University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2016 *Paper* 121

A Weighted Instrumental Variable Estimator to Control for Instrument-Outcome Confounders

Douglas Lehmann* Yun Li†

Rajiv Saran‡ Yi Li**

*The University Of Michigan, lehmannd@umich.edu

†University of Michigan School of Public Health, yunlisph@umich.edu

‡University of Michigan School of Public Health, rsaran@med.umich.edu

**University of Michigan School of Public Health, yili@med.umich.edu

A Weighted Instrumental Variable Estimator to Control for Instrument-Outcome Confounders

Douglas Lehmann, Yun Li, Rajiv Saran, Yi Li

April 10, 2016

Abstract

Instrumental variable (IV) methods are widely used to obtain consistent effect estimates in the presence of unmeasured treatment-outcome confounding, but rely on assumptions that are hard to make and often criticized. Among these is the assumption that the instrument is randomly assigned, which implies that there is no instrument-outcome confounding. This is easy to justify for instruments based on a random process. However, such instruments are rarely available in observational studies and it is more common to find an instrument that meets this assumption only after controlling for various measured confounders.

In this work we develop a weighted estimator based on the IV propensity score to adjust for measured instrument-outcome confounders. The proposed weights reflect the probability that an individual would be selected into a one-to-one match on the IV propensity score. Compared with matching methods, the proposed estimator is more efficient, allows for straightforward variance estimation, and is faster computationally. We study the performance of the estimator through simulation and illustrate its use in a study comparing dialysis session length and mortality in patients undergoing hemodialysis as treatment for end stage renal disease.

1 Introduction

Instrumental variable (IV) methods are widely used to deal with the selection bias or unmeasured confounding often present in observational studies. While IV models can obtain consistent estimates in the presence of unmeasured confounding, they rely on assumptions that are often criticized. A key component of an IV analysis is the instrument, a variable that is believed to encourage individuals toward the treatment or control. The instrument is assumed to be correlated with the treatment, have no direct effect on the outcome outside of its effect on the treatment, and be randomly assigned (Angrist et al., 1996; Baiocchi et al., 2014). The assumption that the instrument is randomly assigned implies that there are no instrument-outcome confounders. This is easily justified when an instrument is based on actual randomization, for example using the treatment a subject is randomly assigned in a randomized trial as an instrument for the treatment a subject ultimately receives. Such instruments are rarely available in observational studies, however, and it is more common to find an instrument that meets this requirement only after controlling for a set of measured instrument-outcome confounders. In other words, the instrument is conditionally distributed "as good as random." For example, regional treatment preference may be a reasonable instrument after controlling for patient characteristics such as race, age, education, income, insurance status and comorbidities, geographic characteristics such as rural/urban status, socioeconomic indicators, and provider characteristics such as procedure volume, supply, and profit or teaching status. violate the assumption that the instrument is randomly assigned. Garabedian et al. (2014) discuss the most commonly used instruments and potential instrument-outcome confounders associated with each, and emphasize that failing to adjust for these can bias estimation.

There are several approaches to controlling for instrument-outcome confounders. They can be included as covariates in two stage regression models. While two stage least squares is

the most common among these, it may be inappropriate for binary outcomes (Bhattacharya et al., 2006). Two stage residual inclusion was proposed in Terza et al. (2008) for use with binary outcomes. Matching on confounders is a common nonparametric alternative to these regression methods, but becomes difficult when there are many confounders or confounders with many discrete levels. Though less common in practice, methods have proposed using the IV propensity score rather than the full set of confounders. Methods using the IV propensity score include the inverse probability weighting estimator of Tan (2006), the matching estimator of Frölich (2007), and the weighting and subclassification methods of Cheng and Lin (2013).

The IV propensity score is the probability that an individual is encouraged, as indicated by their instrument value, toward the treatment. This differs from the usual propensity score, which represents the probability that an individual actually receives the treatment. Like the usual propensity score, the IV propensity score balances the distribution of confounders across instrument groups while reducing the dimension of the adjustment problem (Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004). Unlike the treatment propensity score, which is only used to address measured treatment-outcome confounding, methods based on the IV propensity score can provide consistent effect estimates in the presence of both measured and unmeasured treatment-outcome confounding.

In this work we propose an IV estimator based on the IV propensity score. It is a weighted estimator that uses weights designed to reflect the probability of being selected into a one-to-one IV propensity score match. One-to-one IV propensity score matching involves pairing encouraged subjects to unencouraged subjects with similar scores, usually within a specified range. Often there is no match within this range. Pairing the encouraged subject to an unencouraged subject with a score outside of this range can bias estimation, whereas dropping the encouraged subject reduces sample size and leads to a decrease in efficiency. The proposed estimator avoids this pitfall associated with matching. Weighted estimators also

allow for straightforward variance estimation, whereas the correlation structures introduced by matching algorithms are difficult to account for when estimating the variance of matching estimators (Austin, 2008, 2009, 2011a). Finally, weighting is computationally more efficient than matching.

We further discuss two extensions to the proposed estimator that could prove useful in practice. The first is a modification to the weight function to approximate k:1 matching designs rather than being restricted to a 1:1 design. The second is an alternative formulation of the estimator that provides protection against misspecification of the IV propensity score model. Though this requires the additional specification of an outcome model, this double robust estimator will give consistent estimates if at least one of the IV propensity score or outcome models is correctly specified.

The remainder of this article is organized as follows. In section 2 we define notation, discuss the IV propensity score, and introduce our proposed estimator and the two extensions. Finite-sample performance is reported through simulations in section 3. We illustrate use of the method with a data analysis in section 4, and conclude with a discussion in section 5.

2 Methods

2.1 Notation

We define causal effects using potential outcomes notation (Rubin, 1974; Neyman, 1923; Angrist et al., 1996). For each of $i = 1, ..., n$ subjects, let $Z_i = 1$ if subject i is encouraged toward treatment and $Z_i = 0$ otherwise. Let $D_i(Z_i)$ indicate treatment received for subject i given their encouragement, and let $Y_i(Z_i, D_i)$ indicate the response for subject i given their encouragement and treatment values. $D_i(Z_i)$ and $Y_i(Z_i, D_i)$ are referred to as a subjects potential outcomes. When subject i is encouraged toward treatment, we observe treatment $D_i(1)$ and response $Y_i(1, D_i)$ from subject i, otherwise we observe treatment $D_i(0)$ and

response $Y_i(0, D_i)$. Our interest is in estimating the parameter

$$\lambda = \frac{E(Y_i(1, D_i) - Y_i(0, D_i))}{E(D_i(1) - D_i(0))}. \tag{1}$$

This is the ratio of the instruments effect on the response to its effect on the treatment, and is often referred to as the local average treatment effect (LATE) (Imbens and Angrist, 1994; Angrist et al., 1996). Rather than an average treatment effect over the entire population, the LATE is interpreted as an average effect over a subgroup of the population known as compliers. Depicted in Table 1, compliers are individuals that take the treatment they are encouraged toward, and are one of four population subgroups defined by their response to encouragement.

