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

Tymon Słoczyński[†] Jeffrey M. Wooldridge[‡]

Abstract

In this paper we study doubly robust estimators of various average and quantile treatment effects under unconfoundedness; we also consider an application to a setting with an instrumental variable. 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

[*] We thank Peter Phillips (Editor), Arthur Lewbel (Co-Editor), and three anonymous referees for helpful comments. Tymon Słoczyński also acknowledges financial support from the National Science Centre (grant DEC-2012/05/N/HS4/00395) and the Foundation for Polish Science (a START scholarship). This version: December 15, 2016.

[†] Brandeis University.

[‡] Michigan State University. Correspondence: Department of Economics, Michigan State University, East Lansing, MI 48824-1038, USA. Phone: 517-353-5972. Fax: 517-432-1068. E-mail: wooldri1@msu.edu.

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. In this paper, we unify and extend some of the 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; distribution, quantile, and inequality treatment effects; local average treatment effects in a setting with an instrumental variable; 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 recent studies, our parameters of interest are defined as a solution to a population optimization problem, and not to a moment condition. This has an important advantage when the response or outcome variable has restricted range, as estimators based on moment conditions can produce estimated means outside the range of logical values – unless one makes some rather ad hoc adjustments. Here we extend the approach of Wooldridge (2007), who studied weighted objective functions in a missing data setting, along several useful dimensions. 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 review the statistical and econometric literature on doubly robust estimators. In Section 3, we introduce our notation as well as assumptions and estimands; we also present our main identification result (in Section 3.2). In Section 4, we discuss several applications of this approach. In Section 5, we discuss estimation. Further, we summarize our main findings in Section 6. Finally, in Appendix B, we discuss implementation of our estimation methods.

2 Background

Augmented inverse-probability weighting (AIPW), a standard class of doubly robust estimators, was introduced in the missing data literature by Robins *et al.* (1994).[1] There are two equivalent ways of writing an AIPW estimator: either as an IPW estimator, augmented with an additional term (based on imputation), or as an imputation estimator, augmented with an additional term (based on reweighting of prediction errors).[2] Either of these adjustment terms can be understood as a form of bias correction, but it is this second formulation which arguably provides good intuition for the additional robustness of this estimator. When the conditional mean is correctly specified, the adjustment term – again, based on reweighting of prediction errors – has expectation zero. When the propensity score is correctly specified, this same term consistently estimates the bias of imputation. The resulting estimator is asymptotically normal and locally efficient: when both models are correctly specified, AIPW achieves the semiparametric efficiency bound.

That this estimator is robust to misspecification of at most one of the working models was demonstrated in later work by Scharfstein *et al.* (1999), and the term "doubly robust" was introduced by Robins *et al.* (2000). Doubly robust estimators continue to be an important topic of research in statistics, both in causal inference

[1] However, Kang and Schafer (2007) and Long *et al.* (2012) noted that this approach to estimation goes back at least as far as Cassel *et al.* (1976).

[2] A non-technical introduction to AIPW estimators was provided by Glynn and Quinn (2010). See also Kang and Schafer (2007) and Tan (2007).

and in missing data settings. For example, Bang and Robins (2005) extended this methodology to several panel data models and represented their estimators – which is similar to Scharfstein *et al.* (1999), but otherwise not typical – as imputation estimators, with the inverse of the propensity score included as an additional control variable. Tan (2006a) developed an alternative AIPW estimator, which might provide either an efficiency gain – if the propensity score is correctly specified – or bias reduction – if the propensity score is misspecified, but the conditional mean is correctly specified. Early simulation studies were presented by Lunceford and Davidian (2004) and Kang and Schafer (2007). While the former paper strongly encouraged "routine use of [AIPW] in practice", the latter study was somewhat pessimistic; its results triggered a considerable debate, with comments of Ridgeway and McCaffrey (2007), Robins *et al.* (2007), Tan (2007), Tsiatis and Davidian (2007), and a rejoinder.

Following this debate, further AIPW estimators – with better properties – have been developed. In particular, Cao *et al.* (2009) and Tan (2010) provided new estimators with increased efficiency and improved robustness against very small values of the propensity score. An estimator with more desirable efficiency properties was also studied by Rotnitzky *et al.* (2012). On the other hand, Hu *et al.* (2012) proposed an estimator with increased robustness to model misspecification: first, covariate information is collapsed into a two-dimensional score, with one dimension for the conditional mean of the outcome and the other for the propensity score; second, a regression of the outcome on this score is estimated using nonparametric methods. A sufficient condition for consistency of this estimator is that either of the two dimensions captures the "core" of the corresponding pattern. Finally, Vermeulen and Vansteelandt (2015) focused on bias reduction, and developed an estimator which minimizes the asymptotic bias under misspecification of both working models.

One shortcoming of the AIPW approach – which is clear from the discussion in Kang and Schafer (2007) and Robins *et al.* (2007) – is that the estimated mean functions are not only sensitive to extreme values of the propensity score estimates, but they can actually produce estimates outside the logically consistent range when the response is bounded in some way.

