Doubly Robust Estimation of the Local Average Treatment Effect Curve

Elizabeth L. Ogburn

Johns Hopkins University, Baltimore, USA.

Andrea Rotnitzky

Di Tella University, Buenos Aires, Argentina and Harvard University, Boston, USA.

James M. Robins

Harvard University, Boston, USA.

Abstract. We consider estimation of the causal effect of a binary treatment on an outcome, conditional on covariates, from observational studies or natural experiments in which there is a binary instrument for treatment. We describe a doubly robust, locally efficient estimator of the parameters indexing a model for the local average treatment effect conditional on covariates $\mathbf{V}$ when randomization of the instrument is only true conditional on a high dimensional vector of covariates $\mathbf{X}$, possibly bigger than $\mathbf{V}$. We discuss the surprising result that inference is identical to inference for the parameters of a model for an additive treatment effect on the treated conditional on $\mathbf{V}$ that assumes no treatment-instrument interaction. We illustrate our methods with the estimation of the local average effect of participating in 401(k) retirement programs on savings using data from the U.S. Census Bureau's 1991 Survey of Income and Program Participation.

Keywords: Instrumental variables; Multiplicative effect; LATE; Local efficiency.

1. Introduction

Economists and biostatisticians have long been concerned with the problem of how to estimate the causal effect of a treatment on an outcome of interest, and how this effect is modified by baseline covariates. Estimation of average treatment effects is often facilitated by the unconfoundedness assumption that a vector of measured covariates suffices to control for all confounding of the treatment-outcome relationship. When this assumption is thought implausible, but instrumental variables satisfying the monotonicity assumption given in section 2.1 are available, it is possible to estimate the so called local average treatment effect contrasts. These are treatment effect contrasts for the subpopulation of compliers, i.e. subjects

for whom treatment and instrument agree. Beginning with the seminal paper of Imbens and Angrist (1994), non- and semiparametric instrumental variable methods for estimation of local average treatment effects have received considerable attention in the literature (Angrist and Imbens, 1995; Angrist, Imbens and Rubin, 1996; Angrist, Graddy and Imbens, 2000; Abadie, 2002, 2003; Abadie, Angrist and Imbens, 2002; Froelich, 2007; Tan, 2006a, 2010; Kasy, 2009, Cheng, Small, Tan and Ten Have, 2009, Cheng, Qin and Zhang, 2009).

In this paper we consider estimation of models for the dependence of local average treatment effects on baseline covariates $\mathbf{V}$. We assume that the treatment and instrument are binary and that the outcome support is either the real line, the non-negative real line or the non-negative integers. Like Abadie (2003), Tan (2006a), Froelich (2007), and Uysal (2011), we consider settings in which conditioning on a set of covariates $\mathbf{X}$ is necessary in order for the identifying instrumental variable assumptions to be valid. These settings are important because in practice the instrument may itself be confounded, and conditioning on covariates $\mathbf{X}$ may be required to make the key condition of instrument randomization plausible (Abadie, 2003). We extend the work of these authors to allow $\mathbf{X}$ to be larger than $\mathbf{V}$. This is an important contribution of our methodology, providing desirable flexibility in the definition of the target estimand as often investigators wish to report the treatment effect at low aggregation levels. Specifically, the covariate vector $\mathbf{X}$ is the set of variables that must be conditioned on in order for the instrument-outcome and instrument-treatment relationships to be unconfounded within levels of covariates; however, local average treatment effects conditional on $\mathbf{V}$, a subset of $\mathbf{X}$, may be the relevant contrasts to help guide decision makers who, due to limited resources, will have access only to information about the subset $\mathbf{V}$ of $\mathbf{X}$. For example, consider a study conducted in a sophisticated health maintenance organization (HMO). Suppose that the instrument is the therapy prescribed by the physician, the treatment is the therapy actually followed by the patient, and $\mathbf{X}$ is a vector of measured risk factors for the outcome that were used by the HMO physician to decide on the therapy prescription. The covariates $\mathbf{X}$ could include the results of expensive tests administered to patients at high risk for disease, such as magnetic resonance angiograms, that would not be available to community physicians. Thus, community physicians would need to decide what therapy to prescribe based on just the subset $\mathbf{V}$ of $\mathbf{X}$ that encodes the data available to them. Estimation of effect modification of the local average treatment effects by $\mathbf{V}$ is then critical to enable community physicians to make informed treatment decisions.

The literature on local average treatment effects has primarily focused on the estimation of the local average treatment effect on the additive scale (LATE), defined as the difference in means of the two potential out-

comes (under treatment and under no treatment) in the subpopulation of compliers. Identification of the multiplicative local average treatment effect contrast (MLATE), i.e. the ratio of the potential outcome means among compliers, follows trivially from results of Abadie (2003) but, to our knowledge, estimators of parametric specifications for the dependence of MLATE on covariates has not been discussed in the literature. In this paper we consider estimation of models for LATE and MLATE as functions of $\mathbf{V}$.

When the dimension of the covariate vector $\mathbf{X}$ is large, as will often be required in practice in order for the assumption of a conditionally unconfounded instrument to hold, nonparametric estimation of LATE (Froelich, 2007), of MLATE, and of parametric specifications for the dependence of these contrasts on covariates $\mathbf{V}$ is not feasible, due to the curse of dimensionality. When $\mathbf{V}$ is null, Tan (2006a) and Uysal (2011) derived estimators of LATE that are consistent provided either two models for two specific conditional means given the instrument and $\mathbf{X}$, or a model for the instrument propensity score (the probability that the instrument is equal to 1 conditional on the covariates $\mathbf{X}$) are correctly specified. In this paper we derive a new class of doubly robust estimators of parametric specifications for the dependence of LATE or MLATE on covariates $\mathbf{V}$ which remain consistent and asymptotically normal provided that either the propensity score model or a model for another conditional mean given the instrument and $\mathbf{X}$ are correctly specified. When $\mathbf{V}$ is non-null, the conditional mean models required by our doubly robust estimator are guaranteed to cohere with a parametric specification for the dependence of the local average treatment effect on $\mathbf{V}$. Extensions of the doubly robust methods proposed by Tan and Uysal to the case $\mathbf{V}$ non-null do not have this property.

In Section 2 we introduce the notation, models, and assumptions. We also review existing non and semiparametric methods for estimating local average treatment effects with instruments confounded by $\mathbf{X}$. In Section 3 we describe the proposed doubly robust estimating procedures, discuss efficiency properties and estimation under incorrect specifications for the dependence of LATE or MLATE on $\mathbf{V}$. In Section 4 we explain a surprising result earlier noted in the absence of covariates $\mathbf{X}$ by Clarke and Windmeijer (2010): inference under our models for the local average treatment effects is identical to inference under models proposed by Robins (1994) and Tan (2010) for a very different causal effect measure, namely the treatment effect on the treated. In Section 5 we re-analyze the data used in Poterba, Venti and Wise (1995) and Abadie (2003) with the goal of estimating the causal effect of participating in 401(k) retirement programs on savings using eligibility for a 401(k) program as a binary instrument. Section 6 concludes the article.

2. Background and notation

Suppose that we observe a random sample of size n of the vector $O = (Z, D, \mathbf{X}, Y)$, where D is a binary variable denoting the presence $(D = 1)$ or the absence $(D = 0)$ of a treatment whose effect on the outcome Y we wish to investigate, $\mathbf{X}$ is a vector of baseline covariates, and Z is a binary instrumental variable. Define D_z to be the potential treatment status that would be observed if Z were externally set to z, and define Y_{dz} to be the potential outcome that would be observed if D were externally set to d and Z to z, with $d, z = 0, 1$. Following Angrist et al. (1996), we say a subject is a complier if $D_1 > D_0$, an always taker if $D_1 = D_0 = 1$, a never taker if $D_1 = D_0 = 0$, and a defier if $D_1 < D_0$.

2.1. Assumptions and identification

Following Abadie (2003), Tan (2006a), Froelich (2007), and Uysal (2011), we assume:

(i) Conditional unconfoundedness of the instrument: $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1)$ is conditionally independent of Z given $\mathbf{X}$.

(ii) Exclusion of the instrument: $P(Y_{1d} = Y_{0d}) = 1$ for $d \in \{0, 1\}$.

(iii) Common support of the instrument: $0 < P(Z = 1|\mathbf{X}) < 1$ with probability 1 (w.p.1).

(iv) Instrumentation: $P(D_1 = 1|\mathbf{V}) \neq P(D_0 = 1|\mathbf{V})$ w.p.1.

(v) Monotonicity: $P(D_1 \geq D_0) = 1$.

(vi) Consistency: $Y = DY_1 + (1 - D)Y_0$, $D = ZD_1 + (1 - Z)D_0$, where $Y_d \equiv Y_{1d} = Y_{0d}$ by (ii).

When assumptions (i)-(iv) and (vi) hold, Z is said to be an instrumental variable for the effect of D on Y. Assumption (i) says that, within levels of $\mathbf{X}$, Z is as good as randomly assigned. Assumption (ii) postulates that the effect of Z on the outcome is entirely mediated by D. It implies that Y_{dz} is independent of z, and therefore we write Y_d throughout. Assumption (iii) requires there to be a positive probability of receiving each instrument value within each level of $\mathbf{X}$ or, equivalently, that the support of $\mathbf{X}$ is the same among those with $Z = 1$ and $Z = 0$. Assumption (v) excludes the existence of defiers. Assumption (vi) states that the observed outcome is equal to the potential outcome evaluated at the observed treatment value, and that the observed treatment is equal to the potential treatment evaluated at the observed instrument value. Finally, under assumption (v), assumption (iv) is the same as $P(D_1 = 1|\mathbf{V}) > P(D_0 = 1|\mathbf{V})$ which, in turn, under (i) and (vi) it is the same as $P(D = 1|Z = 1, \mathbf{V}) > P(D = 1|Z = 0, \mathbf{V})$. So it is tantamount to the assumption of positive correlation between Z and D. Abadie (2003) noted that assumptions (i)-(vi) are conditional versions of the assumptions made by Angrist et al. (1996), and Vytlacil (2002) noted that they are equivalent to the assumptions imposed by a nonparametric selection model (Heckman, 1976) in which

treatment is seen as an indicator of whether a latent index, e.g. expected treatment utility, has crossed a particular threshold.