Table 1: Population subgroups defined by the effect of encouragement on treatment. $D(1)$ denotes the treatment a subject will receive if they are encouraged toward treatment, while $D(0)$ denotes the treatment they will receive if they are encouraged toward the control.

		$D(1)$	
		1	0
$D(0)$	1	Always-takers	Defiers
	0	Compliers	Never-takers

The difficulty in estimating λ comes from the fact that we never observe individuals under encouragement and unencouragement, and thus never observe both of their potential outcomes. The data provides, for example, $E(Y_i(1, D_i)|Z_i = 1)$, or the average response under encouragement among encouraged subjects. However, this is not the same as $E(Y_i(1, D_i))$, which is an average response over the entire population if the entire population were observed under encouragement. To recover the expectations in equation (1), and to aid in the interpretation of λ, we make the following five assumptions (Angrist et al., 1996):

A1 - Stable Unit Treatment Value Assumption. The potential outcomes for one subject are unaffected by the potential outcomes, treatment assignment, or encouragement of other

subjects.

A2 - Random assignment of the instrument. The instrument is assumed to be randomly assigned, which implies that there are no unmeasured instrument-outcome confounders. This assumption is often stated conditional on measured instrument-outcome confounders.

A3 - Exclusion restriction. The instrument only affects the outcome through its effect on treatment. This implies that $Y_i(1, D_i = d) = Y_i(0, D_i = d)$ for all i and $d = 0, 1$.

A4 - Nonzero association between instrument and treatment. Assuming a nonzero association between the instrument and treatment implies that $E(D_i(1) - D_i(0)) \neq 0$.

A5 - Monotonicity. This assumption states that there are no defiers, or subjects that always do the opposite of what they are encouraged to do, and implies that $D_i(1) \geq D_i(0)$ for all i.

Assumptions *A1* and *A2* allow for unbiased estimation of the instruments effect on the outcome and treatment, or the numerator and denominator in equation (1). Assumption *A3*-*A5* are added to give a meaningful interpretation to λ. By exclusion restriction, always- and never-takers (Table 1) do not contribute to estimation since their treatment values, and thus response values, do not vary with encouragement. Monotonicity ensures that the group of defiers is empty, while a nonzero association between the instrument and treatment ensures that the group of compliers is not empty. Hence, with the addition of *A3*-*A5*, λ is an average treatment effect among compliers. Unlike the average treatment effect, which is applicable to the entire population, λ is a local effect that only applies to subjects that can be encouraged to switched treatment states, often termed "marginal patients." Further discussion of these assumptions can be found in Imbens and Angrist (1994), Angrist et al. (1996) or Baiocchi et al. (2014), among many others.

2.2 Proposed Estimator

In this section we propose an IV estimator using weights that are based on the IV propensity score. Defined as

$$e(\mathbf{x}) = P(Z = 1|\mathbf{X} = \mathbf{x}), \tag{2}$$

the IV propensity score is similar to the more common treatment propensity score, but represents the probability of receiving encouragement toward treatment rather than actually receiving treatment. From the theorems of Rosenbaum and Rubin (1983), we can say the distribution of covariates $\mathbf{X}$ is balanced across instrument groups conditional on $e(\mathbf{x})$, and if the instrument is independent of unmeasured confounders conditional on $\mathbf{X}$ then it is independent of unmeasured confounders conditional on $e(\mathbf{x})$. Taken together, these imply that conditioning on $e(\mathbf{x})$ is sufficient for adjusting for $\mathbf{X}$.

Define the observed treatment and response values for subject i as $D_i = Z_i D_i(1) + (1 - Z_i)D_i(0)$ and $Y_i = Z_i Y_i(1, D_i) + (1 - Z_i)Y_i(0, D_i)$, respectively. Our proposed estimator, which we refer to as the IV-matching weight (IV-MW) estimator, is given as

$$\lambda_{\text{IV-MW}} = \frac{\sum_i W_i Z_i Y_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) Y_i / \sum_i W_i (1 - Z_i)}{\sum_i W_i Z_i D_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) D_i / \sum_i W_i (1 - Z_i)}, \tag{3}$$

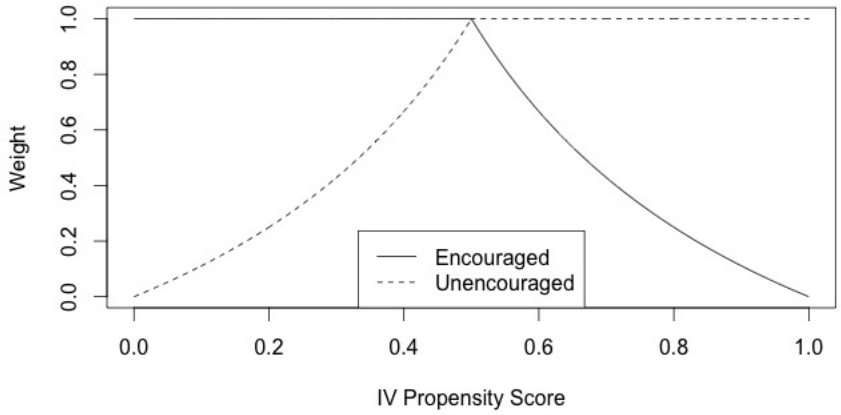
where weights W_i are defined as

$$W_i = \frac{\min(e_i(\mathbf{x}_i), 1 - e_i(\mathbf{x}_i))}{Z_i e_i(\mathbf{x}_i) + (1 - Z_i)(1 - e_i(\mathbf{x}_i))}. \tag{4}$$

W_i is similar to the matching weight of Li and Greene (2013) but defined with the IV propensity score, hence the term IV matching weight estimator. They are referred to as matching weights because they approximate the probability that an individual would be selected into a one-to-one match on the IV propensity score. We show that the asymptotic limits of the IV-MW and one-to-one IV propensity score matching estimators are equal

in the Appendix, but illustrate this idea here with a simple example. Consider a region around the IV propensity score $e = 0.1$ with $m = 100$ individuals. From equation (2), we expect approximately $me = 10$ encouraged and $m(1 - e) = 90$ unencouraged subjects in this region. Therefore, all encouraged subjects are expected to find a match, and $W_i = min(0.1, 0.9)/(1 \times 0.1 + 0 \times 0.9) = 1$. For unencouraged subjects, however, we expect only 10 of the 90 to find a match, and $W_i = min(0.1, 0.9)/(0 \times 0.1 + 1 \times 0.9) = 1/9$. Figure 1 shows the weight assigned to encouraged and unencouraged subjects across the range of IV propensity scores.