In recent years, there has also been substantive interest in doubly robust estimators in the econometric literature. Hirano and Imbens (2001) provided an early application to data on right heart catheterization; however, these authors used linear models for a binary outcome, which necessarily does not fully exploit the double robustness result. Wooldridge (2007) 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.[3] An important benefit of the approach in Wooldridge (2007) is that provided one chooses the conditional mean function so that it coheres with the range of the response variable, estimated counterfactual means are always within the logical range. Cattaneo (2010) used "doubly robust moment conditions" to construct efficient semiparametric estimators of multi-valued treatment effects, also extending the scope of applications to quantile treatment effects (QTEs). More recently, Graham *et al.* (2012) derived a new IPW estimator ("inverse probability tilting"), which replaces the maximum likelihood estimate of the propensity score with a particular method of moments estimate. This new estimator shares the properties of double robustness and local efficiency with previous methods, but it has smaller asymptotic bias under certain conditions.[4]

[3] This approach was applied to a multi-valued treatment effect framework and to a decomposition framework by Uysal (2015) and Kaiser (2016), respectively. Both of these papers also considered an extension to identification and estimation of the average treatment effect on the treated (ATT).

[4] In a different setting, Kline (2011) demonstrated that parametric imputation – also referred to as Oaxaca–Blinder – is also doubly robust in a particular way; namely, it has an alternative IPW representation, in which the weights are based on a linear model for the treatment odds. A similar result had also been demonstrated by Robins *et al.* (2007).

While the majority of these previous papers used parametric models for the conditional mean and the propensity score, and allowed one of these models to be arbitrarily misspecified, Rothe and Firpo (2015) studied the properties of “semiparametric doubly robust estimators” in which this first stage remains fully nonparametric. In this context, misspecification is no longer an issue, but Rothe and Firpo (2015) nevertheless demonstrated that such estimators, which exploit “doubly robust moment conditions” (as in Cattaneo, 2010), have desirable asymptotic properties, since their special structure automatically removes most of the largest “second order” terms. These same moment conditions were also used in recent papers by Belloni *et al.* (2014, 2015) and Farrell (2015) for treatment evaluation with high-dimensional data, including settings with more covariates than observations.

The discussion so far has focused on doubly robust estimators that require the treatment variable to be unconfounded conditional on covariates. A much smaller literature – primarily in statistics – has considered estimation of various parameters of interest in instrumental variable (IV) settings. An early contribution by Tan (2006b) studied doubly robust estimation of the local average treatment effect (LATE) in a model which, unlike Imbens and Angrist (1994), includes additional covariates. The estimator of Tan (2006b) is consistent if either the instrument propensity score is correctly specified, or both the first stage and the conditional mean of the outcome are correctly specified. Doubly robust estimators of the LATE were also studied by Uysal (2011) as well as in an early version of Rothe and Firpo (2015). Finally, a recent paper by Ogburn *et al.* (2015) studied doubly robust estimation of the dependence of the local average treatment effect on a subset of pre-treatment covariates.

Estimation of other parameters of interest has also been considered. In particular, Okui *et al.* (2012) studied doubly robust estimation of a finite-dimensional parameter indexing the dependence of the conditional mean of the outcome on the endogenous treatment variable. Tchetgen Tchetgen and Vansteelandt (2013) discussed a control function approach to estimating the conditional average treatment effect on the treated in an instrumental variable model; subsequently, a similar method for estimating the unconditional ATT was developed by Liu *et al.* (2015).

The current paper contributes to several strands of the literature on doubly robust estimators. We build on the framework of Wooldridge (2007) and expand it in several useful directions by focusing on treatment effects estimation. As mentioned above, we prefer the setup in Wooldridge (2007) because it leads to doubly robust estimation for important nonlinear as well as linear conditional mean specifications, and ensures that estimates of ATEs lie within logical ranges when the response variable is bounded in some way. A related point is that estimators based on a weighted objective function tend to be less sensitive to lots of variation in the estimated propensity score. Wooldridge (2007) only considered estimation of the ATE over the entire population, and did not present results for doubly robust estimation of ATEs for subpopulations. An important aspect of our unified framework is the introduction of an indicator that can select out subpopulations, thereby providing simple doubly robust estimators for a wide variety of treatment effects. This includes not only subpopulations defined by pre-treatment covariates, or average treatment effects for different treatment groups, but it introduces new possibilities. For example, after treatment has been assigned, some units may exhibit observed behavior – such as getting more education – and we can estimate average treatment effects for such populations.

To summarize, our paper shows how doubly robust estimators of various average treatment effects for the most common response variables can be studied in a single framework. This same framework can be used to obtain doubly robust estimators of distribution, quantile, and inequality treatment effects, as well as of the local average treatment effect (in an instrumental variable setting). Previous approaches are limited along one or more dimensions, either focusing only on the ATE, using only linear conditional means, or using moment conditions that can produce nonsensical estimates. As a technical improvement over Wooldridge (2007) and several of the other cited papers, we demonstrate identification using a conditional mean version of unconfoundedness, rather than full conditional independence between the treatment and potential outcomes.

3 Identification

We now introduce our notation and discuss the population parameters of interest. Also, we detail the assumptions that are needed for identification of these estimands as well as outline our main identification result. To avoid confusion with our notation, the discussion of distribution, quantile, and inequality treatment effects – as well as of our extensions to instrumental variable models – is deferred to Section 4. As will be explained, these applications are easily expressed within the following framework.