Abadie (2003) showed that under assumptions (i)-(vi) $E(Y_1|D_1 > D_0, \mathbf{V})$ and $E(Y_0|D_1 > D_0, \mathbf{V})$ are identified, and consequently so is

$$LATE(\mathbf{v}) \equiv E(Y_1|D_1 > D_0, \mathbf{V} = \mathbf{v}) - E(Y_0|D_1 > D_0, \mathbf{V} = \mathbf{v}).$$

Under the additional assumption

(vii) Non-null complier mean under control: $E\{Y_0 \mid D_1 > D_0, \mathbf{V}\} \neq 0$ w.p.1, the contrast

$$MLATE(\mathbf{v}) \equiv E(Y_1|D_1 > D_0, \mathbf{V} = \mathbf{v}) / E(Y_0|D_1 > D_0, \mathbf{V} = \mathbf{v})$$

is well defined with probability 1 and identified.

For conciseness, we will refer to assumptions (i)-(vi) if referring to inference about $LATE(\cdot)$ or (i)-(vii) if referring to inference about $MLATE(\cdot)$ as the instrumental variable (IV) assumptions.

The curves $LATE(\mathbf{v})$ and $MLATE(\mathbf{v})$ describe how treatment effects in the complier subpopulation vary with values $\mathbf{v}$ of $\mathbf{V}$, the first quantifying the effects on an additive scale and the second on a multiplicative scale. Theorem 3.1 in Abadie (2003) implies that under the IV assumptions $LATE(\mathbf{v})$ is equal to the conditional version of the IV estimand,

$$IV(\mathbf{v}) \equiv \frac{E\{E(Y|Z=1,\mathbf{X}) - E(Y|Z=0,\mathbf{X})|\mathbf{V}=\mathbf{v}\}}{E\{E(D|Z=1,\mathbf{X}) - E(D|Z=0,\mathbf{X})|\mathbf{V}=\mathbf{v}\}}, \tag{1}$$

and $MLATE(\mathbf{v})$ is equal to

$$MIV(\mathbf{v}) \equiv -\frac{E\{E(YD|Z=1,\mathbf{X}) - E(YD|Z=0,\mathbf{X})|\mathbf{V}=\mathbf{v}\}}{E[E\{Y(1-D)|Z=1,\mathbf{X}\} - E\{Y(1-D)|Z=0,\mathbf{X}\}|\mathbf{V}=\mathbf{v}]}. \tag{2}$$

To our knowledge, the specific expression (1) for the functional identifying $LATE(\mathbf{V})$ with $\mathbf{V}$ null first appeared in Tan (2006a). The M in front of the acronym MIV is a reminder that this functional identifies a multiplicative treatment effect. The functionals $IV(\cdot)$ and $MIV(\cdot)$ are the target of inference when, as we will assume throughout, the IV assumptions are valid and interest is in estimation of $LATE(\cdot)$ and $MLATE(\cdot)$.

2.2. Review of existing estimators

The estimators that we will propose in Section 3 can accommodate any setting in which $\mathbf{V}$ is a subset of $\mathbf{X}$. Previous proposals for estimators of LATE have generally only considered the special cases in which $\mathbf{V}$ is null or $\mathbf{V}$ is equal to $\mathbf{X}$; to our knowledge the case in which $\mathbf{V}$ is a strict, non-empty subset of $\mathbf{X}$ has not been addressed in the literature.

For the special case in which $\mathbf{V}$ is null, Froelich (2007) studied the asymptotic distribution theory of estimators of the IV functional that

rely on two distinct nonparametric estimation methods for the four curves $E(Y|Z=z,\mathbf{X}=\cdot)$ and $E(D|Z=z,\mathbf{X}=\cdot)$, $z=0,1$, namely local polynomial regression and nonparametric series regression. His estimators, however, suffer from the curse of dimensionality. If the dimension of $\mathbf{X}$ is large, as will be the case in many applications in order to render the unconfoundedness assumption plausible, the IV functional will not in general be estimable in moderately sized samples, essentially because no two units will have values of $\mathbf{X}$ close enough to each other to allow for the borrowing of information needed for the smoothing implicit in these methods. Again for the special case in which $\mathbf{V}$ is null, Tan (2006a) considered estimating the IV functional under parametric models for each of the conditional means $E(Y|D=d,Z=z,\mathbf{X}=\cdot)$ and $E(D|Z=z,\mathbf{X}=\cdot)$, $d,z=0,1$. The consistency of the estimator of the IV functional then hinges on the correct specification of both of these models. See Section 3 for a contrast between these models and the models that must be specified to carry out the doubly robust estimation approach proposed in this paper.

Neither Froelich nor Tan (2006a) addressed the case when $\mathbf{V}$ is a nonempty, strict subset of $\mathbf{X}$, but further difficulties arise for each of their strategies in this case. Extending Froelich's approach to nonparametrically estimate the functionals $IV(\mathbf{V})$ and $MIV(\mathbf{V})$ not only requires smooth estimators of the aforementioned conditional means, but also of the conditional means given $\mathbf{V}$ of the differences involved in the numerators and denominators of these functionals. One possible extension of Tan's (2006a) fully parametric approach along the lines proposed in that paper for the case $\mathbf{X}=\mathbf{V}$, would also require specifying parametric models for the conditional means given $\mathbf{V}$ in the numerator and denominator of the $IV(\mathbf{V})$ functional. As noted by Abadie (2003), this approach will generally produce parametric specifications for the $LATE(\cdot)$ and $MLATE(\cdot)$ curves that are difficult to interpret. For example, linear specifications for each of the four conditional on $\mathbf{V}$ mean functions involved in the $IV(\mathbf{V})$ functional do not imply a linear model for $LATE(\mathbf{V})$. An alternative strategy that avoids this particular difficulty would be to use the approach of Tan (2010); however this latter approach involves specifying working models that may not cohere with the assumed model for $LATE(\cdot)$.

For the special case in which $\mathbf{V}$ is null, and with the goal of reducing sensitivity to model misspecification, Tan (2006a) and Uysal (2011) described doubly robust estimators of the IV functional whose consistency depends on correct parametric specification either of the instrument propensity score or, in the case of Uysal, of $E(Y|Z=z,\mathbf{X}=\cdot)$ and $E(D|Z=z,\mathbf{X}=\cdot)$, $z=0,1$, and, in the case of Tan, of $E(Y|D=d,Z=z,\mathbf{X}=\cdot)$ and $E(D|Z=z,\mathbf{X}=\cdot)$, $d,z=0,1$.

The special case of $\mathbf{V}$ equal to $\mathbf{X}$ was considered by Abadie (2003), Tan (2006a), Hirano et al. (2000), and Little and Yau (1998). Tan's (2006a) estimator of $LATE(\mathbf{X})$ again requires parametric specifications of the four

conditional expectations involved in the $IV(\mathbf{X})$ functional, which results in a specification of $LATE(\mathbf{X})$ that may be difficult to interpret. Hirano et al. (2000) and Little and Yau (1998) specified fully parametric likelihood functions for the observed data and unobserved compliance types (complier, defier, always taker, never taker) and used Bayesian methods to estimate the posterior distribution of Y conditional on compliance type, treatment, and instrument. Abadie (2003) proposed an estimating procedure in which models for $E(Y_d|D_1 > D_0, \mathbf{X} = \cdot)$, $d = 0, 1$ ensure that the resulting model for $LATE(\mathbf{X})$ is easily interpretable. His method hinges on consistent estimation of the instrument propensity score $P(Z = 1|\mathbf{X} = \cdot)$. Abadie considered estimation of the propensity score under a parametric model as well as by nonparametric power series methods. When $\mathbf{X}$ is high dimensional and the sample size is moderate, non-parametric propensity score estimation yields poorly behaved estimators of parametric specifications of $E(Y_d|D_1 > D_0, \mathbf{X} = \cdot)$, $d = 0, 1$ due to the curse of dimensionality.

3. New methods

In this section we describe estimation of the parameters indexing the following parsimonious models for $LATE(\mathbf{V})$ and $MLATE(\mathbf{V})$

$$LATE(\mathbf{v}) \in \mathcal{F}_1 = \{m_1(\mathbf{v}; \beta) : \beta \in B \subset R^p\} \tag{3}$$

and

$$MLATE(\mathbf{v}) \in \mathcal{F}_2 = \{m_2(\mathbf{v}; \beta) : \beta \in B \subset R^p\} \tag{4}$$

for specified functions $m_j(\cdot, \cdot)$ smooth in β, $j = 1, 2$. For inference under model $\mathcal{F}_1$ we assume that Y has unbounded support and for inference under model $\mathcal{F}_2$ we assume that Y has support equal to the non-negative real line or the non-negative integers.

For the special case in which $\mathbf{V}$ is equal to $\mathbf{X}$, Abadie also considered estimation of $LATE(\mathbf{X})$ under a parametric specification for the curve. However, his approach estimates $LATE(\mathbf{X})$ as the difference of the estimators of the means $E(Y_d|D_1 > D_0, \mathbf{X})$, $d = 0, 1$, under separate parametric models for each of them. We prefer estimating $LATE(\mathbf{X})$ under a model that parameterizes just this contrast rather than under separate models for each of the counterfactual means so as to reduce the opportunities of model misspecification.

