DISCUSSION PAPER SERIES

IZA DP No. 8084

A General Double Robustness Result for Estimating Average Treatment Effects

Tymon Słoczyński
Jeffrey M. Wooldridge

March 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

A General Double Robustness Result for Estimating Average Treatment Effects

Tymon Słoczyński
Michigan State University,
Warsaw School of Economics and IZA

Jeffrey M. Wooldridge
Michigan State University
and IZA

Discussion Paper No. 8084
March 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 8084
March 2014

ABSTRACT

A General Double Robustness Result for Estimating Average Treatment Effects[*]

In this paper we study doubly robust estimators of various average treatment effects under unconfoundedness. We unify and extend much of the recent literature by providing a very general identification result which covers binary and multi-valued treatments; unnormalized and normalized weighting; and both inverse-probability weighted (IPW) and doubly robust estimators. We also allow for subpopulation-specific average treatment effects where subpopulations can be based on covariate values in an arbitrary way. Similar to Wooldridge (2007), we then discuss estimation of the conditional mean using quasi-log likelihoods (QLL) from the linear exponential family.

JEL Classification: C13, C21, C31, C51

Keywords: double robustness, inverse-probability weighting (IPW), multi-valued treatments, quasi-maximum likelihood estimation (QMLE), treatment effects

Corresponding author:

Jeffrey M. Wooldridge
Department of Economics
Michigan State University
East Lansing, MI 48824
USA
E-mail: wooldri1@msu.edu

[*] Tymon Słoczyński gratefully acknowledges a START scholarship from the Foundation for Polish Science (FNP).

1 Introduction

In causal inference settings, doubly robust estimators involve models for both the propensity score and the conditional mean of the outcome, and remain consistent if one of these models (but not both) is misspecified. Augmented inverse-probability weighting (AIPW), the standard doubly robust estimator, was introduced in the missing data literature by Robins et al. (1994). Its robustness to misspecification was demonstrated in later work by Scharfstein et al. (1999), and the term "doubly robust" (or "doubly protected") was introduced by Robins et al. (2000). This class of estimators continues to be an important topic of research in statistics, both in causal inference and in missing data settings, with recent contributions by Bang and Robins (2005), Tan (2006), Kang and Schafer (2007), Cao et al. (2009), Tan (2010), Rotnitzky et al. (2012), and others.

In recent years, there has also been substantive interest in doubly robust estimators in the econometric literature. Wooldridge (2007) has developed a general framework for missing data problems and studied doubly robust estimators of the average treatment effect (ATE), including inverse-probability weighted QML estimators with logistic and exponential mean functions. Kaiser (2013) has extended this contribution to decomposition problems and estimating the average treatment effect on the treated (ATT). Cattaneo (2010), Uysal (2012), and Farrell (2013) have considered multi-valued treatment effects, with Uysal (2012) studying parametric doubly robust estimators and Cattaneo (2010) and Farrell (2013) developing (efficient) semiparametric methods.[1] Other recent papers include Kline (2011), Graham et al. (2012), and Rothe and Firpo (2013).

[1] More generally, Farrell (2013) has considered post model selection inference on various average treatment effects of interest when the number of covariates can exceed the number of observations. In this context double robustness allows for accurate coverage even if the outcome model or the propensity score model (but not both) is not sparse.

In this paper, we unify and extend some of this recent literature on doubly robust estimators by providing a very general identification result which accounts for the majority of interesting problems. We cover both binary and multi-valued treatments; the average treatment effect, the average treatment effect on the treated, and average treatment effects for other subpopulations of interest; unnormalized and normalized weighting; and linear, logistic, and exponential mean functions. Inverse-probability weighting (IPW) is also easily shown to be a special case within our approach. As far as we know, this is the first paper to consider all these problems jointly and provide such a general identification result. Moreover, unlike in the majority of these recent studies, our parameters of interest are defined as a solution to a population optimization problem, and not to a moment condition. Our approach also carefully explains the anatomy of double robustness in a very general setting.