3.1 Notation 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 strongest form of unconfoundedness is conditional independence between the treatment assignment and 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

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

Under conditional independence,

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

In order to allow for a wide variety of treatment effects, we introduce a binary variable, D, which we will also assume to be unconfounded with respect to each Y_g. In applications, D might be a deterministic function of X, in which case its inclusion serves to isolate a subset of the population determined by pre-treatment covariates. Another important case is when D is an indicator for a different level of treatment. Yet another possibility is when D indicates a subpopulation formed after the treatment assignment. For example, in the evaluation of a job training program, some individuals might choose to obtain additional education unrelated to the job training program. If D is an indicator representing "more schooling," we would not want to include D in X, as that would generally cause unconfoundedness to be violated – see, for example, Wooldridge (2005). However, we may want to estimate the average treatment effect of the job training program itself for the group that subsequently sought additional schooling. Importantly, we will not have to impose any restrictions on the dependence between W and D. As far as we know, ours is the first framework to consider this possibility.

In what follows we 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

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

If D and Y_g are conditionally independent, then

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

although this is not the version of unconfoundedness we use for our main result.

3.2 A General Result on Weighting

Our general result applies to any function of the potential outcome, Y_g, and the observed covariates. Let $q(Y_g, X)$ denote such a function, where we assume $\mathbb{E}\left[|q(Y_g, X)|\right] < \infty$. In the following lemma – which is crucial for our main result – we demonstrate that, for all g, we can recover $\mathbb{E}\left[q(Y_g, X)|D = 1\right]$ from the distribution of observable variables.

Lemma 3.2: Assume that W_g and D are each unconfounded in conditional mean, that is,

$$\begin{aligned} \mathbb{E}\left[q(Y_g, X)|W_g, X\right] &= \mathbb{E}\left[q(Y_g, X)|X\right] & (11) \\ \mathbb{E}\left[q(Y_g, X)|D, X\right] &= \mathbb{E}\left[q(Y_g, X)|X\right]. & (12) \end{aligned}$$

Define $\eta = \mathbb{P}(D = 1) > 0$. Further, assume that $p_g(x) > 0$ for all $x \in \mathcal{X}$, where $p_g(x)$ is defined in (7). Then,

$$\frac{1}{\eta} \cdot \mathbb{E}\left[\frac{W_g}{p_g(X)} r(X) q(Y_g, X)\right] = \mathbb{E}\left[q(Y_g, X)|D = 1\right]. \square \tag{13}$$

This lemma is fairly general, partly due to the inclusion of the auxiliary variable, D, which defines our subpopulation of interest. The proof of this lemma uses unconfoundedness in the conditional mean separately for W_g and D. Naturally, if

$$\mathbb{E}\left[q(Y_g, X)|W_g, D, X\right] = \mathbb{E}\left[q(Y_g, X)|X\right],$$

then (11) and (12) both hold.

A number of applications of Lemma 3.2 are covered in Section 4. The proof of Lemma 3.2 is straightforward and can be found in Appendix A.

4 Applications

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 of our parameters of interest.

4.1 Average Treatment Effects under Unconfoundedness

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 $q(Y, X) = Y$ and $Y = (1 - W) \cdot Y_0 + W \cdot Y_1$, Lemma 3.2 implies

$$\tau_{ate} = \mathbb{E}(Y_1 - Y_0) = \mathbb{E}\left[\frac{W}{p(X)}Y - \frac{1-W}{1-p(X)}Y\right], \tag{14}$$

with $D = 1$ in both cases. This expression leads directly to the standard IPW estimator (Horvitz and Thompson, 1952). Similarly, we can use Lemma 3.2 to write the average treatment effect on the treated as

$$\tau_{att} = \mathbb{E}(Y_1 - Y_0 | W = 1) = \frac{1}{\mathbb{P}(W = 1)} \cdot \mathbb{E}\left[W \cdot Y - \frac{1-W}{1-p(X)}p(X) \cdot Y\right], \tag{15}$$

because $D = W$, $\eta = \mathbb{P}(W = 1)$, and $r(X) = p(X)$. More generally, we can write the average treatment effect for any subpopulation of interest as

$$\mathbb{E}(Y_1 - Y_0 | D = 1) = \frac{1}{\eta} \cdot \mathbb{E}\left[\frac{W}{p(X)}r(X) \cdot Y - \frac{1-W}{1-p(X)}r(X) \cdot Y\right], \tag{16}$$

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.

Multi-valued treatments: Let $Y = W_0 \cdot Y_0 + W_1 \cdot Y_1 + W_2 \cdot Y_2 + \cdots + W_G \cdot Y_G$. Then, with $q(Y, X) = Y$, Lemma 3.2 suggests that the average gain from switching from the control group to treatment g, $g \in \{1, 2, \ldots, G\}$, is

$$\tau_{g,ate} = \mathbb{E}(Y_g - Y_0) = \mathbb{E}\left[\frac{W_g}{p_g(X)}Y - \frac{W_0}{p_0(X)}Y\right]. \tag{17}$$

Similarly, the average treatment effect on those receiving treatment g, relative to the control group, is

$$\tau_{g,att} = \mathbb{E}(Y_g - Y_0|W = g) = \frac{1}{\mathbb{P}(W = g)} \cdot \mathbb{E}\left[W_g \cdot Y - \frac{W_0}{p_0(X)} p_g(X) \cdot Y\right]. \tag{18}$$