For estimation of $LATE$ and $MLATE$, i.e. when $\mathbf{V}$ is null, the doubly robust estimators that we describe in this section, like the doubly robust estimators proposed by Tan (2006a) and Uysal (2011), are consistent under a correct parametric specification of the propensity score curve $P(Z = 1|\mathbf{X} = \cdot)$. Like the estimators of Tan and Uysal, our estimators remain consistent even under incorrect specification of the propensity score curve provided another set of curves are correctly parameterized. Tan's approach requires modeling $E(Y|Z = \cdot, D = \cdot, \mathbf{X} = \cdot,)$ and

$E(D|Z=\cdot,\mathbf{X}=\cdot)$, and Uysal's approach requires modeling $E(Y|Z=\cdot,\mathbf{X}=\cdot)$ and $E(D|Z=\cdot,\mathbf{X}=\cdot)$. Our approach, by contrast, requires modeling the conditional mean $E\left\{\underline{\varphi}(\mathbf{X})\,|\mathbf{V}=\cdot\right\}$ of a user-specified function $\underline{\varphi}(\mathbf{X})$ (if $\mathbf{V}\neq\mathbf{X}$) and the conditional expectation $E(H_j|Z=\cdot,\mathbf{X}=\cdot)$ ($j=1$ if inference is about $LATE$ and $j=2$ if is about $MLATE$), where

$$H_1 \equiv Y - D\times IV(\mathbf{V})$$

and

$$H_2 \equiv Y \times MIV(\mathbf{V})^{-D}.$$

The issue of which curves must be modeled in the doubly robust procedure, i.e. those in Tan, Uysal or our proposal, is inconsequential when $\mathbf{V}$ is null. However, it is an important issue if $\mathbf{V}$ is non-empty. As shown in the supplementary Web Appendix, when Y has unbounded support, $E\left\{\underline{\varphi}(\mathbf{X})\,|\mathbf{V}=\cdot\right\}$, $E(H_1|Z=\cdot,\mathbf{X}=\cdot)$ and $P(Z=1|\mathbf{X}=\cdot)$ are variation independent with $IV(\cdot)$ and when Y has support equal to $[0,\infty)$ or the non-negative integers, $E\left\{\underline{\varphi}(\mathbf{X})\,|\mathbf{V}=\cdot\right\}$, $E(H_2|Z=\cdot,\mathbf{X}=\cdot)$, and $P(Z=1|\mathbf{X}=\cdot)$ are variation independent with $MIV(\cdot)$. Therefore, our doubly robust procedure offers two genuine independent opportunities to produce consistent estimators of parametric specifications for $LATE(\cdot)$ or $MLATE(\cdot)$, as neither the models for $E\left\{\underline{\varphi}(\mathbf{X})\,|\mathbf{V}=\cdot\right\}$ and $E(H_1|Z=\cdot,\mathbf{X}=\cdot)$ nor the model for $P(Z=1|\mathbf{X}=\cdot)$ can conflict with parametric specifications of $IV(\mathbf{V}=\cdot)$ and, neither the models for $E\left\{\underline{\varphi}(\mathbf{X})\,|\mathbf{V}=\cdot\right\}$ and $E(H_2|Z=\cdot,\mathbf{X}=\cdot)$ nor the model for $P(Z=1|\mathbf{X}=\cdot)$ can conflict with parametric specifications of $MIV(\mathbf{V}=\cdot)$. Essentially, the variation independence of H_1 (H_2) with $IV(\cdot)$ ($MIV(\cdot)$) is a consequence of the fact that the restrictions imposed on the law of H_1 (H_2) by the IV assumptions do not depend on the functional form of $IV(\cdot)$ ($MIV(\cdot)$). In contrast, restrictions on $E(Y|Z=\cdot,\mathbf{X}=\cdot)$ and $E(D|Z=\cdot,\mathbf{X}=\cdot)$ or on $E(Y|Z=\cdot,D=\cdot,\mathbf{X}=\cdot)$ and $E(D|Z=\cdot,\mathbf{X}=\cdot)$ impose restrictions on $IV(\cdot)$ and therefore may conflict with parametric specifications for it.

On the other hand, it is worth noting that $E(Y|Z=\cdot,\mathbf{X}=\cdot)$, $E(D|Z=\cdot,\mathbf{X}=\cdot)$, and $E(Y|Z=\cdot,D=\cdot,\mathbf{X}=\cdot)$ are functionals of the observed data only. Although our proposed method has an important theoretical advantage over methods that rely on correct specifications of these conditional means, a practical advantage of the latter methods is that model building and model checking for these observed data quantities may be more straightforward and intuitive than for $E(H_j|Z=\cdot,\mathbf{X}=\cdot)$, $j=1,2$.

3.1. Estimation of $LATE(\cdot)$ and $MLATE(\cdot)$ under models for the propensity score or outcome regression

The following theorem gives two key expressions for the moment restrictions that are satisfied by the functionals $IV(\mathbf{V})$ and $MIV(\mathbf{V})$ on which our proposed estimators rely.

Theorem 1. For $j \in \{1,2\}$, if the denominators of $IV(\mathbf{V})$ and $MIV(\mathbf{V})$ are non-zero with probability 1, then

$$E\{E(H_j|Z=1,\mathbf{X}) - E(H_j|Z=0,\mathbf{X})|\mathbf{V}\} = 0 \text{ w.p.1} \tag{5}$$

and

$$E\left\{(-1)^{1-Z} p(Z|\mathbf{X})^{-1} H_j|\mathbf{V}\right\} = 0 \text{ w.p.1}, \tag{6}$$

where $p(Z|\mathbf{X}) \equiv P(Z=1|\mathbf{X})^Z \{1-P(Z=1|\mathbf{X})\}^{1-Z}$.

Proof: Equation (5) with $j=1$ follows by algebra from the definition (1) and with $j=2$ it follows from the definition (2). Specifically, to arrive at (5) from (1) when $j=1$ note that the difference between the numerator in the right hand side of (1) and the product of $IV(\mathbf{v})$ with the denominator in the right hand side of (1) is the same as the left hand side of (5). Likewise, to arrive at (5) from (2) when $j=2$ note that the sum of the denominator in the right hand side of (2) with the product of the numerator in the right hand side of (2) times $MIV(\mathbf{v})^{-1}$ is the same as the left hand side of (5). Equation (6) is equivalent to equation (5) because
$E\left\{(-1)^{1-Z} p(Z|\mathbf{X})^{-1} H_j|\mathbf{V}\right\} = E\left[\left\{Zp(Z|\mathbf{X})^{-1} - (1-Z)p(Z|\mathbf{X})^{-1}\right\} H_j|\mathbf{V}\right]$,
$E\left\{Zp(Z|\mathbf{X})^{-1} H_j|\mathbf{V}\right\} = E\{E(H_j|Z=1,\mathbf{X})|\mathbf{V}\}$ and
$E\left\{(1-Z)p(Z|\mathbf{X})^{-1} H_j|\mathbf{V}\right\} = E\{E(H_j|Z=0,\mathbf{X})|\mathbf{V}\}$.

Theorem 1 suggests that well behaved estimators of β can be obtained under parametric specifications of either $P(Z=1|\mathbf{X})$ or $E(H_j|Z,\mathbf{X})$ where throughout we assume $j=1$ if β indexes the parametric specification (3) for $LATE(\mathbf{V})$ and $j=2$ if β indexes the specification (4) for $MLATE(\mathbf{V})$. We now describe such estimators.

Define

$$H_1(\beta) \equiv Y - Dm_1(\mathbf{V};\beta)$$

and

$$H_2(\beta) \equiv Ym_2(\mathbf{V};\beta)^{-D}$$

where $m_1(\mathbf{V};\beta)$ and $m_2(\mathbf{V};\beta)$ are the parametric specifications for $LATE(\mathbf{v})$ defined in (3) and for $MLATE(\mathbf{v})$ defined in (4) respectively. Throughout we let β_0 denote the true the value of β under the given specification (3) or (4).

A consistent and asymptotically normal (CAN) estimator $\widehat{\beta}_{\text{ipw}}$ of β_0 under a parametric class for the instrument probabilities

$$P(Z=1|\mathbf{X}=\mathbf{x}) \equiv \pi(\mathbf{x}) \in \mathcal{P} = \left\{\pi(\mathbf{x};\alpha) : \alpha \in \mathbb{A} \subset R^d\right\} \tag{7}$$

where $\pi(\cdot;\cdot)$ is a specified function smooth in α and $\mathbb{A}$ is a specified subset of R^d, is computed as the solution of

$$E_n\left\{q(\mathbf{V};\beta)(-1)^{1-Z}p(Z|\mathbf{X};\widehat{\alpha})^{-1}H_j(\beta)\right\}=0 \tag{8}$$

where $p(Z|\mathbf{X};\alpha)\equiv\pi(\mathbf{X};\alpha)^Z\{1-\pi(\mathbf{X};\alpha)\}^{1-Z}$, $q(\mathbf{V};\beta)$ is a user specified $p\times 1$ vector valued function (for example $q(\mathbf{V};\beta)=\partial m_j(\mathbf{V};\beta)/\partial\beta$), and

$$\widehat{\alpha}=\arg\max_{\alpha}E_n\left(\log\left[\pi(\mathbf{X};\alpha)^Z\{1-\pi(\mathbf{X};\alpha)\}^{1-Z}\right]\right) \tag{9}$$

is the maximum likelihood estimator of α. Throughout $E_n(\cdot)$ stands for the empirical mean operator. Identity (6) implies that under the IV assumptions, under the parametric specification (3), and with $j=1$ in display (5), $\sqrt{n}\left(\widehat{\beta}_{\text{ipw}}-\beta_0\right)$ converges in law to a mean zero normal distribution when (7) and regularity conditions hold and, in addition, for some σ and $z=0,1$, $P(Z=z|\mathbf{X};\alpha)>\sigma>0$. The same holds under the parametric specification (4) and with $j=2$ in display (5).

Alternatively, one can compute a CAN estimator β_0 under a parametric class for $E(H_j\mid Z,\mathbf{X})$ that respects the constraint (5). To aid the specification of such parametric class, we re-express the constraint (5) as the condition that for some $r(\mathbf{X})$,

$$E(H_j\mid Z=1,\mathbf{X})-E(H_j\mid Z=0,\mathbf{X})=r(\mathbf{X})-E\{r(\mathbf{X})|\mathbf{V}\}.$$

When $\mathbf{V}$ is not equal to $\mathbf{X}$ we derive a flexible parametric specification for $E(H_j\mid Z,\mathbf{X})$ that respects the constraint (5) from the following three specifications:

(1) a linear parametric specification for $r(\mathbf{X})$

$$r(\mathbf{X})\in\mathcal{R}\ =\left\{\rho^T\underline{\varphi}(\mathbf{X}):\rho\in R^K\right\} \tag{10}$$

where $\underline{\varphi}(\mathbf{X})\equiv(\varphi_1(\mathbf{X}),...,\varphi_K(\mathbf{X}))^T$ and φ_s, $s\in\{1,...,K\}$, are user-specified real valued functions,

(2) a linear model for the mean of $\underline{\varphi}(\mathbf{X})$ given $\mathbf{V}$,

$$E\left\{\underline{\varphi}(\mathbf{X})|\mathbf{V}\right\}\in\mathcal{M}\ =\left\{\underline{\phi}(\mathbf{V};\gamma):\gamma\in\Gamma\right\} \tag{11}$$

where $\underline{\phi}(\mathbf{V};\gamma)\equiv(\phi_1(\mathbf{V};\gamma),...,\phi_K(\mathbf{V};\gamma))^T$, Γ is a subset of a Euclidean space and $\phi_k, k\in\{1,...,K\}$, are user-specified real valued functions (when $\mathbf{V}$ is null we set $\underline{\phi}(\mathbf{V};\gamma)=\gamma$ thus leaving $\mathcal{M}$ unrestricted),

(3) a parametric specification for $E(H_j\mid Z=0,\mathbf{X})$, i.e.

$$E(H_j\mid Z=0,\mathbf{X})\in\mathcal{K}\ =\{k(\mathbf{X};\nu):\nu\in\Upsilon\} \tag{12}$$

where $k(\cdot;\cdot)$ is a specified function smooth in ν and Υ is a subset of a Euclidean space.