The remainder of the paper is organized as follows. In Section 2, we introduce notation as well as main assumptions and estimands. In Section 3, we present our identification results and discuss several special cases within this approach. In Section 4, we discuss estimation. Finally, we summarize our main findings in Section 5.

2 Parameters of Interest and Assumptions

We assume some treatment to take on $G+1$ different values, labeled $\{0, 1, 2, \ldots, G\}$. For a given population, let W represent the treatment assignment. Typically, $W = 0$ represents the absence of treatment, but this is not important for what follows. The leading case is $G = 1$, and then $W = 0$ denotes control and $W = 1$ denotes treatment.

For each level of treatment, g, we assume counterfactual outcomes, Y_g, $g \in \{0, 1, 2, \ldots, G\}$. Most of the common treatment effects are defined in terms of the mean values of the Y_g. For example, let

$$\mu_g = \mathbb{E}(Y_g),\ g = 0, 1, 2, \ldots, G \tag{1}$$

denote the mean values of the counterfactual outcomes across the entire population. Assuming $g = 0$ to be the control, the average treatment effect of treatment level g is

$$\tau_{g,ate} = \mathbb{E}(Y_g - Y_0) = \mu_g - \mu_0. \tag{2}$$

We may also be interested in the average treatment effect for units actually receiving this level of treatment, namely

$$\tau_{g,att} = \mathbb{E}(Y_g - Y_0|W = g) = \mathbb{E}(Y_g|W = g) - \mathbb{E}(Y_0|W = g). \tag{3}$$

With more than two treatment levels, we can define similar quantities comparing any two of them. The important point is that our goal is to estimate

$$\mathbb{E}(Y_g) \ \text{ or } \ \mathbb{E}(Y_g|W = h) \tag{4}$$

for treatment levels g and h.

Let X denote a vector of observed, pre-treatment covariates that predict treatment and have explanatory power for the Y_g. We assume that treatment is unconfounded conditional on X. We will refine this assumption when we state the general results; the most restrictive form we use is that treatment is unconfounded with respect to each counterfactual outcome:

$$W \perp Y_g \mid X \ ,\ g = 0, 1, 2, \ldots, G, \tag{5}$$

where "$\perp$" means "independent of" and "$|$" denotes "conditional on". If $\mathbb{D}(\cdot|\cdot)$ denotes conditional distribution, we can write unconfoundedness as $\mathbb{D}(W|Y_g, X) = \mathbb{D}(W|X)$. In estimating the parameter $\tau_{g,att}$, we will see that we only need to assume unconfoundedness with respect to Y_0, the counterfactual in the control state.

In what follows it is helpful to define binary treatment indicators as

$$W_g = 1[W = g],\ g = 0, 1, 2, \ldots, G \tag{6}$$

as well as the generalized propensity score (Imbens, 2000) for treatment level g as

$$p_g(x) = \mathbb{P}(W_g = 1|X = x). \tag{7}$$

By unconfoundeness of treatment,

$$p_g(X) = \mathbb{P}(W_g = 1|Y_g, X). \tag{8}$$

3 Identification Results

Let $q(Y_g, X)$ be any function of the counterfactual response, Y_g, and the covariates, X; we assume $q(Y_g, X)$ to have a finite absolute first moment. Also, let D be a binary variable which – like the treatment W – is unconfounded with respect to each Y_g, conditional on X. In other words, D is independent of Y_g, conditional on X. In applications, D might be a deterministic function of X, in which case its inclusion serves to isolate a subset of the population. Another important case is when D is an indicator for a different level of treatment.

Let $\eta = \mathbb{P}(D = 1)$ be the unconditional probability that $D = 1$ and assume that $\eta > 0$. The special case of $\mathbb{P}(D = 1) = 1$ is important and is allowed. Also, define the propensity score for D as

$$r(x) = \mathbb{P}(D = 1|X = x); \tag{9}$$

by unconfoundedness,

$$r(X) = \mathbb{P}(D = 1|Y_g, X). \tag{10}$$

3.1 A General Result on Weighting

The following lemma is crucial for one half of our general double robustness result. The final step in the proof of the lemma uses the simple identity

$$\mathbb{E}(D \cdot Z) = \mathbb{P}(D = 1) \cdot \mathbb{E}(Z|D = 1) \tag{11}$$

for D a binary variable and Z any random variable with $\mathbb{E}(|Z|) < \infty$.