4.2 Distribution, Quantile, and Inequality Treatment Effects

Distribution treatment effects: In what follows, it is helpful to define an extended set of counterfactual outcomes, $Y_g(y) = 1[Y_g \leq y]$. In other words, we can create a set of binary variables, $Y_g(y)$, where y is any real number and $Y_g(y) = 1$ whenever $Y_g \leq y$. Also, $Y(y) = W_0 \cdot Y_0(y) + W_1 \cdot Y_1(y) + \cdots + W_G \cdot Y_G(y)$. If we let F_{Y_g} denote the unconditional cdf of Y_g, then, from Lemma 3.2, we can write:

$$F_{Y_g}(y) = \mathbb{P}(Y_g \leq y) = \mathbb{E}\left[Y_g(y)\right] = \mathbb{E}\left[\frac{W_g}{p_g(X)} Y(y)\right]. \tag{19}$$

As noted by Foresi and Peracchi (1995), when we vary the value of y, we can provide a useful characterization of F_{Y_g}. This idea was extended in a number of recent papers (especially in Chernozhukov *et al.*, 2013), and we will exploit this later. Now, using (19), we can identify the distribution treatment effect (DTE) of treatment g as

$$\tau_{g,dte}(y) = F_{Y_g}(y) - F_{Y_0}(y) = \mathbb{E}\left[\frac{W_g}{p_g(X)} Y(y) - \frac{W_0}{p_0(X)} Y(y)\right]. \tag{20}$$

Previous studies of distribution treatment effects include Abadie (2002), Lee (2009), Maier (2011), Chernozhukov *et al.* (2013), and Sant'Anna (2016). As far as we know, however, ours is the first paper to consider doubly robust estimation of this parameter.

Quantile treatment effects: There are many papers that consider identification and estimation of quantile treatment effects (QTEs), defined as

$$\tau_{g,qte}(t) = Q_{Y_g}(t) - Q_{Y_0}(t), \tag{21}$$

where $Q_{Y_g}(t)$ is the tth quantile of Y_g. Unlike early contributions, which usually modeled the quantiles directly (see, *e.g.*, Abadie *et al.*, 2002), we first obtain F_{Y_g} using (19). Then, we note that

$$Q_{Y_g}(t) = \inf\left\{u : F_{Y_g}(u) \geq t\right\}. \tag{22}$$

What follows,

$$\tau_{g,qte}(t) = \inf\left\{u : F_{Y_g}(u) \geq t\right\} - \inf\left\{v : F_{Y_0}(v) \geq t\right\}. \tag{23}$$

This approach to identifying quantile treatment effects was used by Cattaneo (2010), Frandsen *et al.* (2012), Chernozhukov *et al.* (2013), Frölich and Melly (2013), Donald and Hsu (2014), and Sant'Anna (2016), among others. However, only Cattaneo (2010) discussed doubly robust estimation of QTEs, which we allow in Section 5. Also, our approach to estimation differs from that in Cattaneo (2010).

Inequality treatment effects: A recent paper by Firpo and Pinto (2016) introduced an alternative approach to studying distributional impacts of interventions. Namely, such effects were modeled as differences in inequality measures between two marginal distributions of potential outcomes. Let $\upsilon : \mathcal{F}_\upsilon \rightarrow \mathbb{R}$ be an inequality measure, say the coefficient of variation, the interquartile range, the Theil index, or the Gini coefficient. As before, we first obtain F_{Y_g} using (19). Then we note, following

Firpo and Pinto (2016), that the inequality treatment effect (ITE) can be written as

$$\tau_{g,ite} = \upsilon(F_{Y_g}) - \upsilon(F_{Y_0}). \tag{24}$$

In the concluding section of their paper, Firpo and Pinto (2016) suggested that studying doubly robust estimators of ITEs would be an interesting avenue for further research. Such estimators of these parameters follow from Section 5.

4.3 Extensions to Instrumental Variable Estimation

Our general result can be applied in contexts where we require instrumental variables to provide exogenous variation in treatment assignment. In the following discussion we need to introduce new notation. In particular, let some instrumental variable – which satisfies the usual exclusion restriction – take on $H+1$ different values, labeled $\{0, 1, 2, \dots, H\}$. Let Z represent the instrument assignment. We also define

$$Z_h = 1[Z = h],\ h = 0, 1, 2, \dots, H. \tag{25}$$

Further, we introduce the instrument propensity score for each Z_h:

$$s_h(x) = \mathbb{P}(Z_h = 1|X = x). \tag{26}$$

If Z is unconfounded with respect to each counterfactual outcome and each counterfactual treatment, conditional on X, we can use Lemma 3.2 to separately identify the numerator and the denominator of the usual formula for the local average treatment effect (LATE). More precisely, and similarly to Tan (2006b), we can identify this parameter of interest as

$$\tau_{h,late} = \frac{\mathbb{E}\left[\frac{Z_h}{s_h(X)}Y - \frac{Z_0}{s_0(X)}Y\right]}{\mathbb{E}\left[\frac{Z_h}{s_h(X)}W - \frac{Z_0}{s_0(X)}W\right]}, \tag{27}$$

where both the numerator and the denominator are simply equal to the average effects of Z_h, compared with Z_0, on the outcome and the treatment, respectively. In other words, we identify these objects separately, using (17) in both cases.