Figure 1: Matching weights for encouraged and unencouraged subjects by IV propensity score.



A benefit of the proposed estimator over one-to-one matching is that all subjects contribute a fraction of themselves to estimation, avoiding the situation where individuals are removed after not finding a suitable match. This avoids a decrease in sample size and efficiency. A benefit of the proposed estimator over inverse probability weighting using the IV propensity score (Tan, 2006) is that the weights are bounded between 0 and 1, whereas inverse probability weights can "blow up" near probabilities 0 or 1, causing an increase in variance of the estimate (Li et al., 2014). We thus expect the proposed estimator to be more efficient than both one-to-one matching and inverse probability weighting using the IV

propensity score.

An additional benefit of weighting over matching based estimators is that they allow for straightforward variance estimation. Matching algorithms introduce complicated correlation structures that are difficult to account for when estimating the variance of matching estimators, and often the matched nature of the data is ignored entirely (Austin, 2008, 2009, 2011a). Following Lunceford and Davidian (2004) and Li and Greene (2013), a sandwich type variance estimator is obtained using estimating equations

$$\mathbf{0} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \phi_{\mathbf{i}}(\theta) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \begin{bmatrix} W_i Z_i (Y_i - \mu_{y1}) \\ W_i (1 - Z_i)(Y_i - \mu_{y0}) \\ W_i Z_i (D_i - \mu_{d1}) \\ W_i (1 - Z_i)(D_i - \mu_{d0}) \\ \mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta}) \end{bmatrix}, \tag{5}$$

where $\theta = (\mu_{y1}, \mu_{y0}, \mu_{d1}, \mu_{d0}, \boldsymbol{\eta}')$, with $\mu_{y1} = E(W_i Z_i Y_i)/E(W_i Z_i)$, $\mu_{y0} = E(W_i(1-Z_i)Y_i)/E(W_i(1-Z_i))$ and similar for μ_{d1} and μ_{d0}. $\mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ represent estimating equations for coefficients $\boldsymbol{\eta}$ from the model used to estimate the IV propensity score, often a logistic regression. An estimate of $var(\hat{\theta})$ is obtained as $n^{-1}\hat{A}_n^{-1}\hat{B}_n(\hat{A}_n^T)^{-1}$, where $\hat{A}_n = \sum_{i=1}^n \partial\phi_i(\theta)/\partial\theta|_{\theta=\hat{\theta}}$ and $\hat{B}_n = \sum_{i=1}^n \phi_i(\theta)\phi_i^T(\theta)|_{\theta=\hat{\theta}}$. With an estimate of $var(\hat{\theta})$, the multivariate delta method can be applied for an estimate of $var(\hat{\lambda})$. This procedure allows for simultaneous estimation of the IV propensity score and λ, and is used for variance estimation for all estimators compared in Sections 3 and 4. For the IV propensity score matching procedure, this sandwich variance estimate ignores the matched nature of the sample. This typically leads to overestimated variance (Austin, 2009, 2011a), though in simulations reported in Section 3 estimated and empirical standard deviations are similar.

Note that W_i is not differentiable everywhere with respect to $\boldsymbol{\beta}$ due to the minimum function in the numerator. To apply the variance estimation procedure, rewrite the weight

function as

$$W_i = \frac{e_i(\mathbf{x}_i)I[e_i(\mathbf{x}_i) \leq 0.5] + (1 - e_i(\mathbf{x}_i))I[e_i(\mathbf{x}_i) > 0.5]}{Z_i e_i(\mathbf{x}_i) + (1 - Z_i)(1 - e_i(\mathbf{x}_i))}. \quad (6)$$

The indicator functions are then replaced with cumulative distribution functions to create a smooth, differentiable function for W_i (Horowitz, 1992).

In the following two sections we extend the IV-MW estimator in ways that could prove useful in practice. We first modify the weight function in equation (4) to approximate a $k:1$ match rather than a one-to-one match. We then modify equation (3) for a double robust IV-MW estimator that protects against misspecification of the IV propensity score model. This requires the additional specification of an outcome model but will give consistent estimates if at least one of the IV propensity score or outcome models is correctly specified.

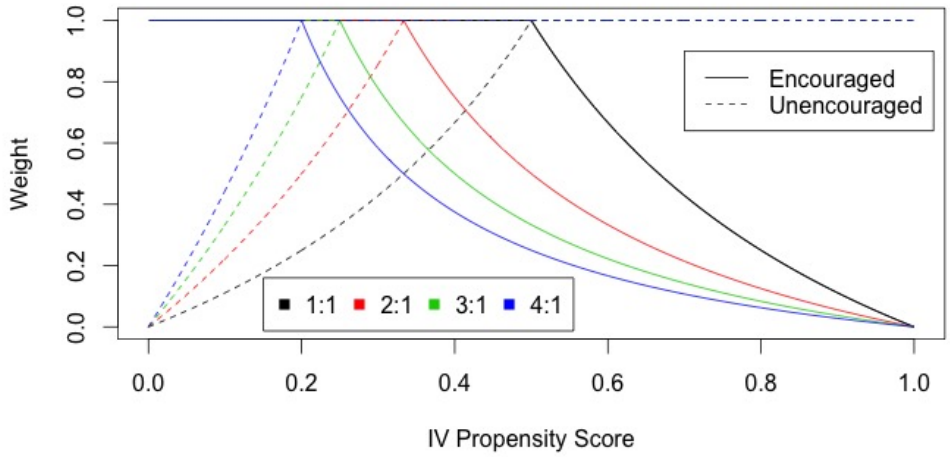
2.2.1 Extension to k:1 IV Matching

The weights in equation (4) are designed to approximate a one-to-one match on the IV propensity score. While one-to-one matching is the most common in practice (Austin, 2008), if the pool of unencouraged subjects is large enough we might consider matching multiple unencouraged subjects to each encouraged subject. Increasing the number of unencouraged subjects has the benefit of increasing the sample size and thereby decreasing the variability in estimation. We can extend the IV matching weight estimator in equation (3) to approximate k:1 matching designs with weights

$$W_i = \frac{\min(ke_i(\mathbf{x}_i), 1 - e_i(\mathbf{x}_i))}{Z_i ke_i(\mathbf{x}_i) + (1 - Z_i)(1 - e_i(\mathbf{x}_i))}. \quad (7)$$

As the number of unencouraged subjects to be matched increases, the probability that they will be selected into a match for any given IV propensity score increases, while decreasing the probability that encouraged subjects will be able to find k matches. Figure 2 graphs this weight function for up to 4:1 matching.