Lemma 1: Assume that W_g and D are each unconfounded with respect to Y_g, conditional on X. Define $\eta = \mathbb{P}(D = 1)$ and assume $\eta > 0$. Further, $p_g(x) > 0$ for all $x \in \mathcal{X}$, where $p_g(x)$ is defined in (7). Then,

$$\frac{1}{\eta} \cdot \mathbb{E}\left[\frac{W_g}{p_g(X)} r(X) q(Y_g, X)\right] = \mathbb{E}\left[q(Y_g, X)|D = 1\right]. \tag{12}$$

Proof: The proof that

$$\mathbb{E}\left[\frac{W_g}{p_g(X)} r(X) q(Y_g, X)\right] = \mathbb{E}\left[r(X) q(Y_g, X)\right] \tag{13}$$

is similar to Wooldridge (2007), and is an implication of iterated expectations and unconfoundedness

$$\begin{aligned}\mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)q(Y_g,X)\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)q(Y_g,X)\middle| Y_g,X\right]\right\}\\ &= \mathbb{E}\left\{\left[\frac{\mathbb{E}(W_g|Y_g,X)}{p_g(X)}r(X)q(Y_g,X)\right]\right\}\\ &= \mathbb{E}\left\{\left[\frac{\mathbb{E}(W_g|X)}{p_g(X)}r(X)q(Y_g,X)\right]\right\}\\ &= \mathbb{E}\left[r(X)q(Y_g,X)\right], \end{aligned} \tag{14}$$

because $\mathbb{E}(W_g|X) = p_g(X)$. Next, we show that

$$\mathbb{E}\left[r(X)q(Y_g,X)\right] = \mathbb{E}\left[D \cdot q(Y_g,X)\right] \tag{15}$$

which again follows by iterated expectations and unconfoundedness of D:

$$\begin{aligned}\mathbb{E}\left[D \cdot q(Y_g,X)\right] &= \mathbb{E}\left\{\mathbb{E}\left[D \cdot q(Y_g,X)\right| Y_g,X]\right\}\\ &= \mathbb{E}\left\{\left[\mathbb{E}(D|Y_g,X)q(Y_g,X)\right]\right\}\\ &= \mathbb{E}\left\{\left[\mathbb{E}(D|X)q(Y_g,X)\right]\right\}\\ &= \mathbb{E}\left[r(X)q(Y_g,X)\right]. \end{aligned} \tag{16}$$

Finally,

$$\begin{aligned}\mathbb{E}\left[D \cdot q(Y_g,X)\right] &= (1-\eta)\cdot\mathbb{E}\left[D \cdot q(Y_g,X)|D=0\right] + \eta\cdot\mathbb{E}\left[D \cdot q(Y_g,X)|D=1\right]\\ &= \eta\cdot\mathbb{E}\left[q(Y_g,X)|D=1\right]. \end{aligned} \tag{17}$$

Combining the three pieces gives

$$\mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)q(Y_g,X)\right]=\eta\cdot\mathbb{E}\left[q(Y_g,X)|D=1\right], \tag{18}$$

which completes the proof, because $\eta>0$ is assumed. □

3.2 Unnormalized versus Normalized Weighting

In the previous setup, given a random sample $\{(W_{ig},D_i,X_i,Y_i):i=1,2,\ldots,N\}$, Lemma 1 suggests how to consistently estimate $\mu_{g,1}\equiv\mathbb{E}\left[q(Y_g,X)|D=1\right]$:

$$\frac{1}{\hat{\eta}}\left[N^{-1}\sum_{i=1}^{N}\frac{W_{ig}}{p_g(X_i)}r(X_i)q(Y_i,X_i)\right], \tag{19}$$

where $\hat{\eta}\xrightarrow{p}\eta>0$. One simple, unbiased and consistent estimator of η is

$$\hat{\eta}=N^{-1}\sum_{i=1}^{N}D_i=N_D/N, \tag{20}$$

where N_D is the number of observations with $D_i=1$. The estimator of $\mu_{g,1}$ is then

$$\hat{\mu}_{g,1,unnormalized}=N_D^{-1}\sum_{i=1}^{N}\frac{W_{ig}}{p_g(X_i)}r(X_i)q(Y_i,X_i). \tag{21}$$