Specifications (10) , (11) , (12) imply the following model respects the constraint (5) ,

$$E\left(H_j \mid Z=z, \mathbf{X}=\mathbf{x}\right) \in \mathcal{H}=\left\{h\left(z, \mathbf{x}; \eta, \gamma\right): \eta \in R^K \times \Upsilon\ , \gamma \in \Gamma\right\} \quad (13)$$

where $\eta \equiv (\rho, \nu)$ and $h\left(z, \mathbf{x}; \eta, \gamma\right) \equiv k\left(\mathbf{x}; \nu\right)+\rho^T\left\{\underline{\varphi}\left(\mathbf{x}\right)-\underline{\phi}\left(\mathbf{v}; \gamma\right)\right\} z$.

When $\mathbf{V}=\mathbf{X}$, we ignore (11) and replace the specification (13) with

$$E\left(H_j \mid Z=z, \mathbf{X}=\mathbf{x}\right) \in \mathcal{H}\ =\left\{h\left(\mathbf{X}; \eta\right): \eta \in \Upsilon\right\} \quad (14)$$

where $h\left(\cdot;\cdot\right)$ is a specified function smooth in η and Υ is a subset of a Euclidean space. This specification also respects the constraint (5) because when $\mathbf{V}=\mathbf{X}$ this constraint is the same as the condition that $E\left(H_j \mid Z, \mathbf{X}=\mathbf{x}\right)$ does not depend on Z.

An estimator $\widehat{\beta}_{\text{reg}}$ consistent and asymptotically normal (CAN) for β_0 under specifications (11) and (13) when $\mathbf{V} \neq \mathbf{X}$ or specification (14) when $\mathbf{V}=\mathbf{X}$ can be computed as the first component of the vector $\left(\widehat{\beta}_{\text{reg}}, \widehat{\eta}\right)$ solving

$$E_n\left\{l\left(Z, \mathbf{X}; \beta, \eta, \widehat{\gamma}\right) \varepsilon_j\left(\beta, \eta, \widehat{\gamma}\right)\right\}=0 \quad (15)$$

where $l\left(\cdot,\cdot;\cdot,\cdot,\cdot\right)$ is a user-specified vector-valued function of the same dimension as (β, η) ,

$$\varepsilon_j\left(\beta, \eta, \gamma\right) \equiv H_j\left(\beta\right)-h\left(Z, \mathbf{X}; \eta, \gamma\right)$$

and $\widehat{\gamma}$ solves $E_n\left[\left\{\partial \underline{\phi}\left(\mathbf{V}; \gamma\right)^T / \partial \gamma\right\}\left\{\underline{\varphi}\left(\mathbf{X}\right)-\underline{\phi}\left(\mathbf{V}; \gamma\right)\right\}\right]=0$ if $\mathbf{V} \neq \mathbf{X}$, and $\varepsilon_j\left(\beta, \eta, \gamma\right) \equiv H_j\left(\beta\right)-h\left(\mathbf{X}; \eta\right)$ if $\mathbf{V}=\mathbf{X}$. One practical choice of $l\left(Z, \mathbf{X}; \beta, \eta, \widehat{\gamma}\right)$ is

$$l\left(Z, \mathbf{X}; \beta, \eta, \gamma\right)=\left[\begin{array}{c} l_\eta\left(Z, \mathbf{X}; \eta, \gamma\right) \\ l_\beta\left(Z, \mathbf{X}; \beta\right) \end{array}\right]=\left[\begin{array}{c} \partial h\left(Z, \mathbf{X}; \eta, \gamma\right) / \partial \eta \\ Z \times \partial m\left(\mathbf{V}; \beta\right) / \partial \beta \end{array}\right]. \quad (16)$$

Under (11) and (13) when $\mathbf{V} \neq \mathbf{X}$ or (14) when $\mathbf{V}=\mathbf{X}$, the IV assumptions and the parametric specification (3) if $j=1$ or (4) if $j=2$, $E\left\{\varepsilon_j\left(\beta_0, \eta_0, \gamma_0\right)|Z, \mathbf{X}\right\}=0$ where (η_0, γ_0) are the true values of (η, γ) , so $\sqrt{n}\left(\widehat{\beta}_{\text{reg}}-\beta_0\right)$ converges in law to a mean zero normal distribution provided standard regularity conditions for convergence of $M-$ estimators hold.

Selection of the parametric class for $E\left(H_j|Z, \mathbf{X}\right)$ can be aided with the following $\alpha-$level score type test of the null hypothesis $\mathbb{H}_0: \eta_2=0$ where $\eta=\left(\eta_1^T, \eta_2^T\right)^T$ and η_2 is of dimension, say, d_2. Let $R_n=E_n\left[\left\{\partial h\left(\widetilde{\beta}_{\text{reg}}, \widetilde{\eta}_1, \eta_2, \widehat{\gamma}\right) / \partial \eta_2\Big|_{\eta_2=0}\right\} \varepsilon_j\left(\widetilde{\beta}_{\text{reg}}, \widetilde{\eta}_1, 0, \widehat{\gamma}\right)\right]$ where $\left(\widetilde{\beta}_{\text{reg}}, \widetilde{\eta}_1\right)$ solves $E_n\left\{\left[\partial h\left(Z, \mathbf{X}; \eta_1, 0, \widehat{\gamma}\right) / \partial \eta_1^T, l_\beta\left(Z, \mathbf{X}; \beta\right)^T\right]^T \varepsilon_j\left(\beta, \eta_1, 0, \widehat{\gamma}\right)\right\}=0$.

Under $\mathbb{H}_0$, $\sqrt{n}R_n$ converges in law to a mean zero d_2-variate normal distribution with variance covariance matrix, say, J. Thus, if $\widehat{J}$ is a consistent estimator of J, a test that rejects $\mathbb{H}_0$ when $R_n^T \widehat{J} R_n > \chi_{1-\alpha,d_2}$ where $\chi_{1-\alpha,d_2}$ is the $1-\alpha$ quantile of a chi-squared distribution with d_2 degrees of freedom is an asymptotic $\alpha-$ level test of $\mathbb{H}_0$. A consistent variance estimator $\widehat{J}$ can be derived from standard Taylor expansion arguments for M estimators (Stefanski and Boos, 2002).

3.2. Doubly robust estimation of $LATE(\cdot)$ and $MLATE(\cdot)$

In this section we derive a doubly robust estimator $\widehat{\beta}_{\mathrm{dr}}$ of β which satisfies that $\sqrt{n}\left(\widehat{\beta}_{\mathrm{dr}} - \beta_0\right)$ converges to a mean zero normal distribution under the IV assumptions and regularity conditions provided one of the following two conditions (i) or (ii) holds, even if both don't hold simultaneously:

(i) specifications (11) and(13) are correct when $\mathbf{V} \neq \mathbf{X}$, or specification (14) is correct when $\mathbf{V} = \mathbf{X}$,

(ii) specification (7) is correct.

The estimator $\widehat{\beta}_{\mathrm{dr}}$ solves the estimating equations

$$E_n\left[q(\mathbf{V};\beta)(-1)^{1-Z} p(Z|\mathbf{X};\widehat{\alpha})^{-1}\{H_j(\beta) - a(\mathbf{X};\widehat{\alpha},\widehat{\eta}(\beta),\widehat{\gamma})\}\right] = 0 \tag{17}$$

where, for each fixed β, $\widehat{\eta}(\beta)$ solves $E_n\{l_\eta(Z,\mathbf{X};\beta,\eta,\widehat{\gamma})\varepsilon_j(\beta,\eta,\widehat{\gamma})\} = 0$ with l_η defined as in (16) and

$$a(\mathbf{X};\alpha,\eta,\gamma) \equiv \{1-\pi(\mathbf{X};\alpha)\}h(1,\mathbf{X};\eta,\gamma) + \pi(\mathbf{X};\alpha)h(0,\mathbf{X};\eta,\gamma)$$

if $\mathbf{V} \neq \mathbf{X}$ or $a(\mathbf{X};\alpha,\eta,\gamma) \equiv h(\mathbf{X};\eta)$ if $\mathbf{V} = \mathbf{X}$.

The estimator $\widehat{\beta}_{\mathrm{dr}}$ is consistent for β_0 when (ii) holds because $E\left\{q(\mathbf{V};\beta)(-1)^{1-Z} p(Z|\mathbf{X};\alpha_0)^{-1} a(\mathbf{X};\alpha,\eta,\gamma)\right\} = 0$ for all β since $E\left\{(-1)^{1-Z} p(Z|\mathbf{X};\alpha_0)|\mathbf{X}\right\} = 0$.

On the other hand, consistency when (i) holds can be seen after re-expressing equation (17) as

$$E_n\left\{q(\mathbf{V};\beta)\frac{(-1)^{1-Z}}{p(Z|\mathbf{X};\widehat{\alpha})}\varepsilon_j(\beta,\widehat{\eta}(\beta),\widehat{\gamma})\right\} +$$
$$E_n\left[q(\mathbf{V};\beta)\{h(1,\mathbf{X};\widehat{\eta}(\beta),\widehat{\gamma}) - h(0,\mathbf{X};\widehat{\eta}(\beta),\widehat{\gamma})\}\right] = 0$$

and noting that, by virtue of equality (5) of Theorem 1, $E\left[q(\mathbf{V};\beta)\{h(1,\mathbf{X};\eta_0,\gamma_0) - h(0,\mathbf{X};\eta_0,\gamma_0)\}\right] = 0$ and by $E\{\varepsilon_j(\beta,\eta_0,\gamma_0)|Z,\mathbf{X}\} = 0$, $E\left[b(Z,\mathbf{X})\varepsilon_j(\beta,\eta_0,\gamma_0)\right] = 0$ for all $b(Z,\mathbf{X})$ and, in particular, for $b(Z,\mathbf{X}) = q(\mathbf{V};\beta)(-1)^{1-Z} p(Z|\mathbf{X};\alpha)^{-1}$ with arbitrary α.