Figure 2: Weights for encouraged and unencouraged subjects across IV propensity scores for 1:1, 2:1, 3:1 and 4:1 matching designs. Weights for 1:1 matching are the same as Figure 1



2.2.2 Double Robust Estimation

The IV-MW estimator presented in equation (3) requires correct specification of the IV propensity score model for consistent estimation. Here we present a double robust version that protects against a misspecified IV propensity score model. While this requires the additional specification of an outcome model, the double robust (IV-MW$_{DR}$) estimator will provide consistent estimates if at least one of the IV propensity score or outcome models is correctly specified, but does not require both.

Let $m_0(\mathbf{X}_i) = E\{Y_i(0, D_i)|\mathbf{X}_i, Z_i = 0\}$ be the outcome model among the unencouraged group and $m_1(\mathbf{X}_i)$ be the outcome model among the unencouraged group. Following from Lunceford and Davidian (2004) and Li and Greene (2013), a double robust version of the estimator presented in (3) is obtained as

$$\lambda_{\text{IV-MW}_{DR}} = \frac{A + B - C}{\sum_i W_i Z_i D_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) D_i / \sum_i W_i (1 - Z_i)}, \tag{8}$$

where

$$A = \sum_i W_i\{m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i)\} / \sum_i W_i,$$
$$B = \sum_i W_i Z_i\{Y_i - m_1(\mathbf{X}_i)\} / \sum_i W_i Z_i,$$
$$C = \sum_i W_i(1 - Z_i)\{Y_i - m_0(\mathbf{X}_i)\} / \sum_i W_i(1 - Z_i).$$

Variance is estimated using the procedure discussed in section 2.2, with estimating equations

$$\mathbf{0} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \phi_{\mathbf{i}}(\theta) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \begin{bmatrix} W_i\{m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i) - \mu_A\} \\ W_i Z_i\{Y_i - m_1(\mathbf{X}_i) - \mu_B\} \\ W_i(1 - Z_i)\{Y_i - m_0(\mathbf{X}_i) - \mu_C\} \\ W_i Z_i(D_i - \mu_{d1}) \\ W_i(1 - Z_i)(D_i - \mu_{d0}) \\ \mathbf{S}_1(\boldsymbol{\alpha}_1) \\ \mathbf{S}_0(\boldsymbol{\alpha}_0) \\ \mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta}) \end{bmatrix}, \tag{9}$$

where $\theta = (\mu_A, \mu_B, \mu_C, \mu_{d1}, \mu_{d0}, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_0', \boldsymbol{\eta}')$. μ_A, μ_B, μ_C, μ_{d1}, and μ_{d0} correspond to the limits of A, B, C, and the averages in the denominator of equation (8). $\mathbf{S}_1(\boldsymbol{\alpha}_1)$ and $\mathbf{S}_0(\boldsymbol{\alpha}_0)$ represent the estimating equations for the parameters in $m_1(\mathbf{X}_i)$ and $m_0(\mathbf{X}_i)$, and $\mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ the estimating equations for the parameters in the IV propensity score model.

3 Simulation

3.1 Setup

We report simulation results to investigate the finite-sample performance of the proposed estimator (IV-MW). We compare with two alternatives that make use of the IV propensity score: the inverse probability weighting (IV-IPW) estimator of Tan (2006) and one-to-one IV propensity score matching (IV-PSM). The IV-IPW estimator has the same form as the IV-MW estimator in equation (3) but with the numerator of equation (4) replaced with 1. For the IV-PSM procedure, we match on the logit of the IV propensity score, using an optimal one-to-one match with a caliper of width equal to one fourth the standard deviation of logit of the IV propensity scores. For information about caliper selection, see Cochran and Rubin (1973), Raynor (1983), Rosenbaum and Rubin (1985), or Austin (2011b).

We generate 1,000 datasets with binary outcome, treatment, and instrument for $i = 1, ..., n$ individuals from

$$P(Y_i = 1|D_i, X_{1i}, X_{2i}) = logit^{-1}(\beta D_i + \delta_1 X_{1i} + \delta_2 X_{2i} + \epsilon_i^y), \quad (10)$$

$$P(D_i = 1|Z_i) = logit^{-1}(\gamma_0 + \gamma_1 Z_i + \epsilon_i^d), \quad (11)$$

$$P(Z_i = 1|X_{1i}, X_{2i}) = logit^{-1}(\psi_0 + \psi_1 X_{1i} + \psi_2 X_{2_i}). \quad (12)$$

(ϵ^d, ϵ^y) are drawn from a bivariate normal with correlation $\rho = .8$, creating the effect of unmeasured confounding between the treatment and the outcome. X_1 and X_2 represent instrument-outcome confounders and are randomly drawn from standard normal distributions. β is varied from 0 to 1 and sample size from 500 to 2,000. Results reported set parameters $\psi_0 = \gamma_0 = -1$, $\psi_1 = \delta_1 = -0.25$, $\gamma_1 = 1$ and $\psi_2 = \delta_2 = 0.25$.

3.2 Results

Estimation and coverage results are reported in Table 2. Each of the three estimators are found to be approximately unbiased. This is expected because we have adjusted for all instrument-outcome confounders (X_1 and X_2) in this simulation. Coverage rates for each method are converging to the nominal rate as sample size increases. While both weighted estimators have lower mean squared errors (MSE) than IV-PSM, the proposed IV-MW estimator has the lowest MSE in each scenario. The 2:1 and 3:1 matching scenarios confirm that the IV-MW estimator remains unbiased with the lowest MSE. A comparison of standard deviations in Table 3 confirm that the sandwich variance estimates (ASD) approximate the empirical standard deviations (ESD).

Table 2: Bias, MSE, and 95% coverage probabilities of IV-MW, IV-IPW, and IV-PSM for estimation of λ. Reported results are multiplied by 100.