In special cases, several papers have discouraged empirical researchers from using $\hat{\eta}=N_D/N$, because it leads to a weighted average where the weights do not sum to unity. In particular, the weight for observation i is

$$\frac{1}{N_D}\frac{W_{ig}}{p_g(X_i)}r(X_i), \tag{22}$$

and these do not usually sum to unity across i. It is a simple adjustment to obtain a consistent estimator whose weights are guaranteed to sum to unity. To choose such weights, note that we can apply Lemma 1 to $q(Y_g, X) \equiv 1$ to get

$$\eta = \mathbb{E}\left[\frac{W_g}{p_g(X)} r(X)\right], \tag{23}$$

and so an alternative unbiased and consistent estimator of η is

$$\hat{\eta} = N^{-1} \sum_{i=1}^{N} \frac{W_{ig}}{p_g(X_i)} r(X_i). \tag{24}$$

When we plug this estimator in (19) for $\hat{\eta}$, we obtain

$$\hat{\mu}_{g,1,normalized} = \left[\sum_{i=1}^{N} \frac{W_{ig}}{p_g(X_i)} r(X_i)\right]^{-1} \sum_{i=1}^{N} \frac{W_{ig}}{p_g(X_i)} r(X_i) q(Y_i, X_i) \tag{25}$$

and now the weights,

$$\left[\sum_{j=1}^{N} \frac{W_{jg}}{p_g(X_j)} r(X_j)\right]^{-1} \frac{W_{ig}}{p_g(X_i)} r(X_i), \tag{26}$$

necessarily sum to unity across i.

Many applications of inverse-probability weighted (IPW) estimators, including those to doubly robust estimation, use normalized weights because the weights are applied to an objective function, such as a squared residual or a quasi-log likelihood function. For example, to estimate $\mu_g = \mathbb{E}(Y_g)$, we can solve

$$\min_{m_g \in \mathbb{R}} \sum_{i=1}^{N} \frac{W_{ig}}{p_g(X_i)} (Y_i - m_g)^2, \tag{27}$$

and the solution is easily seen to be the estimator with normalized weights.

3.3 Special Cases

Before considering doubly robust estimation, it is useful to see how some important special cases in the literature fit into the current framework. We are primarily interested in showing the population moments that establish identification, but the formulas also suggest simple estimators – using either unnormalized or normalized weights.

Binary treatments: Let $G = 1$, $W_0 = 1 - W_1 = 1 - W$, and $p_0(X) = 1 - p_1(X) = 1 - p(X)$. Then, with $q(Y_g, X) = Y_g$ and $Y = (1 - W) \cdot Y_0 + W \cdot Y_1$, Lemma 1 implies

$$\tau_{ate} = \mathbb{E}(Y_1 - Y_0) = \mathbb{E}\left[\frac{W \cdot Y}{p(X)} - \frac{(1 - W) \cdot Y}{1 - p(X)}\right], \tag{28}$$

with $D = 1$ in both cases. This expression leads directly to the Horvitz and Thompson (1952) estimator. An expression that leads to normalized weights is

$$\tau_{ate} = \frac{\mathbb{E}\left[\frac{W \cdot Y}{p(X)}\right]}{\mathbb{E}\left[\frac{W}{p(X)}\right]} - \frac{\mathbb{E}\left[\frac{(1-W) \cdot Y}{1-p(X)}\right]}{\mathbb{E}\left[\frac{1-W}{1-p(X)}\right]}, \tag{29}$$