We leave applications to identifying other parameters via instrumental variables to future research.

4.4 Unnormalized and Normalized Weights

In the previous setup, given a random sample $\{(W_{ig}, D_i, X_i, Y_i) : i = 1, 2, \ldots, N\}$, Lemma 3.2 suggests how to consistently estimate $\mu_{g,1} \equiv \mathbb{E}\left[q(Y_g, X)|D = 1\right]$:

$$\frac{1}{\hat{\eta}}\left[N^{-1}\sum_{i=1}^{N}\frac{W_{ig}}{p_g(X_i)}r(X_i)q(Y_i, X_i)\right], \tag{28}$$

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

$$\hat{\eta} = N^{-1}\sum_{i=1}^{N} D_i = N_D/N, \tag{29}$$

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

$$\hat{\mu}_{g,1,unnormalized} = N_D^{-1}\sum_{i=1}^{N}\frac{W_{ig}}{p_g(X_i)}r(X_i)q(Y_i, X_i). \tag{30}$$

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

$$\frac{1}{N_D}\frac{W_{ig}}{p_g(X_i)}r(X_i), \tag{31}$$

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 3.2 to $q(Y_g, X) \equiv 1$ to get

$$\eta = \mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)\right], \tag{32}$$

and so an alternative unbiased and consistent estimator of η is

$$\hat{\eta} = N^{-1}\sum_{i=1}^{N}\frac{W_{ig}}{p_g(X_i)}r(X_i). \tag{33}$$

When we plug this estimator in (28) for $\hat{\eta}$, we obtain

$$\hat{\mu}_{g,1,normalized} = \left[\sum_{i=1}^{N}\frac{W_{ig}}{p_g(X_i)}r(X_i)\right]^{-1}\sum_{i=1}^{N}\frac{W_{ig}}{p_g(X_i)}r(X_i)q(Y_i, X_i) \tag{34}$$

and now the weights,

$$\left[\sum_{j=1}^{N}\frac{W_{jg}}{p_g(X_j)}r(X_j)\right]^{-1}\frac{W_{ig}}{p_g(X_i)}r(X_i), \tag{35}$$

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

$$\min_{m_g \in \mathbb{R}} \sum_{i=1}^{N} \frac{W_{ig}}{p_g(X_i)} (Y_i - m_g)^2, \tag{36}$$

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

5 Estimation

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

$$\mu_{g,1} \equiv \mathbb{E}(Y_g | D = 1). \tag{37}$$

As we saw in Section 4, 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 3.2 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 5.2.

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.

5.1 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 $\theta_g^o \in \Theta_g$,

$$\mathbb{E}(Y_g|X = x) = m_g(x, \theta_g^o),\ x \in \mathcal{X}, \tag{38}$$

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 it can be written as

$$q(Y_g, X; \theta_g) = a\left[m_g(X, \theta_g)\right] + b(Y_g) + Y_g c\left[m_g(X, \theta_g)\right]. \tag{39}$$

Gourieroux *et al.* (1984) use this structure to show that θ_g^o is a solution to

$$\max_{\theta_g \in \Theta_g}\ \mathbb{E}[q(Y_g, X; \theta_g)|X] \tag{40}$$

for all outcomes X, which means

$$\mathbb{E}[q(Y_g, X; \theta_g^o)|X] \geq \mathbb{E}[q(Y_g, X; \theta_g)|X]. \tag{41}$$

For our purposes, another important feature of the LEF family is that it will suffice to assume that treatment assignment W_g is unconfounded in the mean:

$$\mathbb{E}\left(Y_g|W_g,X\right)=\mathbb{E}\left(Y_g|X\right); \tag{42}$$

we do not need the stronger conditional independence assumption in (5). This is because $b(Y_g)$ does not depend on the parameters, and so we can drop it from the objective function. The only other place in which Y_g appears is to multiply a function of X (and θ_g), and so it will suffice to impose (42).

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 standard binary response models by MLE. (More precisely, by quasi-MLE because we allow the binary response models to be misspecified.)

The weighted objective function for estimating θ_g^o – where W_{ig} selects the units in treatment group g – is

$$N^{-1}\sum_{i=1}^{N}\frac{W_{ig}}{F_g(X_i;\hat{\gamma}_g)}J(X_i;\hat{\psi})\cdot q(Y_i,X_i;\theta_g). \tag{43}$$

Using standard convergence results – for example, Newey and McFadden (1994) and Wooldridge (2010, Chapter 12) – (43) converges in probability to

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

An argument very similar to Lemma 3.2 shows that

$$\mathbb{E}\left[\frac{W_g}{F_g(X;\gamma_g^*)}J(X;\psi^*)\cdot q(Y_g,X;\theta_g)\right]=\mathbb{E}\left\{\frac{p_g(X)J(X;\psi^*)}{F_g(X;\gamma_g^*)}\mathbb{E}[q(Y_g,X;\theta_g)|X]\right\}. \quad (44)$$

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

$$\frac{p_g(X)J(X;\psi^*)}{F_g(X;\gamma_g^*)}\mathbb{E}[q(Y_g,X;\theta_g^o)|X]\geq\frac{p_g(X)J(X;\psi^*)}{F_g(X;\gamma_g^*)}\mathbb{E}[q(Y_g,X;\theta_g)|X] \quad (45)$$