				Weighting						Matching		
				IV-MW			IV-IPW			IV-PSM		
k	N	β	λ	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	CP
1:1	500	0.0	0.00	-0.39	6.11	97.4	-0.51	6.38	97.2	0.50	12.53	97.2
		0.5	0.12	-0.09	5.88	97.4	-0.38	6.17	97.1	1.05	9.42	98.2
		1.0	0.22	-0.65	5.18	97.9	-0.40	5.50	97.4	-0.13	8.61	98.3
1:1	1000	0.0	0.00	0.16	2.59	96.2	0.27	2.62	96.2	0.69	4.14	96.5
		0.5	0.12	0.42	2.50	96.9	0.67	2.57	96.5	0.70	3.84	97.0
		1.0	0.22	0.31	2.17	97.0	0.81	2.32	96.5	1.19	3.59	96.8
1:1	2000	0.0	0.00	-0.04	1.28	95.6	-0.03	1.32	95.3	0.38	1.85	95.4
		0.5	0.12	0.22	1.15	96.1	0.35	1.19	95.9	0.39	1.61	96.8
		1.0	0.22	-0.71	1.26	95.2	-0.27	1.32	94.6	-0.19	1.87	95.5
2:1	2000	0.00	-0.00	-0.10	1.25	96.1	-0.12	1.27	95.7	0.30	1.83	95.8
		0.50	0.12	0.51	1.29	94.8	0.46	1.31	94.7	0.38	1.84	95.1
		1.00	0.22	0.18	1.19	95.0	0.36	1.22	95.4	0.13	1.69	96.0
3:1	2000	0.00	0.00	-0.25	1.27	95.8	-0.31	1.27	96.3	-0.27	1.87	95.8
		0.50	0.12	0.52	1.21	95.0	0.51	1.22	95.2	0.55	1.72	96.4
		1.00	0.22	-0.05	1.16	95.7	-0.09	1.18	95.6	-0.16	1.65	97.5

We performed additional simulations to study the performance of the double robust IV-MW estimator (IV-MW$_{DR}$) proposed in Section 2.2.2. The generating equations for Y and

Table 3: Comparison of standard deviations obtained empirically (ESD) and using the sandwich variance technique of Section 2.2 (ASD). Reported results are multiplied by 100.

				Weighting				Matching	
				IV-MW		IV-IPW		IV-PSM	
k	N	β	λ	ASD	ESD	ASD	ESD	ASD	ESD
1:1	500	0.0	0.00	24.58	24.71	25.03	25.25	33.01	35.40
		0.5	0.12	23.97	24.26	24.68	24.84	31.22	30.67
		1.0	0.22	23.03	22.74	23.84	23.45	29.56	29.33
1:1	1000	0.0	0.00	16.37	16.08	16.62	16.19	20.06	20.35
		0.5	0.12	16.06	15.79	16.32	16.02	19.76	19.59
		1.0	0.22	15.40	14.74	15.79	15.21	19.03	18.91
1:1	2000	0.0	0.00	11.35	11.30	11.48	11.50	13.66	13.59
		0.5	0.12	11.14	10.73	11.34	10.90	13.45	12.69
		1.0	0.22	10.84	11.20	11.12	11.50	13.20	13.68
2:1	2000	0.0	-0.00	11.27	11.20	11.52	11.29	13.74	13.52
		0.5	0.12	11.00	11.35	11.31	11.43	13.42	13.54
		1.0	0.22	10.67	10.90	11.03	11.02	13.07	12.99
3:1	2000	0.0	0.00	11.34	11.26	11.52	11.28	13.71	13.67
		0.5	0.12	11.17	11.00	11.36	11.03	13.44	13.12
		1.0	0.22	10.85	10.78	11.05	10.84	13.09	12.85

Z are modified to include the interaction term X_1X_2 with coefficient 1. This interaction term is ignored for an incorrectly specified model.

Results in Table 4 confirm that IV-MW$_{DR}$ provides consistent estimates and maintains nominal coverage rates if at least one of the outcome or IV propensity score models is correctly specified. The original IV-MW estimator only provides unbiased estimates if the IV propensity score model is correctly specified, and the performance suffers greatly when it is misspecified. An interesting finding in Table 4 is that even though IV-MW$_{DR}$ requires the additional specification of an outcome model, performance is not damaged compared to IV-MW in situations in which the double robust property would not be needed, i.e. when the IV propensity score model is correctly specified.

The simulations reported throughout this section demonstrate that the proposed IV-MW estimator performs well compared with alternatives. It provided consistent estimates,

Table 4: Comparison of estimation using IV-MW and IV-MW$_{DR}$ under correctly and incorrectly specified IV propensity score and outcome models. Results reported are for $n = 1,000$ and $\beta = 1$ and are multiplied by 100.

P(Z)	P(Y)	Estimator	% Bias	MSE	ASD	ESD	95% CP
Correct	Correct	IV-MW	-2.50	0.56	7.52	7.47	94.9
		IV-MW$_{DR}$	-2.50	0.56	7.80	7.48	95.5
Correct	Incorrect	IV-MW	-3.12	0.52	7.57	7.18	96.3
		IV-MW$_{DR}$	-3.13	0.52	7.66	7.18	96.2
Incorrect	Correct	IV-MW	84.16	3.86	7.51	7.36	30.1
		IV-MW$_{DR}$	-6.18	0.55	7.66	7.33	95.8
Incorrect	Incorrect	IV-MW	85.14	3.96	7.49	7.56	30.6
		IV-MW$_{DR}$	87.65	4.17	7.57	7.56	29.6

achieved the lowest MSE, and maintained approximately nominal coverage. Both weighting estimators (IV-MW and IV-IPW) had lower MSE than the matching estimator (IV-PSM). IV-MW had lower MSE than IV-IPW, possibly because IV-IPW weights can "blow up" near probabilities of 0 or 1 (Li et al., 2014). In simulations, while IV-MW weights are bounded between 0 and 1, IV-IPW weights ranged from 1.05 to 20.5. Additionally, the IV-MW$_{DR}$ estimator proposed for protection against misspecification of the IV propensity score model performed as expected, providing consistent estimates if at least one of the IV propensity score or outcome models were correctly specified, but did not require both to be correct. The weighting estimators additionally saw computational benefits over IV propensity score matching. Using a MacBook Pro with a 2 GHz Intel Core i7 processor, we observed the following times (in seconds) to complete 1,000 simulations for n = 500, 1,000, and 2,000, respectively: IV-MW - 6.3, 9.9, and 23.6, IV-IPW - 6.0, 9.4, and 21.4, IV-PSM - 65.4, 166.1, and 547.9. While computing time was not an important issue in these simulations, it quickly becomes one for matching estimators as sample size increases.

4 Data Example

We illustrate use of these methods with data from the United States Renal Data System (USRDS) to study the association between dialysis session length and mortality among incident hemodialysis patients in the United States. It is thought that longer dialysis sessions decrease mortality risk by reducing the risk of intradialytic hypotension and and better controlling volume excess and serum phosphorous (Daugirdas, 2013), but this relationship is likely confounded. Shorter dialysis sessions are often prescribed to smaller patients, and smaller patients tend to have higher mortality rates. Several observational studies have found a significant increase in mortality in patients receiving shorter dialysis time (Flythe et al., 2013; Brunelli et al., 2010; Saran et al., 2006). However, a 2002 randomized trial found no significant relationship between dialysis time and mortality (Eknoyan et al., 2002), and Brunelli et al. (2010) found longer dialysis times to be associated with higher or lower mortality depending on whether the treatment was considered time dependent. These conflicting results suggest that unmeasured confounding may be present and an IV analysis may be useful.