where the denominators in both expressions are equal to unity but their sample counterparts will not be. Similarly, we can use Lemma 1 to write the average treatment effect on the treated as

$$\tau_{att} = \mathbb{E}(Y_1 - Y_0 | W = 1) = \frac{\mathbb{E}(W \cdot Y)}{\mathbb{P}(W = 1)} - \frac{\mathbb{E}\left[\frac{1-W}{1-p(X)} p(X) \cdot Y\right]}{\mathbb{P}(W = 1)}, \tag{30}$$

because $D = W$, $\eta = \mathbb{P}(W = 1)$, and $r(X) = p(X)$. Instead of dividing by $\mathbb{P}(W = 1)$, we can divide the second expectation by

$$\mathbb{E}\left[\frac{(1 - W) \cdot p(X)}{1 - p(X)}\right] \tag{31}$$

to obtain an estimator with normalized weights. In particular, the estimate of τ_{att} is obtained from the simple regression,

$$Y_i \text{ on } 1, W_i \ (i = 1, 2, \ldots, N), \tag{32}$$

using weights

$$W_i + (1 - W_i)\frac{p(X_i)}{1 - p(X_i)}\frac{1 - \eta}{\eta}, \tag{33}$$

an estimator suggested by Busso et al. (2009). (In practice, we do not know the propensity score and we would replace it with a consistent estimator.)

More generally, we can write the average treatment effect for any subpopulation of interest as

$$\mathbb{E}(Y_1 - Y_0 | D = 1) = \frac{1}{\eta} \cdot \mathbb{E}\left[\frac{W}{p(X)}r(X) \cdot Y - \frac{1 - W}{1 - p(X)}r(X) \cdot Y\right], \tag{34}$$

as long as this subpopulation is defined by D, a binary variable which is unconfounded with respect to potential outcomes, conditional on X. A leading case is when D is a deterministic function of X, so we are looking at a subpopulation determined by the conditioning variables that appear in the propensity score.

Dose-response function: Let $D = 1$ and define the dose-response function as $\mu = (\mu_0, \mu_1, \ldots, \mu_G)$. See also Imbens (2000). Then, we can use Lemma 1, along with

$$Y = W_0 \cdot Y_0 + W_1 \cdot Y_1 + \ldots + W_G \cdot Y_G, \tag{35}$$

to write the dose-response function as

$$\mu = \left(\mathbb{E}\left[\frac{W_0}{p_0(X)} Y\right], \mathbb{E}\left[\frac{W_1}{p_1(X)} Y\right], \ldots, \mathbb{E}\left[\frac{W_G}{p_G(X)} Y\right] \right). \tag{36}$$

In estimating the mean μ_g, we can write an expression that leads directly to normalized weights, namely

$$\mu_g = \frac{\mathbb{E}\left[\frac{W_g}{p_g(X)} Y\right]}{\mathbb{E}\left[\frac{W_g}{p_g(X)}\right]}, \tag{37}$$

where the denominator is unity for all g.

Average effects of multi-valued treatments: The expression in (37) suggests that the average gain in going from the control, $g = 0$, to treatment level g is:

$$\tau_{g,ate} = \mathbb{E}(Y_g - Y_0) = \frac{\mathbb{E}\left[\frac{W_g}{p_g(X)} Y\right]}{\mathbb{E}\left[\frac{W_g}{p_g(X)}\right]} - \frac{\mathbb{E}\left[\frac{W_0}{p_0(X)} Y\right]}{\mathbb{E}\left[\frac{W_0}{p_0(X)}\right]}. \tag{38}$$

Similarly, the average treatment effect on those receiving treatment level g, relative to no treatment, is:

$$\tau_{g,att} = \mathbb{E}(Y_g - Y_0 | W = g) = \frac{\mathbb{E}(W_g \cdot Y)}{\mathbb{P}(W = g)} - \frac{\mathbb{E}\left[\frac{W_0}{p_0(X)} p_g(X) \cdot Y\right]}{\mathbb{E}\left[\frac{W_0}{p_0(X)} p_g(X)\right]}. \tag{39}$$

4 Doubly Robust Estimators

We now develop doubly robust (DR) estimators of various average treatment effects by considering estimation of

$$\mu_{g,1} \equiv E(Y_g|D=1). \tag{40}$$

As we saw in Section 3, various average treatment effects can be obtained by appropriate choice of D, where $D=1$ simply defines a subpopulation of interest.