The convergence of $\sqrt{n}\left(\widehat{\beta}_{\mathrm{dr}}-\beta_0\right)$ to a normal distribution follows after noticing that $\left(\widehat{\beta}_{\mathrm{dr}},\widehat{\eta},\widehat{\gamma},\widehat{\alpha}\right)$ where $\widehat{\eta}\equiv\widehat{\eta}\left(\widehat{\beta}_{\mathrm{dr}}\right)$ is an M-estimator, i.e. it solves a joint system of estimating equation. The accuracy of this asymptotic result in finite samples hinges on the strength of the instrument Z, i.e. on how close $\Delta(\mathbf{V})=E\left\{E\left(D|Z=1,\mathbf{X}\right)-E\left(D|Z=0,\mathbf{X}\right)|\mathbf{V}\right\}$ is to 0. Theoretical results exploring the asymptotic distribution of $\widehat{\beta}_{\mathrm{dr}}$ as $\Delta(\mathbf{V})$ shrinks to zero at different rates with sample size, similar to those in the conventional IV literature, should be explored but are beyond the scope of this paper.

The asymptotic variance of $\widehat{\beta}_{\mathrm{dr}}$ can be consistently estimated with the standard empirical sandwich variance estimator (Stefanski and Boos, 2002) or with the nonparametric bootstrap (Gill, 1989).

In the special case of estimation of $\beta_0\equiv LATE$, i.e. when $\mathbf{V}$ is null, we have that $H_1(\beta)=Y-\beta D$ and our doubly robust estimator is similar to that in Tan (2006a) and that in Uysal (2011), except that these authors replace $h\left(Z,\mathbf{X};\widehat{\eta}(\beta),\widehat{\gamma}\right)$ with $\widehat{E}\left(Y|Z,\mathbf{X}\right)-\beta\widehat{E}\left(D|Z,\mathbf{X}\right)$. Tan computes estimators $\widehat{E}\left(Y|Z,\mathbf{X}\right)$ and $\widehat{E}\left(D|Z,\mathbf{X}\right)$ under parametric models for $E\left(Y|D=d,Z=z,\mathbf{X}=\cdot\right)$ and $E\left(D|Z=z,\mathbf{X}=\cdot\right),\ d,z=0,1$ whereas Uysal (2011) under parametric models for $E\left(Y|Z=z,\mathbf{X}=\cdot\right)$ and $E\left(D|Z=z,\mathbf{X}=\cdot\right),\ z=0,1$.

3.3. Local efficiency under correct parametric specification of the propensity score model

In addition to $\widehat{\beta}_{\mathrm{ipw}}$ and $\widehat{\beta}_{\mathrm{dr}}$, there are other consistent and asymptotically normal estimators of β_0 under the propensity score specification (7) and the IV assumptions. Specifically, given a user-specified $p\times 1$ function $s(\mathbf{x};\beta)$, consider the estimator $\widehat{\beta}_{\mathrm{s}}$ solving

$$E_n\left[q(\mathbf{V};\beta)(-1)^{1-Z}p\left(Z|\mathbf{X};\widehat{\alpha}\right)^{-1}\left\{H_j(\beta)-s(\mathbf{X})\right\}\right]=0.$$

Because $E\left\{q(\mathbf{V};\beta)(-1)^{1-Z}p\left(Z|\mathbf{X}\right)^{-1}s(\mathbf{X})\right\}=0$ it follows that under regularity conditions, when (7) holds,$\sqrt{n}\left(\widehat{\beta}_{\mathrm{s}}-\beta_0\right)$ converges to a mean zero normal distribution with variance $\Sigma_{\mathrm{q,s}}$, where $\Sigma_{\mathrm{q,s}}$ depends on $q(\cdot)$ and on $s(\cdot)$. Invoking the theory of inverse probability weighted estimation in Robins and Rotnitzky (1992), in the supplementary Web Appendix we show that for each fixed $q(\cdot)$ the optimal choice $s_{\mathrm{opt,j}}(\mathbf{X})$, in the sense that $\Sigma_{\mathrm{q,s}}-\Sigma_{\mathrm{q},s_{\mathrm{opt,j}}}\geq 0$ (i.e. semipositive definite), is given by

$$s_{\mathrm{opt,j}}(\mathbf{X})=\{1-\pi(\mathbf{X})\}E\left(H_j\mid Z=1,\mathbf{X}\right)+\pi(\mathbf{X})E\left(H_j\mid Z=0,\mathbf{X}\right).$$

In the supplementary Web Appendix we also show that when the specifications (11), (13) and (7) hold if $\mathbf{V}$ is not equal to $\mathbf{X}$ or when the

specifications (14) and (7) hold if $\mathbf{V} = \mathbf{X}$, the limiting distribution of $\sqrt{n}\left(\widehat{\beta}_{\mathrm{dr}} - \beta_0\right)$ has variance precisely equal to the bound $\Sigma_{\mathrm{q,s}_{\mathrm{opt,j}}}$. The estimator $\widehat{\beta}_{\mathrm{dr}}$, however, may have asymptotic variance even larger than that of $\widehat{\beta}_{\mathrm{ipw}}$ if specification (11) and/or (13) is incorrect when $\mathbf{V} \neq \mathbf{X}$ or if specification (14) is incorrect when $\mathbf{V} = \mathbf{X}$. Using ideas similar to those in Tan (2006b, 2010) we can construct another doubly robust estimator $\widetilde{\beta}_{\mathrm{dr}}$ that remedies this flaw. The estimator $\widetilde{\beta}_{\mathrm{dr}}$ is computed by solving

$$E_n\left[(-1)^{1-Z} p\left(Z|\mathbf{X};\widehat{\alpha}\right)^{-1}\left\{H_j\left(\beta\right) Id - a\left(\mathbf{X};\widehat{\alpha},\widehat{\eta}\left(\beta\right),\widehat{\gamma}\right)\widehat{C}\left(\beta\right)^{\mathrm{T}}\right\} q\left(\mathbf{V};\beta\right)\right] = 0, \tag{18}$$

where Id is the $p \times p$ identity matrix and $\widehat{C}\left(\beta\right)$ is the $p \times p$ matrix formed by the first p columns of the $p \times (p+d)$ matrix

$$E_n\left\{(-1)^{1-Z} p\left(Z|\mathbf{X};\widehat{\alpha}\right)^{-1} H_j\left(\beta\right) q\left(\mathbf{V};\beta\right) \times K\left(\beta\right)\right\} \times$$
$$E_n\left\{\left[\begin{array}{c} q\left(\mathbf{V};\beta\right)(-1)^{1-Z} p\left(Z|\mathbf{X};\widehat{\alpha}\right)^{-1} h\left(Z,\mathbf{X};\widehat{\eta}\left(\beta\right),\widehat{\gamma}\right) \\ \partial \log p\left(Z|\mathbf{X};\alpha\right)/\partial\alpha\big|_{\alpha=\widehat{\alpha}} \end{array}\right] \times K\left(\beta\right)\right\}^{-1}$$

with

$$K\left(\beta\right) = \left\{q\left(\mathbf{V};\beta\right)^{\mathrm{T}}(-1)^{1-Z} p\left(Z|\mathbf{X};\widehat{\alpha}\right)^{-1} a\left(\mathbf{X};\widehat{\alpha},\widehat{\eta}\left(\beta\right),\widehat{\gamma}\right),\ \ \partial \log p\left(Z|\mathbf{X};\alpha\right)/\partial\alpha^{\mathrm{T}}\big|_{\alpha=\widehat{\alpha}}\right\}.$$

Like $\widehat{\beta}_{\mathrm{dr}}$, the estimator $\widetilde{\beta}_{\mathrm{dr}}$ is doubly robust and has asymptotic variance equal to $\Sigma_{\mathrm{q,s}_{\mathrm{opt,j}}}$ when specifications (11), (13) and (7) are correct ((14) and (7) are correct if $\mathbf{V} = \mathbf{X}$), but unlike $\widehat{\beta}_{\mathrm{dr}}$, it is guaranteed to be the most efficient estimator, asymptotically, among the class of estimators solving equations of the form (18) with $\widehat{C}\left(\beta\right)$ replaced by an arbitrary $p \times p$ constant matrix C. In particular, letting $C = 0$ we conclude that under model (7), $\widetilde{\beta}_{\mathrm{dr}}$ is never less efficient asymptotically than $\widehat{\beta}_{\mathrm{ipw}}$. See the supplementary Web Appendix for a sketch of the proof of the asymptotic properties of $\widetilde{\beta}_{\mathrm{dr}}$.

A further result, derived in the supplementary Web Appendix, establishes that for $j \in \{1,2\}$ the optimal function $q_{\mathrm{opt,j}}\left(\cdot\right)$, in the sense that $\Sigma_{\mathrm{q,s}_{\mathrm{opt,j}}} - \Sigma_{\mathrm{q}_{\mathrm{opt,j}},\mathrm{s}_{\mathrm{opt,j}}} \geq 0$ for any $q\left(\cdot\right)$, is

$$q_{\mathrm{opt,j}}\left(\mathbf{V};\beta\right) = \left\{\partial m_j\left(\mathbf{V};\beta\right)/\partial\beta\right\} c_j\left(\mathbf{V};\beta\right)$$

where

$$c_j\left(\mathbf{V};\beta\right) = -m_j\left(\mathbf{V};\beta\right)^{2(1-j)} E\left\{\left(-1\right)^{1-Z} p\left(Z|\mathbf{X}\right)^{-1} DY^{j-1}\Big|\,\mathbf{V}\right\} \times$$
$$E\left[p\left(Z|\mathbf{X}\right)^{-2}\left\{H_j - s_{\mathrm{opt,}j}\left(\mathbf{X}\right)\right\}^2 \mid \mathbf{V}\right]^{-1}.$$