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

$$\max_{\theta_g\in\Theta_g}\mathbb{E}\left[\frac{W_g}{F_g(X;\gamma_g^*)}J(X;\psi^*)\cdot q(Y_g,X;\theta_g)\right] \quad (46)$$

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,

$$\mathbb{E}(Y_g|X)=m_g(X,\theta_g^o). \quad (47)$$

Because D is unconfounded conditional on X,

$$\mathbb{E}(Y_g|X,D)=\mathbb{E}(Y_g|X) \quad (48)$$

and so

$$\mathbb{E}(Y_g|D=1)=\mathbb{E}[m_g(X,\theta_g^o)|D=1]. \quad (49)$$

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

$$\hat{\mu}_{g,1} = N_D^{-1} \sum_{i=1}^{N} D_i \cdot m_g(X_i, \hat{\theta}_g), \tag{50}$$

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

5.2 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 $\mathbb{P}(D = 1|X = x)$:

$$\mathbb{P}(W_g = 1|X = x) = F(x, \gamma_g^o) \tag{51}$$

$$\mathbb{P}(D = 1|X = x) = J(x, \psi^o), \tag{52}$$

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 3.2 and the structure of the QLL in the LEF – see (39) – we know that because

$$\frac{1}{\eta} \cdot \mathbb{E}\left[\frac{W_g}{F(X, \gamma_g^o)} J(X, \psi^o) \cdot q(Y_g, X; \theta_g)\right] = \mathbb{E}\left[q(Y_g, X; \theta_g)|D = 1\right] \tag{53}$$

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

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

By the convergence arguments in Section 5.1, the solution $\hat{\theta}_g$ to (43) is consistent for θ_g^*. So it remains to be shown 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

$$\mathbb{E}(Y_g|D=1) = \mathbb{E}[m_g(X, \theta_g^*)|D=1]. \tag{55}$$

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

$$\mathbb{E}(Y_g|D=1) = \mathbb{E}(\alpha_g^* + X\beta_g^*|D=1). \tag{56}$$

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

$$m_g(x, \theta_g) = \Lambda(\alpha_g + x\beta_g), \tag{57}$$

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 (2016) 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. In addition, D can be an indicator of being in a subpopulation formed after treatment assignment, thereby allowing us to estimate treatment effects for groups that have certain behavioral responses to treatment. This appears to be novel.

6 Summary Discussion and Wider Issues

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 also extend this initial result to distribution, quantile, and inequality treatment effects, as well as to the local average treatment effect in a setting with an instrumental variable. Further, 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.

Our contribution is essentially methodological, and we do not directly contribute to the philosophical underpinnings of causal inference, or to the debate on the proper way to define causality. Nevertheless, our setup applies to several of the causal frameworks discussed in the recent issue of this *journal* in honor of Trygve Haavelmo's contributions to structural equations and causal inference. As was often the case in econometrics, the early giants in the field had a coherent, deep understanding of their area of research, and Haavelmo is a leading example. He understood that, when economists specify, say, a supply and demand system, each has a hypothetical or counterfactual interpretation, and identification is synonymous with making assumptions about how observables are generated from the underlying economic equations. Over the years there was slippage in the empirical application of simultaneous equations models in that researchers applied them to settings where the underlying equations did not satisfy Ragnar Frisch's *autonomy* requirement, as discussed recently by Heckman and Pinto (2015). See also the *Econometric Theory* interview with Arthur Goldberger (Kiefer, 1989), who lamented that the progress made by Frisch, Haavelmo, and others at the Cowles Commission had eroded.

The papers by Heckman and Pinto (2015) and Pearl (2015) explicitly use a counterfactual setting of the type we use here. Pearl's *do*-calculus is what our framework captures in the counterfactual outcomes Y_g for different treatment levels g. Had we used a different notation, say $Y(w)$, where $Y(w)$ is the random outcome with, say, the price or policy set at w, then studying changes in $Y(w)$ as w changes is precisely the purpose of Pearl's *do*-calculus for changing w. Heckman and Pinto (2015, p. 118–119) provide a very clear discussion of Haavelmo's hypothetical function in the context of the linear model. While we do not consider continuous "treatment" variables W in this paper, the idea of hypothetical or potential outcomes is paramount. Future research could study extensions of our results in a setting that allows for multiple "treatment" variables that can be continuous, discrete, or some mixture.

A Proofs

Proof of Lemma 3.2: The proof that

$$\mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)q(Y_g,X)\right]=\mathbb{E}\left[r(X)q(Y_g,X)\right] \tag{58}$$

is similar to Wooldridge (2007). However, Wooldridge assumes conditional independence rather than the weaker conditional mean independence we use here. The current proof is an implication of iterated expectations and unconfoundedness in mean:

$$\begin{aligned}\mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)q(Y_g,X)\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)q(Y_g,X)\middle| W_g,X\right]\right\}\\ &= \mathbb{E}\left\{\frac{W_g}{p_g(X)}r(X)\mathbb{E}\left[q(Y_g,X)|W_g,X\right]\right\}\\ &= \mathbb{E}\left\{\frac{W_g}{p_g(X)}r(X)\mathbb{E}\left[q(Y_g,X)|X\right]\right\}\\ &\equiv \mathbb{E}\left\{\frac{W_g}{p_g(X)}r(X)h_g(X)\right\},\end{aligned} \tag{59}$$

where we use $\mathbb{E}\left[q(Y_g,X)|W_g,X\right]=\mathbb{E}\left[q(Y_g,X)|X\right]\equiv h_g(X)$. Now apply iterated expectations again:

