delta - X'c)V + E(-VX')\eta_\psi. \]

This gives the asymptotic variance in (3.9).

Turning to (3.10) adopting \( E(\delta|X) = \Phi(X'\psi) \) for binary \( \delta \), we have

\[ m(a, b, c, g) \equiv \{ \delta - \Phi(X'c) \} \left[ Y - g_0 - g_1X'c - g_2(X'c)^2 - b \{ D - \Phi(X'a) \} \right], \]

\[ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \psi} = -\{ \delta - \Phi(X'c) \} \varepsilon, \]

\[ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \alpha'} = \{ \delta - \Phi(X'c) \} \{-\gamma_1 - \gamma_22X'\alpha + \beta_{sp}\phi(X'\alpha)\}X', \]

\[ \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \beta'} = -\phi(X'\psi)X', \quad \frac{\partial m(\alpha, \beta_{sp}, \psi, \gamma)}{\partial \gamma'} = -\{ \delta - \Phi(X'\psi) \} W' \]

\[ \implies \eta = \{ \delta - \Phi(X'c) \} V + E\{-\phi(X'\psi)X'\} \eta_\psi. \]

which leads to the asymptotic variance in (3.10).
REFERENCES


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


Lee, M.J., 2018, Simple least squares estimator for treatment effects using propensity score residuals, Biometrika 105, 149-164.


Rosenbaum, P.R., 2002, Observational studies, 2nd ed., Springer.