The optimal function $q_{\mathrm{opt,j}}\left(\cdot\right)$ depends on the unknown observed data distribution and hence it is not available for data analysis. However, we

can estimate it under working parametric specifications for its unknown constituents,

$$E\left\{(-1)^{1-Z} p(Z|\mathbf{X})^{-1} DY^{j-1} \middle| \mathbf{V}\right\} \in \mathcal{E}_j = \{e_j(\mathbf{V};\delta) : \delta \in \Delta\} \quad (19)$$

and

$$E\left[p(Z|\mathbf{X})^{-2} \{H_j - s_{\text{opt},j}(\mathbf{X})\}^2 \mid \mathbf{V}\right] \in \mathcal{T}_j = \{t_j(\mathbf{V};\omega) : \omega \in \Omega\} \quad (20)$$

where $e_j(\cdot;\cdot)$ and $t_j(\cdot)$ are smooth functions and Δ and Ω are included in Euclidean spaces. To do so we estimate δ and ω with the weighted least squares estimators $\widehat{\delta}$ and $\widehat{\omega}$ by regressing $(-1)^{1-Z} p(Z|\mathbf{X};\widehat{\alpha})^{-1} DY^{j-1}$ and $p(Z|\mathbf{X};\widehat{\alpha})^{-2} \left\{H_j\left(\widehat{\beta}_{\text{dr}}\right) - a\left(\mathbf{X};\widehat{\alpha},\widehat{\eta}\left(\widehat{\beta}_{\text{dr}}\right),\widehat{\gamma}\right)\right\}^2$ on $\mathbf{V}$ under models (19) and (20) respectively, where $\widehat{\beta}_{\text{dr}}$ is a preliminary doubly robust estimator of β computed using an arbitrary $q(\mathbf{V};\beta)$. We then estimate $q_{\text{opt,j}}(\mathbf{V};\beta)$ with $\widehat{q}_{\text{opt,j}}(\mathbf{V};\beta) \equiv -\{\partial m(\mathbf{V};\beta)/\partial\beta\} \times m_j(\mathbf{V};\beta)^{2(1-j)} e_j\left(\mathbf{V};\widehat{\delta}\right) t_j(\mathbf{V};\widehat{\omega})^{-1}$.

When specification (7) is correct and $P(Z=z|\mathbf{X}) > \sigma > 0$ for $z = 0$ or 1, the estimators $\widehat{\beta}_{\text{dr}}$ and $\widetilde{\beta}_{\text{dr}}$ that use $\widehat{q}_{\text{opt,j}}(\mathbf{V};\beta)$ for $q(\mathbf{V};\beta)$ and the estimator $\widetilde{\beta}_{\text{C}}$ that solves (18) with $\widehat{C}(\beta)$ replaced by an arbitrary $p \times p$ constant matrix C and with $\widehat{q}_{\text{opt,j}}(\mathbf{V};\beta)$ instead of $q(\mathbf{V};\beta)$ satisfy under regularity conditions:

(a) $\sqrt{n}\left\{\widehat{\beta}_{\text{dr}} - \beta_0\right\}$, $\sqrt{n}\left\{\widetilde{\beta}_{\text{dr}} - \beta_0\right\}$ and $\sqrt{n}\left\{\widetilde{\beta}_{\text{C}} - \beta_0\right\}$ converge to mean zero normal distributions with variances $\sum_{\text{dr}}$, $\sum_{\text{better.dr}}$ and $\sum_{\text{C}}$ respectively. Furthermore, $\sum_{\text{better.dr}} - \sum_{\text{C}} \leq 0$ and $\sum_{\text{better.dr}} - \sum_{\text{dr}} \leq 0$.

(b) If, additionally, the specifications (11) and (13) are correct when $\mathbf{V} \neq \mathbf{X}$, or the specification (14) is correct when $\mathbf{V} = \mathbf{X}$, then $\sum_{\text{dr}} = \sum_{\text{better.dr}} = \Sigma_{q_{\text{opt,j}},s_{\text{opt,j}}}$.

3.4. Estimation of least squares approximations under incorrect specifications of local average treatment effect curves.

A slight modification of the procedure for computing $\widehat{\beta}_{\text{dr}}$ and $\widetilde{\beta}_{\text{dr}}$ yields estimators that are doubly robust for least squares approximations of the true local average treatment effect curves when the parametric specifications for these curves are incorrect.

Given a real valued function $w(\mathbf{v})$, the w-weighted least squares approximation of the $LATE(\cdot)$ curve is

$$\beta_{w,0} \equiv \arg\min_{\beta} E\left[w(\mathbf{V})\{LATE(\mathbf{V}) - m_1(\mathbf{V};\beta)\}^2 | D_1 > D_0\right]. \quad (21)$$

In the supplementary Web Appendix we show that under the IV-conditions, $\beta_{w,0}$ satisfies

$$E\left\{q_w(\mathbf{V})(-1)^{1-Z}p(Z|\mathbf{X})^{-1}H_1(\beta_{w,0})\right\}=0 \tag{22}$$

where $q_w(\mathbf{V}) \equiv w(\mathbf{V})\,\partial m_1(\mathbf{V};\beta)/\partial\beta|_{\beta=\beta_{w,0}}$. Arguing as in section 3.2, we conclude that when condition (ii) of section 3.2 holds (i.e.when the propensity score specification (7) is correct), the estimators $\widehat{\beta}_{\mathrm{dr}}$ and $\widetilde{\beta}_{\mathrm{dr}}$ that use $q(\mathbf{V};\beta)$ equal to $q_w(\mathbf{V};\beta) \equiv w(\mathbf{V})\,\partial m_1(\mathbf{V};\beta)/\partial\beta$ converge in probability to $\beta_{w,0}$ even if the specification (3) is incorrect.

On the other hand, unfortunately, $\widehat{\beta}_{\mathrm{dr}}$ and $\widetilde{\beta}_{\mathrm{dr}}$ need not converge to $\beta_{w,0}$ for any w when the propensity score model is incorrect even if condition (i) of section 3.2 holds. This happens essentially because (22) is equivalent to

$$E\left[q_w(\mathbf{V})\left[E\left\{H_1(\beta_{w,0})|Z=1,\mathbf{X}\right\}-E\left\{H_1(\beta_{w,0})|Z=0,\mathbf{X}\right\}\right]\right]=0, \tag{23}$$

which involves $E\{H_1(\beta_{w,0})|Z,\mathbf{X}\}$ but not $E(H_1|Z,\mathbf{X})$. Nevertheless, the equality (23) suggests that CAN estimators of $\beta_{w,0}$ under parametric models for $E\{H_1(\beta_{w,0})|Z,\mathbf{X}\}$ should exist. However, some care must be taken in formulating such models. For instance, one cannot postulate that $E\{H_1(\beta_{w,0})|Z,\mathbf{X}\} \in \mathcal{H}$ where $\mathcal{H}$ is defined in (13) with $j=1$ since this specification is necessarily wrong if the model (11) is correct. This happens because $\mathcal{H}$ respects the constraint (5) but $E\{H_1(\beta_{w,0})|Z,\mathbf{X}\}$ does not, since of all random variables of the form $H_1(m)=Y-m(\mathbf{V})D$ for any $m(\mathbf{V})$, only $H_1=Y-IV(\mathbf{V})D$ satisfies the constraint (5) as this constraint identifies the $IV(\cdot)$ curve.

A slight modification to the class $\mathcal{H}$ yields a new class that respects the constraint (23) but not necessarily the stronger constraint $E[[E\{H_1(\beta_{w,0})|Z=1,\mathbf{X}\}-E\{H_1(\beta_{w,0})|Z=0,\mathbf{X}\}]|\mathbf{V}]=0$ and thus gives the opportunity of formulating a correctly specified model for $E\{H_1(\beta_{w,0})|Z,\mathbf{X}\}$. Specifically, the parametric specification

$$\begin{aligned}&E\{H_1(\beta_{w,0})\mid Z=z,\mathbf{X}=\mathbf{x}\}\in\mathcal{H}_w=\\&\left\{k(\mathbf{x};\nu)+\lambda^T\left\{\underline{\varphi}(\mathbf{x})-\theta q_w(\mathbf{v})\right\}z:\lambda\in R^K,\nu\in\Upsilon\right\}\end{aligned} \tag{24}$$

where $\underline{\varphi}(\cdot)$ and $k(\cdot;\cdot)$ are user-chosen functions as defined in section 3.1 and

$$\theta=E\left\{\underline{\varphi}(\mathbf{X})q_w(\mathbf{V})^T\right\}E\left\{q_w(\mathbf{V})q_w(\mathbf{V})^T\right\}^{-1} \tag{25}$$

necessarily respects the constraint (23) but not the aforementioned stronger constraint.

A modification in the computation of $\widehat{\beta}_{\mathrm{dr}}$ yields a new estimator $\widehat{\widehat{\beta}}_{\mathrm{dr}}$, described below, that satisfies for a given, user-specified, weight function $w(\cdot)$ the following two conditions:

a $\sqrt{n}\left(\widehat{\widehat{\beta}}_{\rm dr}-\beta_0\right)$ converges to a Normal distribution if the parametric specification (3) for $LATE(\cdot)$ is correct and either condition (i) or condition (ii) of section 3.2 hold, and

b $\sqrt{n}\left(\widehat{\widehat{\beta}}_{\rm dr}-\beta_{w,0}\right)$ converges to a Normal distribution if the parametric specification (3) for $LATE(\cdot)$ is incorrect but either condition (ii) of section 3.2 or the parametric specification (24) hold.