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

because $\mathbb{E}\left(W_g|X\right)=p_g(X)$. Another application of iterated expectations gives the result because

$$\mathbb{E}\left[r(X)q(Y_g,X)\right]=\mathbb{E}\left\{r(X)\mathbb{E}\left[q(Y_g,X)|X\right]\right\}=\mathbb{E}\left[r(X)h_g(X)\right].$$

Next, we show that

$$\mathbb{E}\left[r(X)q(Y_g,X)\right]=\mathbb{E}\left[D\cdot q(Y_g,X)\right] \tag{60}$$

which again follows by iterated expectations and unconfoundedness in mean:

$$\begin{aligned}\mathbb{E}\left[D\cdot q(Y_g,X)\right]&=\mathbb{E}\left\{\mathbb{E}\left[D\cdot q(Y_g,X)\right|D,X]\right\}\\&=\mathbb{E}\left\{D\cdot\mathbb{E}\left[q(Y_g,X)|W_g,X\right]\right\}\\&=\mathbb{E}\left[D\cdot h_g(X)\right]\\&=\mathbb{E}\left[r(X)h_g(X)\right]=\mathbb{E}\left[D\cdot q(Y_g,X)\right].\end{aligned} \tag{61}$$

Finally,

$$\begin{aligned}\mathbb{E}\left[D\cdot q(Y_g,X)\right]=(1-\eta)\cdot\mathbb{E}\left[D\cdot q(Y_g,X)|D=0\right]+\eta\cdot\mathbb{E}\left[D\cdot q(Y_g,X)|D=1\right]\\=\eta\cdot\mathbb{E}\left[q(Y_g,X)|D=1\right].\end{aligned} \tag{62}$$

Combining the three pieces gives

$$\mathbb{E}\left[\frac{W_g}{p_g(X)}r(X)q(Y_g,X)\right]=\eta\cdot\mathbb{E}\left[q(Y_g,X)|D=1\right], \tag{63}$$

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

B Implementation in Stata

We now discuss possible applications of our estimation methods in Stata, as well as previous implementations of doubly robust estimators in this software package, such as Emsley *et al.* (2008), Cattaneo *et al.* (2013), and the `teffects aipw` and `teffects ipwra` commands in Stata 13.

In the first implementation that we are aware of, Emsley *et al.* (2008) introduced a Stata command for the standard AIPW estimator of the ATE (`dr`). On the other hand, Cattaneo *et al.* (2013) provided a Stata command (`poparms`) for estimating treatment effects of multivalued treatments, including QTEs, but excluding, for example, the ATT. The implemented (semiparametric) estimators are based on earlier work of Cattaneo (2010), and hence both the conditional mean and the propensity score are estimated using series estimators. When no polynomials are included in these models, `poparms` will overlap with `dr`.

The standard AIPW estimator of the ATE is also implemented as `teffects aipw` in Stata 13. Again, when no polynomials are included in `poparms`, it will overlap with `teffects aipw`. Because our estimation methods are not AIPW, none of these implementations will overlap with our approach. To the best of our knowledge, the only exception is `teffects ipwra` in Stata 13 – which implements the estimation methods of Wooldridge (2007) as well as some extensions of these methods which we discuss in this paper. We provide further discussion in the following.

For simplicity, let us consider the case of a binary treatment. Let `ovar` and `tvar` be the names of the outcome and treatment variables, respectively, and let `xvars` be the list of names of control variables. We start with estimating the propensity scores.

```
.  logit tvar xvars
.  predict pscore
```

When the mean function is linear, we can estimate the ATE in the following way:

```
.  regress ovar xvars if tvar == 1 [pw = 1/pscore]

.  predict ot

.  regress ovar xvars if tvar == 0 [pw = 1/(1-pscore)]

.  predict oc

.  generate te = ot-oc

.  summarize te
```

An identical estimate can be obtained by typing:

```
.  teffects ipwra (ovar xvars) (tvar xvars)
```

When the mean function is logistic or exponential, we can replace `regress` in the previous procedure with `logit` or `poisson`, respectively. We can also obtain the same estimates by simply typing:

```
.  teffects ipwra (ovar xvars, logit) (tvar xvars)

.  teffects ipwra (ovar xvars, poisson) (tvar xvars)
```

Alternatively, with a linear mean function, we can estimate the ATT in the following way:

```
.  regress ovar xvars if tvar == 1

.  predict ot

.  regress ovar xvars if tvar == 0 [pw = pscore/(1-pscore)]

.  predict oc

.  generate te = ot-oc

.  summarize te if tvar == 1
```

Again, an identical estimate can be obtained by typing:

```
.  teffects ipwra (ovar xvars) (tvar xvars), atet
```

Extensions to logistic and exponential mean functions are analogous and straightforward. On the other hand, quantile and inequality treatment effects as well as various parameters in instrumental variable models cannot be readily estimated using `teffects ipwra`. A comprehensive implementation of estimators for all these cases is beyond the scope of this paper, and we conclude this discussion with an implementation of our estimation method for DTEs and QTEs.