We obtained complete data on 319,168 adults initiating hemodialysis (HD) between January 1, 2010 and December 31, 2013 from the USRDS database. We restricted the analysis to patients on a thrice-weekly dialysis schedule (98% of all incident HD patients). We conducted an intention-to-treat analysis, defined treatment as being prescribed dialysis sessions of four hours or longer, and defined the outcome as death in the first year. Mean treatment usage in the hospital service area (HSA) from 2007 to 2009 is used as the IV (Figure 3). An HSA is a geographic region representing a collection of zip codes whose residents receive most of their healthcare within that region (Dartmouth, 2016). Preference-based instruments such as this one are among the most common in health research (Garabedian et al., 2014), and are thought to measure treatment preferences that are independent of patient

level confounders (Brookhart and Schneeweiss, 2007; Li et al., 2015).

Among the 3,336 HSAs in the data, mean treatment usage varied from 0 to 100% with a mean of 74%. The correlation coefficient between the mean treatment usage from 2007-2009 and mean treatment usage from 2010-2013 in an HSA was almost 90%. This indicates that preferences in an HSA are relatively stable through time, and that mean treatment usage in an HSA from 2007-2009 is a strong instrument for treatment in the 2010-13 data. To fit the methods of this chapter, we dichotomize the instrument, considering HSAs with above average treatment usage to be encouraging subjects toward longer dialysis sessions and HSAs with below average usage to be encouraging their subjects toward shorter sessions.

Figure 3: Distribution of longer dialysis session usage by hospital service area. Longer dialysis sessions are defined as being prescribed sessions of four hours or more.

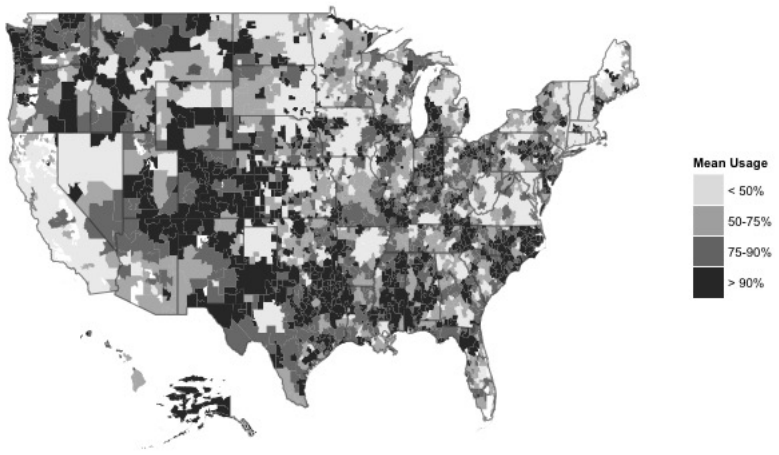


Table 5 reports the distribution of covariates by treatment and by instrument. Patients receiving longer dialysis sessions tend to have higher BMI and are more likely to be male, black, and younger compared with patients receiving shorter sessions. They are also more likely to be treated at for profit facilities in poorer, less educated areas. The improved covariate balance across instrument levels is evidence that HSA treatment usage may serve as a valid IV, although some imbalances remain in facility and zip code level variables.

Table 5: Distribution of covariates across treatment (long vs short dialysis sessions) and instrument groups (high vs low treatment usage at HSA level). Reported in the table is the mean and absolute standardized difference, d, between groups. An absolute standardized difference of more than 10 is generally considered to indicate an imbalance (Love, 2002).

	Hours on Dialysis			Usage in HSA		
	< 4	≥ 4	d	Low	High	d
Treatment						
4+ hour sessions	-	-	-	60.3%	88.8%	69.1
Outcome						
Death w/in 1st year	23%	21%	6.2	22%	21%	0.8
Patient Level Covariates						
Age	65.9	63.0	19.6	64.3	63.3	6.8
Male	51%	59%	16.7	57%	56%	2.1
BMI	28.0	30.4	28.2	29.4	30.1	8.3
Serum Creatinine	6.4	6.7	1.6	6.7	6.6	0.3
Hemoglobin	10.0	9.9	0.7	9.9	9.9	0.1
Black	23%	31%	16.6	26%	32%	12.42
Hispanic	17%	14%	8.4	17%	11%	10.9
Pre-ESRD 6+ Months	41%	44%	5.6	41%	45%	8.3
Employed	8.3%	9.1%	3.0	9.1%	8.8%	0.9
No Insurance	5%	8%	10.9	6%	8%	10.3
# Comorbidities	2.5	2.6	5.4	2.5	2.6	5.7
Facility Level Covariates						
# Nurses	7.3	7.3	0.3	7.7	7.0	12.2
# Patient techs	8.9	8.8	1.1	9.1	8.5	8.1
# HD stations	20.8	21.7	10.2	21.2	21.7	5.7
For profit	81%	86%	14.9	82%	88%	16.5
Zip Code Level Covariates						
Median income	$54,551	$49,286	25.3	$53,358	$47,960	26.9
Bachelors degree +	25.5%	22.9%	18.0	24.6%	22.6%	14.9

We first fit unadjusted and covariate adjusted logistic regression models. These indicate a significant decrease in the odds of first year mortality among patients with longer dialysis sessions, with estimated odds ratio and 95% confidence intervals of 0.86 (0.84, 0.87) and 0.95 (0.93, 0.97), respectively. Age, sex, race, ethnicity, BMI, number of comorbidities, access type, profit status of the facility and median income in the zip code were included in the covariate adjusted model. These logistic regressions will be biased if there are confounders between the treatment and the outcome that are not included in the model.

To implement the methods discussed in this article, we first model the IV propensity score, or the probability of being in an "encouraging" HSA. We use a logistic regression model including HSA level covariates mean age, BMI, number of comorbidities, percentage of males, blacks, hispanics, and patients without insurance, and median income. Using the estimated IV propensity score, a weight is assigned to each subject for the IV-MW and IV-IPW procedures. For the IV-MW procedure, this weight is given in equation (4). For the IV-IPW procedure the weight is defined similarly as equation (4), but with the numerator replaced with 1. For the IV-PSM procedure, we specified a one-to-one optimal match on the IV propensity score with a caliper of 0.05.