Consider first the case $\mathbf{V}\neq\mathbf{X}$. The estimator $\widehat{\widehat{\beta}}_{\rm dr}$ solves equation (17) with $q_w(\mathbf{V};\beta)$ instead of $q(\mathbf{V};\beta)$, and with $a(\mathbf{X};\widehat{\alpha},\widehat{\eta}(\beta),\widehat{\gamma})$ replaced by

$$b(\mathbf{X};\alpha,\eta,\gamma,\theta)\equiv\{1-\pi(\mathbf{X};\alpha)\}\,h_w(1,\mathbf{X};\beta,\eta,\gamma,\theta)+\pi(\mathbf{X};\alpha)\,h_w(0,\mathbf{X};\beta,\eta,\gamma,\theta),$$

where $\eta=(\nu,\rho,\lambda)$,

$$h_w(z,\mathbf{x};\beta,\eta,\gamma,\theta)\equiv k(\mathbf{x};\nu)+\rho^T\left\{\underline{\varphi}(\mathbf{x})-\underline{\phi}(\mathbf{v};\gamma)\right\}z+\lambda^T\left\{\underline{\varphi}(\mathbf{x})-\theta q_w(\mathbf{v};\beta)\right\}z, \tag{26}$$

$\widehat{\eta}(\beta)$ solves $E_n\left[\left\{\partial h_w\left(Z,\mathbf{X};\beta,\eta,\widehat{\gamma},\widehat{\theta}(\beta)\right)/\partial\eta\right\}\varepsilon_w\left(\beta,\eta,\widehat{\gamma},\widehat{\theta}(\beta)\right)\right]=0$ with

$$\varepsilon_w(\beta,\eta,\gamma,\theta)\equiv H_1(\beta)-h_w(z,\mathbf{x};\beta,\eta,\gamma,\theta),$$

$\widehat{\gamma}$ solves $E_n\left[\left\{\partial\underline{\phi}(\mathbf{V};\gamma)^T/\partial\gamma\right\}\left\{\underline{\varphi}(\mathbf{X})-\underline{\phi}(\mathbf{V};\gamma)\right\}\right]=0$ and $\widehat{\theta}(\beta)\equiv E_n\left\{\underline{\varphi}(\mathbf{X})\,q_w(\mathbf{V})^T\right\}E_n\left\{q_w(\mathbf{V})\,q_w(\mathbf{V})^T\right\}^{-1}$. When $\mathbf{V}=\mathbf{X}$, $\widehat{\widehat{\beta}}_{\rm dr}$ is computed analogously except that ρ is set to 0 and γ is absent.

The desired properties (a) and (b) of the estimator $\widehat{\widehat{\beta}}_{\rm dr}$ are deduced from the following considerations. When condition (ii) holds the estimator $\widehat{\widehat{\beta}}_{\rm dr}$ is CAN for $\beta_{w,0}$ regardless of whether or not (3) holds because $E_n\left[q_w(\mathbf{V};\beta)(-1)^{1-Z}\,p(Z|\mathbf{X};\widehat{\alpha})^{-1}\,b\left(\mathbf{X};\beta,\widehat{\alpha},\widehat{\eta}(\beta),\widehat{\gamma},\widehat{\theta}(\beta)\right)\right]$ converges to zero in probability for all β. On the other hand, the convergence of $\widehat{\widehat{\beta}}_{\rm dr}$ to β_0 when (3) and condition (i) hold, and the convergence of $\widehat{\widehat{\beta}}_{\rm dr}$ to $\beta_{w,0}$ when (3) is incorrect but (24) holds follows arguing as in section 3.2 for the convergence of $\widehat{\beta}_{\rm dr}$.to β_0 when condition (i) holds, after noticing that the class

$$\mathcal{H}^{\mathbf{ext}}\equiv\left\{h_w(z,\mathbf{x};\beta,\eta,\gamma,\theta):\rho\in R^K,\lambda\in R^K,\beta\in R^p,\gamma\in\Gamma\right\}$$

with θ defined as in (25) includes both the class $\mathcal{H}$ (corresponding to $\lambda=0$) and the class $\mathcal{H}_w$ (corresponding to $\rho=0$).

An estimator $\widetilde{\widehat{\beta}}_{\rm dr}$ satisfying (a) and (b) and additionally guaranteed to be at least as efficient asymptotically as $\widehat{\beta}_{\rm ipw}$ is constructed just as $\widetilde{\beta}_{\rm dr}$ in

section 3.2 but replacing $a(\mathbf{X};\widehat{\alpha},\widehat{\eta}(\beta),\widehat{\gamma})$ with $b\left(\mathbf{X};\beta,\widehat{\alpha},\widehat{\eta}(\beta),\widehat{\gamma},\widehat{\theta}(\beta)\right)$, $q(\mathbf{V};\beta)$ with $q_w(\mathbf{V};\beta)$ and $h(Z,\mathbf{X};\widehat{\eta}(\beta),\widehat{\gamma})$ with $h_w\left(Z,\mathbf{X};\beta,\eta,\widehat{\gamma},\widehat{\theta}(\beta)\right)$. In the supplementary Web Appendix we also describe an estimator $\widehat{\widehat{\beta}}_{\text{opt,dr}}$ which satisfies property (a) and has limiting normal distribution with variance equal to $\Sigma_{\text{q}_{\text{opt},1},s_{\text{opt},1}}$ when conditions (i) and (ii) of section 3.2 hold and yet converges to a weighted least squares approximation when the specification (3) for $LATE(\mathbf{V})$ is wrong.

For estimation of the $MLATE(\cdot)$ curve in the supplementary Web Appendix we show that the estimator $\widehat{\widehat{\beta}}_{\text{dr}}$ computed using $H_2(\beta)$ instead of $H_1(\beta)$ and with $q_w(\mathbf{V};\beta)$ redefined as $m_2(\mathbf{V};\beta)\times\{\partial m_2(\mathbf{V};\beta)/\partial\beta\}\times w(\mathbf{V})$ satisfies (a) and (b) where in the statements of these properties, specifications (3) and (24) are replaced with (4) and the specification that $E\{H_2(\beta_{w,0})\mid Z=z,\mathbf{X}=\mathbf{x}\}\in\mathcal{H}_w$ respectively, and $\beta_{w,0}$ is redefined as

$$\beta_{w,0}\equiv\arg\min_{\beta}E\left[e_0(\mathbf{V})\,w(\mathbf{V})\left\{MLATE(\mathbf{V})-m_2(\mathbf{V};\beta)\right\}^2|D_1>D_0\right],$$

with $e_0(\mathbf{v})\equiv E(Y_0|D_1>D_0,\mathbf{V}=\mathbf{v})$. Note that, unlike the definition (21), $\beta_{w,0}$ is now a weighted least squares approximation with weights that are unknown to the data analyst since they depend on the unknown function $e_0(\mathbf{V})$. It does not appear to be possible to construct doubly robust estimators of weighted least squares approximations to the $MLATE(\cdot)$ curve for known, i.e. user-specified, weights.

4. Connections to models for the treatment effect on the treated

Robins (1994) and Tan (2010) considered estimation of the so-called additive treatment effect on the treated contrast

$$ATT(z,\mathbf{v})\equiv E(Y_1|D_z=1,\mathbf{V}=\mathbf{v})-E(Y_0|D_z=1,\mathbf{V}=\mathbf{v}).$$

This contrast quantifies the effect of treatment D on the subset of the subpopulation with baseline covariates $\mathbf{V}=\mathbf{v}$ comprised of subjects who would be treated with $D=1$ if Z were set to z. Robins (1994) showed for $\mathbf{V}=\mathbf{X}$ and Tan (2010) showed for $\mathbf{V}$ a strict subset of $\mathbf{X}$, that $ATT(z,\mathbf{v})$ is identified under the IV assumptions assumptions (i)-(iv) and (vi) and specific restrictions on $ATT(\cdot,\cdot)$. In particular, Robins (1994) showed that when $\mathbf{V}=\mathbf{X}$, $ATT(z,\mathbf{v})$ is identified under the assumptions (i)-(iv), (vi), and the assumption

(v-ATT) No additive treatment-instrument interaction on the treated: $ATT(z,\mathbf{v})=ATT(\mathbf{v})$ does not depend on z.

Remarkably, Robins showed that under these assumptions $ATT(\mathbf{v})$ is equal to the $IV(\mathbf{v})$.

In fact, it is easy to show that the preceding assertions remain true when $\mathbf{V}$ is a strict subset of $\mathbf{X}$. We thus see that under assumptions (i)-(iv) and (vi), the structural interpretation of the observed data functional $IV(\mathbf{v})$ depends on which of the assumptions (v) or (v-ATT) is adopted. The only exception is when $P(D_0 = 1) = 0$, or equivalently when $P(D = 1|Z = 0) = 0$, since in such case the complier subpopulation is the same as the subpopulation defined by condition $D_1 = 1$, and consequently $LATE(\mathbf{v}) = ATT(\mathbf{v})$.

A further deep connection exists between the works of Robins (1994) and Tan (2010) and the problem addressed in this article. For short, refer to the model defined by assumptions (i)-(vi) as "our additive model" and to the model defined by assumptions (i)-(iv), (vi) and (v-ATT) as the "Robins-Tan additive model". Remarkably, the problem of estimating the parameter β indexing a parametric specification $m_1(\mathbf{v}; \beta)$ for $LATE(\mathbf{v})$ under our additive model is formally identical to the problem of estimating the parameters β indexing a parametric specification $m_1(\mathbf{v}; \beta)$ for $ATT(\mathbf{v})$ under the Robins-Tan additive model. This surprising fact is explained by the following three results whose proofs will be sketched below:

(a) under the intersection model that assumes (i)-(vi) and (v-ATT), i.e. the model that makes simultaneously the assumptions of our additive model and of the Robins-Tan additive model, $LATE(\mathbf{v})$ and $ATT(\mathbf{v})$ are indeed identical causal effect contrasts,

(b) our model is statistically indistinguishable from the intersection model. That is, given our model, the intersection model imposes restrictions that always fit the observed data perfectly and hence cannot be rejected by any statistical test,

(c) the restrictions imposed on the observed data law by the intersection model and not imposed by the Robins-Tan additive model are only inequality constraints.

Results (a) and (b) imply that a functional of the observed data law is equal to $LATE(\mathbf{v}) = ATT(\mathbf{v})$ under the intersection model if and only if it is equal to $LATE(\mathbf{v})$ under our additive model. If this were not the case, there would be some observed data law functional equal to $LATE(\mathbf{v})$ under the intersection model but not under our additive model (the opposite is not possible because our additive model is bigger than the intersection model). But in such case, there would be a restriction, specifically the restriction that sets the new functional equal to $LATE(\mathbf{v})$, that would be satisfied under the intersection model but not under our additive model, thus contradicting (b).

Result (c) implies that a functional of the observed data law is equal to $ATT(\mathbf{v})$ under the intersection model if and only if it is equal to $ATT(\mathbf{v})$

under the Robins-Tan additive model. If this were not the case, the intersection model would satisfy an equality constraint not satisfied by the Robins-Tan additive model, namely the constraint that sets a new functional of the observed data law equal to $ATT(\mathbf{v})$, thus contradicting (c).

Results (a)-(c) then imply that any functional of the observed data law that is equal to $ATT(\mathbf{v})$ under the Robins-Tan must be equal to $LATE(\mathbf{v})$ under our additive model and vice versa. This, in turn, proves that the problem of conducting inference about the parameters β of models $m_1(\mathbf{v};\beta)$ for $ATT(\mathbf{v})$ under the Robins-Tan assumptions is formally the same as the problem of conducting inference about the parameters β indexing a parametric specification $m_1(\mathbf{v};\beta)$ for $LATE(\mathbf{v})$ under our additive model.