It is helpful to divide the argument into two subsections. The first part of the DR result is when a conditional mean function is correctly specified, and here we need to draw on important results from the literature on quasi-MLE estimation of correctly specified conditional means. The second part requires an application of Lemma 1 and a basic understanding of the linear exponential family of distributions.

The setting is that for a counterfactual outcome Y_g a parametric mean function is specified, which we write as $\{m_g(x,\theta_g) : x \in \mathcal{X}, \theta_g \in \Theta_g\}$. Along with the specification of the mean function, we choose as an objective function a quasi-log likelihood (QLL) from the linear exponential family (LEF). As discussed in Gourieroux et al. (1984) – see also Wooldridge (2010, Chapter 13) – the LEF has the feature that it identifies the parameters in a correctly specified conditional mean. What is somewhat less known is that if the QLL is chosen so that the conditional mean function represents the so-called canonical link, then the unconditional mean is consistently estimated even if the conditional mean function is misspecified. We use this fact in Section 4.2.

In what follows we assume regularity conditions such as smoothness of the conditional mean functions in β_g and enough finite moments so that standard consistency and asymptotic normality results hold for quasi-maximum likelihood estimation.

4.1 Part 1: The Conditional Mean Is Correctly Specified

In this subsection we assume that the conditional mean is correctly specified which means that, for some vector $\theta_g^o \in \Theta_g$,

$$\mathbb{E}(Y_g|X = x) = m_g(x, \theta_g^o),\ x \in \mathcal{X}, \tag{41}$$

where $\mathcal{X}$ is the support of X. As shown in Gourieroux et al. (1984), if $q(Y_g, X; \theta_g)$ is a QLL from a density in the LEF with mean function $m_g(x, \theta_g)$, then θ_g^o is a solution to

$$\max_{\theta_g \in \Theta_g}\ \mathbb{E}[q(Y_g, X; \theta_g)|X] \tag{42}$$

for all outcomes X, which means

$$\mathbb{E}[q(Y_g, X; \theta_g^o)|X] \geq \mathbb{E}[q(Y_g, X; \theta_g)|X]. \tag{43}$$

We use parametric models for the propensity scores, $p_g(x)$, say $F_g(x; \gamma_g)$. We allow this model to be misspecified, but assume that the estimator settles down to a limit: $\hat{\gamma}_g \xrightarrow{p} \gamma_g^*$ where γ_g^* is sometimes called the "pseudo-true value". Similarly, $\mathbb{P}(D = 1|X = x)$ is modeled parametrically as $J(x; \psi)$ with $\hat{\psi} \xrightarrow{p} \psi^*$. In obtaining $\hat{\gamma}_g$ and $\hat{\psi}$ we would almost certainly use the Bernoulli log likelihood. In other words, we estimate stanard binary response models by MLE. (More precisely, by quasi-MLE because we allow the binary response models to be misspecified.)

Then the weighted objective function for estimating θ_g^o is

$$N^{-1} \sum_{i=1}^{N} \frac{W_{ig}}{F_g(X_i; \hat{\gamma}_g)} J(X_i; \hat{\psi}) \cdot q(Y_i, X_i; \theta_g). \tag{44}$$

Using standard convergence results – for example, Newey and McFadden (1994) and

Wooldridge (2010, Chapter 12), (44) converges in probability to

$$\begin{aligned}\mathbb{E}\left[\frac{W_g}{F_g(X;\gamma_g^*)}J(X;\psi^*)\cdot q(Y_g,X;\theta_g)\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{W_g}{F_g(X;\gamma_g^*)}J(X;\psi^*)\cdot q(Y_g,X;\theta_g)\middle| X\right]\right\} \\ &= \mathbb{E}\left\{\frac{\mathbb{E}(W_g|X)}{F_g(X;\gamma_g^*)}J(X;\psi^*)\cdot \mathbb{E}[q(Y_g,X;\theta_g)|X]\right\} \\ &= \mathbb{E}\left\{\frac{p_g(X)J(X;\psi^*)}{F_g(X;\gamma_g^*)}\mathbb{E}[q(Y_g,X;\theta_g)|X]\right\}. \quad (45)\end{aligned}$$