Table 6: Instrumental variable estimates and 95% confidence intervals for the effect of longer dialysis sessions on first year mortality. Negative estimates suggest less first year mortality among the patients receiving longer dialysis sessions.

	$\hat{\lambda}$	95% CI
IV-MW	-0.015	(-0.028, -0.002)
IV-IPW	-0.006	(-0.018, 0.006)
IV-PSM	-0.015	(-0.028, 0.001)

Results in Table 6 suggest a small protective effect of longer dialysis sessions. These results corroborate those found using logistic regression, although the estimated effects appear smaller and are insignificant for the IV-IPW and IV-PSM procedures. These estimates can be interpreted as follows; for $\hat{\lambda} = -0.015$, for example, we expect 1.5 less deaths in the first year for every 100 patients that could be encouraged to take long dialysis sessions. Note that IV-MW and IV-PSM gave similar results, with IV-MW obtaining a slightly shorter confidence interval. This is in line with the idea that the IV-MW estimator is a more efficient approximation to the IV-PSM process.

5 Conclusion

A key assumption in instrumental variable analyses is that the instrument is randomly assigned, which requires that there are no unmeasured confounders between the instrument and the outcome. Unfortunately, unless the instrument is based on actual randomization, this assumption is unlikely to hold without conditioning on a set of known, measured confounders. The researcher must then argue that the instrument is distributed as good as random after controlling for these instrument-outcome confounders. Garabedian et al. (2014) emphasize that the most commonly used instruments have potential instrument-outcome confounders associated with them and that failing to adjust for these confounders can bias estimation.

In this article we developed a weighted IV estimator based on the IV propensity score to adjust for instrument-outcome confounders. The weights reflected the probability that an individual would be selected into a one-to-one IV propensity score match, and modified weights for approximating k:1 matching designs were provided as well. We further presented a double robust version of the estimator that protects against misspecification of the IV propensity score model. Though this required the additional specification of an outcome model, the double robust estimator provided consistent estimates if only one of the outcome or IV propensity score models was correctly specified, but did not require both to be correct.

One-to-one IV propensity score matching involves pairing each encouraged subject to an unencouraged subject with a similar IV propensity scores, often within a specified range. If a match cannot be found within this range, pairing the encouraged subject with an unencouraged subject with a substantially different IV propensity score can bias estimation, whereas removing that subject from the analysis leads a loss of efficiency. The proposed estimator avoids these pitfalls, leading to more efficient estimation as every individual contributes. Additional benefits over matching include straightforward variance estimation and computational efficiency. Through simulation, the proposed estimator was found to outperform

alternatives, being equally unbiased with uniformly smaller mean squared errors.

Methods were illustrated using USRDS data to study the association between dialysis session length and first year mortality among hemodialysis patients in the United States. While longer dialysis sessions are thought to decrease risk of mortality, it is a difficult research question as the relationship between session length and mortality is likely confounded, as smaller patients with higher mortality risk are more likely to be prescribed shorter dialysis sessions. This might explain the lack of consensus among previous studies. Using the IV methods of this article, a small protective effect of longer dialysis sessions was found, suggesting 1.5 fewer first year deaths for every 100 dialysis patients encouraged to switch from shorter to longer dialysis sessions.

Appendix

In this section we show that the IV-MW and IV-PSM estimators have the same limit as $n \to \infty$. Following Li and Greene (2013), we will assume that the IV propensity score takes finitely many values c_k for $k = 1, ..., K$ with $c_k \in (0, 1)$. This assumption is to allow exact matching on the IV propensity score and avoid unnecessary complications of working with other matching algorithms. For the IV-PSM estimator we assume one-to-one exact matching without replacement on the IV propensity score. Additionally, we simplify the notation of section 2, letting $Y(1, D_i) = Y_i^1$, $Y_i(0, D_i) = Y_i^0$, $D_i(1) = D_i^1$, $D_i(0) = D_i^0$, $Y_i = Z_i Y_i^1 + (1 - Z_i)Y_i^0$, $D_i = Z_i D_i^1 + (1 - Z_i)D_i^0$, and $e_i(\mathbf{x}_i) = e_i$. We further denote $P(e_i = c_k) = \tau_k$, with $\sum_k \tau_k = 1$.

We begin with the IV-MW estimator, defined as

$$\begin{aligned} \lambda_{\text{IV-MW}} &= \frac{\sum_i W_i Z_i Y_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) Y_i / \sum_i W_i (1 - Z_i)}{\sum_i W_i Z_i D_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) D_i / \sum_i W_i (1 - Z_i)} \\ &\equiv \frac{A/F - B/G}{C/F - D/G}. \end{aligned}$$

The limit for A is

$$\begin{aligned} n^{-1} \sum_i W_i Z_i Y_i &\to_p E\{W_i Z_i Y_i^1\} \\ &= E\left\{ E(\frac{min(e_i, 1 - e_i)}{e_i} I(Z_i = 1) Y_i^1 | \mathbf{x}_i) \right\} \\ &= E\{min(e_i, 1 - e_i) E(Y_i^1 | \mathbf{x}_i)\}. \end{aligned}$$

Similarly, the limits for B, C, and D are given by

$$n^{-1}\sum_i W_i(1-Z_i)Y_i \to_p E\{min(e_i, 1-e_i)E(Y_i^0|\mathbf{x}_i)\},$$
$$n^{-1}\sum_i W_i Z_i D_i \to_p E\{min(e_i, 1-e_i)E(D_i^1|\mathbf{x}_i)\},$$
$$n^{-1}\sum_i W_i(1-Z_i)D_i \to_p E\{min(e_i, 1-e_i)E(D_i^0|\mathbf{x}_i)\}.$$

Taking the limit of F and G gives

$$\begin{aligned} n^{-1}\sum_i W_i Z_i &\to_p E\{W_i Z_i\} \\ &= E\left\{\frac{min(e_i, 1-e_i)}{e_i} I(Z_i = 1)\right\} \\ &= E\{min(e_i, 1-e_i)\} \end{aligned}$$

and

$$n^{-1}\sum_i W_i(1-Z_i) \to_p E\{min(e_i, 1-e_i)\}$$

Combining these and reducing, the limit of the IV-MW as $n \to \infty$ is given as

$$\hat{\lambda}_{\text{IV-MW}} \to_p \frac{E\{min(e_i, 1-e_i)(E(Y_i^1|\mathbf{x}_i) - E(Y_i^0|\mathbf{x}_i))\}}{E\{min(e_i, 1-e_i)(E(D_i^1|\mathbf{x}_i) - E(D_i^0|\mathbf{x}_i))\}}.$$