Let `y` be the value of the outcome variable in which we are interested, that is, let it be equal to y in $\tau_{1,dte}(y)$. We can then estimate the DTE in the following way:

```
.  generate oy = ovar<=y
.  logit oy xvars if tvar == 1 [pw = 1/pscore]
.  predict ot
.  logit oy xvars if tvar == 0 [pw = 1/(1-pscore)]
.  predict oc
.  generate te = ot-oc
.  summarize te
```

To obtain the QTE for a given quantile, it is sufficient to vary the value of y (`y`) and estimate average values of $Y_1(y)$ (`ot`) and $Y_0(y)$ (`oc`) for each value. We can then estimate $Q_{Y_1}(t)$ or $Q_{Y_0}(t)$ by finding the smallest y for which the estimated average value of $Y_1(y)$ or $Y_0(y)$, respectively, is greater than or equal to t.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association* 97, 284–292.

Abadie, A., Angrist, J. & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70, 91–117.

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

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2015). Program evaluation with high-dimensional data. Unpublished.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81, 608–650.

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

Cassel, C. M., Särndal, C. E. & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 63, 615–620.

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

Cattaneo, M. D., Drukker, D. M. & Holland, A. D. (2013). Estimation of multivalued treatment effects under conditional independence. *Stata Journal* 13, 407–450.

Chernozhukov, V., Fernández-Val, I. & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica* 81, 2205–2268.

Donald, S. G. & Hsu, Y.-C. (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics* 178, 383–397.

Emsley, R., Lunt, M., Pickles, A. & Dunn, G. (2008). Implementing double-robust estimators of causal effects. *Stata Journal* 8, 334–353.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 189, 1–23.

Firpo, S. & Pinto, C. (2016). Identification and estimation of distributional impacts of interventions using changes in inequality measures. *Journal of Applied Econometrics* 31, 457–486.

Foresi, S. & Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association* 90, 451–466.

Frandsen, B. R., Frölich, M. & Melly, B. (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics* 168, 382–395.

Frölich, M. & Melly, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics* 31, 346–357.

Glynn, A. N. & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* 18, 36–56.

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

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

Heckman, J. & Pinto, R. (2015). Causal analysis after Haavelmo. *Econometric Theory* 31, 115–151.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology* 2, 259–278.

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

Hu, Z., Follmann, D. A. & Qin, J. (2012). Semiparametric double balancing score estimation for incomplete data with ignorable missingness. *Journal of the American Statistical Association* 107, 247–257.

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

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

Kaiser, B. (2016). Decomposing differences in arithmetic means: A doubly robust estimation approach. *Empirical Economics* 50, 873–899.

KANG, J. D. Y. & 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.

KIEFER, N. M. (1989). The ET interview: Arthur S. Goldberger. *Econometric Theory* 5, 133–160.

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

LEE, M.-J. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses. *Journal of the Royal Statistical Society: Series B* 71, 243–264.

LIU, L., MIAO, W., SUN, B., ROBINS, J. & TCHETGEN TCHETGEN, E. (2015). Doubly robust estimation of a marginal average effect of treatment on the treated with an instrumental variable. Unpublished.

LONG, Q., HSU, C.-H. & LI, Y. (2012). Doubly robust nonparametric multiple imputation for ignorable missing data. *Statistica Sinica* 22, 149–172.

LUNCEFORD, J. K. & 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.

MAIER, M. (2011). Tests for distributional treatment effects under unconfoundedness. *Economics Letters* 110, 49–51.

NEWEY, W. K. & MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In: *Handbook of Econometrics* (ENGLE, R. F. & MCFADDEN, D., eds.), vol. 4. North Holland.

Ogburn, E. L., Rotnitzky, A. & Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B* 77, 373–396.

Okui, R., Small, D. S., Tan, Z. & Robins, J. M. (2012). Doubly robust instrumental variable regression. *Statistica Sinica* 22, 173–205.

Pearl, J. (2015). Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory* 31, 152–179.

Ridgeway, G. & McCaffrey, D. F. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22, 540–543.

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

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

Robins, J. M., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science* 22, 544–559.

Rothe, C. & Firpo, S. (2015). Semiparametric two-step estimation using doubly robust moment conditions. Unpublished.

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

Sant'Anna, P. H. C. (2016). Program evaluation with right-censored data. Unpublished.

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

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

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

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science* 22, 560–568.

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

Tchetgen Tchetgen, E. J. & Vansteelandt, S. (2013). Alternative identification and inference for the effect of treatment on the treated with an instrumental variable. Unpublished.

Tsiatis, A. A. & Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22, 569–573.

Uysal, S. D. (2011). Doubly robust IV estimation of the local average treatment effect. Unpublished.

Uysal, S. D. (2015). Doubly robust estimation of causal effects with multivalued treatments: An application to the returns to schooling. *Journal of Applied Econometrics* 30, 763–786.

Vermeulen, K. & Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* 110, 1024–1036.

Wooldridge, J. M. (2005). Violating ignorability of treatment by controlling for too many factors. *Econometric Theory* 21, 1026–1028.

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 ed.