But $p_g(X)J(X;\psi^*)/F_g(X;\gamma_g^*) \geq 0$ so

$$\frac{p_g(X)J(X;\psi^*)}{F_g(X;\gamma_g^*)}\mathbb{E}[q(Y_g,X;\theta_g^o)|X] \geq \frac{p_g(X)J(X;\psi^*)}{F_g(X;\gamma_g^*)}\mathbb{E}[q(Y_g,X;\theta_g)|X] \quad (46)$$

for all X. By iterated expectations, θ_g^o is a solution to

$$\max_{\theta_g\in\Theta_g}\mathbb{E}\left[\frac{W_g}{F_g(X;\gamma_g^*)}J(X;\psi^*)\cdot q(Y_g,X;\theta_g)\right] \quad (47)$$

and, provided the mean function is well specified and the distribution of X is sufficiently rich, θ_g^o will be the unique solution. The conclusion is that, even if $\mathbb{P}(W_g = 1|X)$ and $\mathbb{P}(D = 1|X)$ are misspecified, we consistently estimate the parameters θ_g^o in the correctly specified conditional mean,

$$\mathbb{E}(Y_g|X) = m_g(X,\theta_g^o). \quad (48)$$

Because D is unconfounded conditional on X,

$$\mathbb{E}(Y_g|X,D) = \mathbb{E}(Y_g|X) \quad (49)$$

and so

$$\mathbb{E}(Y_g|D=1) = \mathbb{E}[m_g(X,\theta_g^o)|D=1]. \tag{50}$$

It follows that a consistent estimator of $\mu_{g,1} = \mathbb{E}(Y_g|D=1)$ is

$$\hat{\mu}_{g,1} = N_D^{-1}\sum_{i=1}^{N} D_i \cdot m_g(X_i,\hat{\theta}_g), \tag{51}$$

where N_D is the number of observations with $D_i = 1$.

4.2 Part 2: The Propensity Score Is Correctly Specified

We are still interested in consistently estimating $\mu_{g,1} = \mathbb{E}(Y_g|D=1)$. Now we assume that we have correctly specified parametric models for the propensity scores and $\mathbb{P}(D=1|X=x)$:

$$\mathbb{P}(W_g=1|X=x) = F(x,\gamma_g^o) \tag{52}$$

$$\mathbb{P}(D=1|X=x) = J(x,\psi^o), \tag{53}$$

and we still maintain unconfoundedness with respect to Y_g. In some cases we will not estimate $\mathbb{P}(D=1|X=x)$. From Lemma 1 we know that because

$$\frac{1}{\eta}\cdot\mathbb{E}\left[\frac{W_g}{F(X,\gamma_g^o)}J(X,\psi^o)\cdot q(Y_g,X;\theta_g)\right] = \mathbb{E}\left[q(Y_g,X;\theta_g)|D=1\right] \tag{54}$$

for all θ_g, the minimizer θ_g^* of $\mathbb{E}\left[q(Y_g,X;\theta_g)|D=1\right]$, which we assume is unique, is also the minimizer of

$$\mathbb{E}\left[\frac{W_g}{F(X,\gamma_g^o)}J(X,\psi^o)\cdot q(Y_g,X;\theta_g)\right]. \tag{55}$$

By the convergence arguments in Section 4.1, the solution $\hat{\theta}_g$ to (44) is consistent for θ_g^*. So it remains to show that, for estimating $\mu_{g,1}$, having a consistent estimator of θ_g^* suffices.