Next we consider the IV-PSM estimator, which we write as

$$\hat{\lambda}_{\text{IV-PSM}} = \frac{\left\{\frac{\sum_k \sum_i Y_i I(i\in S_{1k})}{\sum_k \sum_i I(i\in S_{1k})}\right\} - \left\{\frac{\sum_k \sum_i Y_i I(i\in S_{0k})}{\sum_k \sum_i I(i\in S_{0k})}\right\}}{\left\{\frac{\sum_k \sum_i D_i I(i\in S_{1k})}{\sum_k \sum_i I(i\in S_{1k})}\right\} - \left\{\frac{\sum_k \sum_i D_i I(i\in S_{0k})}{\sum_k \sum_i I(i\in S_{0k})}\right\}} \equiv \frac{A/F - B/G}{C/F - D/G},$$

where S_{1k} and S_{0k} represent the sets of encouraged and unencouraged subjects matched at

c_k, respectively. The limit of A is then

$$\begin{aligned} n^{-1}\sum_k\sum_i Y_i I(i\in S_{1k}) &= n^{-1}\sum_k\sum_i Y_i^1 I(i\in S_{1k}) \\ &\to_p E\left\{\sum_k Y_i^1 I(i\in S_{1k})\right\} \\ &= E\left\{E(Y_i^1|\mathbf{x}_i)E(\sum_k I(i\in S_{1k})|\mathbf{x}_i)\right\} \\ &= E\left\{E(Y_i^1|\mathbf{x}_i)\sum_k \tau_k e_i\frac{min(e_i,1-e_i)}{e_i}\right\} \\ &= E\{min(e_i,1-e_i)E(Y_i^1|\mathbf{x}_i)\}. \end{aligned}$$

Similarly, the limits for B, C, and D are given as

$$\begin{aligned} n^{-1}\sum_k\sum_i Y_i I(i\in S_{0k}) &\to_p E\{min(e_i,1-e_i)E(Y_i^0|\mathbf{x}_i)\}, \\ n^{-1}\sum_k\sum_i D_i I(i\in S_{1k}) &\to_p E\{min(e_i,1-e_i)E(D_i^1|\mathbf{x}_i)\}, \\ n^{-1}\sum_k\sum_i D_i I(i\in S_{0k}) &\to_p E\{min(e_i,1-e_i)E(D_i^0|\mathbf{x}_i)\}. \end{aligned}$$

Finally, for F we have

$$\begin{aligned} n^{-1}\sum_k\sum_i I(i\in S_{1k}) &\to_p E\left\{\sum_k I(i\in S_{1k})\right\} \\ &= E\left\{min(e_i,1-e_i)\sum_k \tau_k\right\} \\ &= E\{min(e_i,1-e_i)\}, \end{aligned}$$

and similarly for G

$$n^{-1}\sum_k\sum_i I(i\in S_{0k}) \to_p E\{min(e_i,1-e_i)\}.$$

Combining everything and reducing, the limit of the IV-PSM estimator as $n \to \infty$ is found to be

$$\hat{\lambda}_{\text{IV-PSM}} \to_p = \frac{E\{min(e_i, 1-e_i)(E(Y_i^1|\mathbf{x}_i) - E(Y_i^0|\mathbf{x}_i))\}}{E\{min(e_i, 1-e_i)(E(D_i^1|\mathbf{x}_i) - E(D_i^0|\mathbf{x}_i))\}},$$

which is the same as that of the IV-MW estimator.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 444–455.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* **27,** 2037–2049.

Austin, P. C. (2009). Type i error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics* **5,**.

Austin, P. C. (2011a). Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in Medicine* **30,** 1292–1301.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* **10,** 150–161.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine* **33,** 2297–2340.

Bhattacharya, J., Goldman, D., and McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in medicine* **25,** 389–413.

Brookhart, M. A. and Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The International Journal of Biostatistics* **3,**.

Brunelli, S. M., Chertow, G. M., Ankers, E. D., Lowrie, E. G., and Thadhani, R. (2010).

Shorter dialysis times are associated with higher mortality among incident hemodialysis patients. *Kidney international* **77,** 630–636.

Cheng, J. and Lin, W. (2013). Understanding causal effects in observational studies with instrumental propensity scores. *Joint Statistical Meeting* .

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *SankhyÄ: The Indian Journal of Statistics, Series A (1961-2002)* **35,** 417–446.

Dartmouth (2016). *The Dartmouth Atlas of Healthcare*. Trustees of Dartmouth College.

Daugirdas, J. T. (2013). Dialysis time, survival, and dose-targeting bias. *Kidney Int* **83,** 9–13.

Eknoyan, G., Beck, G. J., Cheung, A. K., Daugirdas, J. T., Greene, T., Kusek, J. W., Allon, M., Bailey, J., Delmez, J. A., Depner, T. A., et al. (2002). Effect of dialysis dose and membrane flux in maintenance hemodialysis. *New England Journal of Medicine* **347,** 2010–2019.

Flythe, J. E., Curhan, G. C., and Brunelli, S. M. (2013). Shorter length dialysis sessions are associated with increased mortality, independent of body weight. *Kidney international* **83,** 104–113.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics* **139,** 35–75.

Garabedian, L. F., Chu, P., Toh, S., Zaslavsky, A. M., and Soumerai, S. B. (2014). Potential bias of instrumental variable analyses for observational comparative effectiveness research. *Annals of Internal Medicine* **161,** 131–138.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60,** 505–531.

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

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2014). Balancing covariates via propensity score weighting. *arXiv:1404.1785 [stat.ME]* .

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics* **9,** 215–234.

Li, Y., Lee, Y., Wolfe, R. A., Morgenstern, H., Zhang, J., Port, F. K., and Robinson, B. M. (2015). On a preference-based instrumental variable approach in reducing unmeasured confounding-by-indication. *Statistics in Medicine* **34,** 1150–1168.

Love, T. (2002). Displaying covariate balance after adjustment for selection bias. In *Joint Statistical Meetings*.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* **23,** 2937–2960.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science* **5,** 463–480.

Raynor, W. J. (1983). Caliper pair-matching on a continuous variable in case-control studies. *Communications in Statistics - Theory and Methods* **12,** 1499–1509.

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39,** 33–38.

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

Saran, R., Bragg-Gresham, J., Levin, N., Twardowski, Z., Wizemann, V., Saito, A., Kimata, N., Gillespie, B., Combe, C., Bommer, J., et al. (2006). Longer treatment time and slower ultrafiltration in hemodialysis: associations with reduced mortality in the dopps. *Kidney international* **69,** 1222–1228.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* **101,** 1607–1618.

Terza, J., Basu, A., and Rathouz, P. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of health economics* **27,** 531–543.