In order to recover $\mu_{g,1}$ from $m_g(X, \theta_g^*)$, we need to know some further properties of the LEF family. As discussed in Wooldridge (2007), certain combinations of QLLs and mean functions generate the important result

$$\mathbb{E}(Y_g|D = 1) = \mathbb{E}[m_g(X, \theta_g^*)|D = 1]. \tag{56}$$

The key is that for a given LEF we choose the canonical link function to obtain the conditional mean model. For the normal distribution, which leads to OLS as the estimation method, the canonical link function leads to a mean linear in parameters. It is well-known from linear regression analysis that, as long as an intercept is included in the equation, the average of the fitted values is the same as the average of the dependent variable. The population result also holds. Thus, if we use a linear model $m_g(x, \theta_g) = \alpha_g + x\beta_g$, then it is always true that

$$\mathbb{E}(Y_g|D = 1) = \mathbb{E}(\alpha_g^* + X\beta_g^*|D = 1). \tag{57}$$

The same is true for the Bernoulli QLL when we use a logistic function for the mean:

$$m_g(x, \theta_g) = \Lambda(\alpha_g + x\beta_g), \tag{58}$$

which means that if Y_g is binary or fractional, then we should use the Bernoulli QMLE with a logistic mean function. A third useful case is when $Y_g \geq 0$, in which case the QLL-mean pair that delivers double robustness is the Poisson QLL and an exponential mean function: $m_g(x, \theta_g) = \exp(\alpha_g + x\beta_g)$. These cases are discussed in more detail in Wooldridge (2007). See also Kaiser (2013) for an application of the Poisson QMLE with an exponential mean function to decomposition problems. The new twist here is that the claims hold for any population we choose to define via $D = 1$, and because D can be a treatment indicator or an indicator based on X, we have a single double robustness result for a broad class of average treatment effects.

5 Summary

In this paper we unify the current literature on doubly robust estimators by establishing identification of a large class of average treatment effects under unconfoundedness. We cover binary and multi-valued treatments as well as the average treatment effect, the average treatment effect on the treated, and average treatment effects for other subpopulations of interest (based on covariates). We allow for both unnormalized and normalized weighting, and cover standard inverse-probability weighted (IPW) estimators as a special case.

Because doubly robust estimators involve models for both the conditional mean and the propensity score, and require that at least one of these models is correctly specified in order to remain consistent, we carefully describe each of these cases. Similar to Wooldridge (2007), we consider estimation of the propensity score using Bernoulli QMLE as well as estimation of the conditional mean using various QLLs from the linear exponential family. More precisely, we consider three cases: OLS with a linear mean function; Bernoulli QMLE with a logistic mean function; and Poisson QMLE with an exponential mean function. These nonlinear mean functions have typically been ignored in recent work, even though they might provide a useful alternative to a linear model for many outcome variables of interest.

References

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

Busso, M., DiNardo, J., and McCrary, J. (2009). Finite sample properties of semiparametric estimators of average treatment effects. Unpublished.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155:138–154.

Farrell, M. H. (2013). Robust inference on average treatment effects with possibly more covariates than observations. Unpublished.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52:681–700.

Graham, B. S., Campos de Xavier Pinto, C., and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79:1053–1079.

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

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

Kaiser, B. (2013). Decomposing differences in arithmetic means: A doubly-robust estimation approach. Unpublished.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:523–539.

Kline, P. (2011). Oaxaca-Blinder as a reweighting estimator. *American Economic Review: Papers & Proceedings*, 101:532–537.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4. North Holland.

Robins, J. M., Rotnitzky, A., and van der Laan, M. (2000). Comment. *Journal of the American Statistical Association*, 95:477–482.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Rothe, C. and Firpo, S. (2013). Semiparametric estimation and inference using doubly robust moment conditions. IZA Discussion Paper no. 7564.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99:439–456.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Rejoinder. *Journal of the American Statistical Association*, 94:1135–1146.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101:1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97:661–682.

Uysal, S. D. (2012). Doubly robust estimation of causal effects with multivalued treatments. *Journal of Applied Econometrics*, forthcoming.

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

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd edition.