A further result (result (d) stated below) implies that $IV(\mathbf{v})$ is indeed the only functional of the observed data law that is equal to $LATE(\mathbf{v})$ under our additive model, and consequently, the only observed data functional equal to $ATT(\mathbf{v})$ under the Robins-Tan additive model.

(d) The only restrictions imposed on the observed data law by our additive model are inequality constraints on certain conditional distributions.

As indicated, result (d) implies that no functional of the observed data law other than $IV(\mathbf{v})$ can be equal to $LATE(\mathbf{v})$ under our additive model. If this were not the case, then the observed data law would satisfy an equality constraint under our model, namely the equality that sets $IV(\mathbf{v})$ equal to the other functional that agrees with $LATE(\mathbf{v})$, thus contradicting (d).

We now demonstrate results (a)-(d). Results (a) and (b) are a consequence of the fact that the intersection model can be equivalently defined as the model that imposes restrictions (i)-(vi) and the additional restriction

$$E(Y_1 - Y_0|T = co, \mathbf{V}) = E(Y_1 - Y_0|T = at, \mathbf{V}) \tag{27}$$

where T denotes compliance type, i.e. $T = at$ iff $D_1 = D_0 = 1$ (always taker), $T = nt$ iff $D_1 = D_0 = 0$ (never taker), $T = co$ iff $D_1 > D_0$ (complier) and $T = de$ iff $D_1 < D_0$ (defier). This equivalence holds because assumption (v-ATT) is the same as the assumption that

$$E(Y_1 - Y_0|T \in \{at, co\}, \mathbf{V}) = E(Y_1 - Y_0|T \in \{at, de\}, \mathbf{V}). \tag{28}$$

Thus, when no defiers exist, i.e. when assumption (v) holds, (28) is equivalent to (27).

Result (a) follows because restriction (27) implies that $ATT(\mathbf{v}) \equiv E(Y_1 - Y_0|T \in \{co, at\}, \mathbf{V} = \mathbf{v}) = E(Y_1 - Y_0|T = co, \mathbf{V} = \mathbf{v}) \equiv LATE(\mathbf{v})$, so under the intersection model, $LATE(\mathbf{v})$ is indeed equal to $ATT(\mathbf{v})$. Result (b) follows because under assumptions (i)-(vi), a test of the intersection model is a test that restriction (27) holds. No test can be constructed

with power to detect departures from (27) because $E(Y_0|T=at,\mathbf{V})$ is not identified and the law of the observed data does not bound its range, when, as we have assumed throughout Y has unbounded support.

Results (c) and (d) are a consequence of the following Lemmas whose proofs are given in the supplementary Web Appendix.

Lemma 1: The only restrictions on the observed data law encoded by our additive model are $0 < P(Z=1|\mathbf{X}) < 1$ and the following inequality constraints. For any $y < y'$,

$$\Pr(y < Y \leq y', D=1|Z=1,\mathbf{X}) - \Pr(y < Y \leq y', D=1|Z=0,\mathbf{X}) \geq 0 \tag{29}$$

$$\Pr(y < Y \leq y', D=0|Z=0,\mathbf{X}) - \Pr(y < Y \leq y', D=0|Z=1,\mathbf{X}) \geq 0 \tag{30}$$

$$E\{E(D|Z=1,\mathbf{X})|\mathbf{V}\} - E\{E(D|Z=0,\mathbf{X})|\mathbf{V}\} > 0. \tag{31}$$

Lemma 2: the only restrictions on the observed data law imposed by the Robins-Tan additive model are $0 < P(Z=1|\mathbf{X}) < 1$ and $E\{E(D|Z=1,\mathbf{X})|\mathbf{V}\} - E\{E(D|Z=0,\mathbf{X})|\mathbf{V}\} \neq 0$.

It is interesting to contrast the structural interpretation of the functional $E(H_1|Z,\mathbf{X})$ under our additive model and the Robins-Tan additive models. In the supplementary Web Appendix we show that under the Robins-Tan additive model,

$$E(H_1|Z=z,\mathbf{X}) = E(Y_0|\mathbf{X}) - \{ATT(\mathbf{V}) - ATT(z,\mathbf{X})\} P(D_z=1|\mathbf{X})$$

and under our additive model,

$$\begin{aligned} E(H_1|Z=z,\mathbf{X}) &= E(Y_0|\mathbf{X}) + \{E(Y_0-Y_1|\mathbf{X},T=at) - LATE(\mathbf{X})\} \times \\ &\quad P(T=at|\mathbf{X}) + \{LATE(\mathbf{X}) - LATE(\mathbf{V})\} \times \\ &\quad \{zP(T\in\{at,\ co\}|\mathbf{X}) + (1-z)P(T=ne|\mathbf{X})\}. \end{aligned} \tag{32}$$

Abadie (2003) has previously derived (32) in the special case $\mathbf{V}=\mathbf{X}$ under our additive model. Observe that only under the Robins-Tan additive model and only for the special case $\mathbf{V}=\mathbf{X}$, $E(H_1|Z,\mathbf{X})$ has a simple structural interpretation, namely as $E(Y_0|\mathbf{X}=\mathbf{x})$ (since by v-ATT implies $ATT(z,\mathbf{X}) = ATT(\mathbf{X})$ when $\mathbf{V}=\mathbf{X}$). No simple structural meaning can be given to $E(H_1|Z,\mathbf{X})$ in all other cases. It is this counterintuitive aspect of the functional $E(H_1|Z,\mathbf{X})$ that we believe may have delayed the discovery of the doubly robust estimators of β proposed in this article.

Robins (1994) and Tan (2010) also discussed inference about models for the multiplicative treatment effect on the treated curve $MTT(z,\mathbf{v}) \equiv E(Y_1|D_z=1,\mathbf{V}=\mathbf{v})/E(Y_0|D_z=1,\mathbf{V}=\mathbf{v})$. Deep connections along the lines made in this section also exist between the work of these authors for inference about $MTT(z,\mathbf{v})$ and the proposal for estimation about $MLATE(\mathbf{v})$ in this paper.

5. Data Analysis

We apply the procedures discussed in this paper to estimate the local average treatment effect of participation in 401(k) programs on household saving. 401(k) tax-deferred retirement plans were introduced in the 1980s with the goal of encouraging household saving; they have since grown to be the most popular retirement plans in the United States. But economists have hypothesized that 401(k) plans may not represent increased saving, rather they may replace other modes of saving for those who participate. Among people who are eligible to participate in 401(k) plans, those who choose to participate are likely more inclined to save than those who choose not to participate. Therefore, standard methods for examining the effect of 401(k) participation on savings based on covariate adjustment are inappropriate as underlying saving preference is an unmeasured confounder of the treatment-outcome relationship. Using 401(k) eligibility as an instrument for 401(k) participation, estimation of the local average treatment effect of 401(k) participation on savings is feasible.

Poterba et al. (1994, 1995) and Abadie (2003) analyzed data from the U.S. Census Bureau's 1991 Survey of Income and Program Participation (SIPP) to test whether participation in 401(k) plans increases household savings. Here we reanalyze the data analyzed by Abadie (2003), consisting of a sample of 9,725 household reference subjects aged 25 to 64 and their spouses, with annual income between $10,000 and $200,000. In our analysis as in Abadie's, the outcome Y is net financial assets, the instrument Z is an indicator of 401(k) eligibility, the treatment D is an indicator of 401(k) participation, and the vector of covariates is $\mathbf{X} = (X_1, X_2, X_3, X_4)$ where X_1 is age (approximated to the closest integer year *after* subtracting off the minimum age in the sample), X_2 is an indicator of marital status (married or not), X_3 is family size, and X_4 is annual household income (in $1000).

In this example, the instrumentation assumption (iv) and monotonicity assumption (v) hold trivially because it is not possible to choose to participate in 401(k) plans if not eligible to do so ($D_0 = 0$ with probability 1). The exclusion restriction (ii) is very plausible because 401(k) plans are run through employers with only some employers granting eligibility to their employees; evidence suggests that the effect of an employer's offer of 401(k) eligibility on an employee's saving behavior operates only through the employee's choice to participate or not in the program (Poterba et al., 1995). Finally, the randomization assumption is also likely to hold when we include in $\mathbf{X}$ the measured predictors income, age, marital status, and family size of eligibility and savings. Because $D_0 = 0$ there can be no defiers or always takers and the complier subpopulation is comprised of all eligible subjects who chose to participate; consequently $LATE(\cdot) = ATT(\cdot)$ is estimable with the SIPP data.

To illustrate our methodology we considered estimation of the parameters indexing models for $LATE(\mathbf{V})$ for two choices of $\mathbf{V}$, namely $\mathbf{V} = X_4$ (income) and $\mathbf{V} =$ null. We will see that the analysis when $\mathbf{V} = X_4$ showed that income was a significant determinant of LATE. This gave us the opportunity to explore the behavior of the proposed estimators under misspecification of the model for the $LATE(\cdot)$ curve. Specifically, we applied the procedures in this paper to estimate a scalar parameter β under the specification $m(\mathbf{X};\beta) = \beta$, i.e. under a, likely misspecified, model that assumes that $LATE(\mathbf{X})$ does not depend on income or any of the other covariates in $\mathbf{X}$. This specification was also used to analyze this data in Abadie (2003).

Table 1 reports the estimators of β with their bootstrap standard errors in parenthesis in the case $\mathbf{V} = X_4$ under the specification $m(X_4;\beta) = \beta_0 + \beta_1 X_4$. The table reports results for eight estimators: five doubly robust estimators $\widehat{\beta}_{dr}$, two IPW estimators $\widehat{\beta}_{ipw}$ and one outcome regression estimator $\widehat{\beta}_{reg}$. The estimator $\widehat{\beta}_{reg}$ was computed using the function $l(Z,\mathbf{X};\beta,\eta,\gamma)$ given in (16). Three of the doubly robust estimators, denoted with $\widehat{\beta}^{opt}_{dr}, \widehat{\beta}^{opt}_{dr,\pi-fixed}, \widehat{\beta}^{opt}_{dr,h-fixed}$, used $q(\mathbf{V})$ equal to $\widehat{q}_{opt,1}(\mathbf{V})$ as defined in section 3.3. In the calculation of $\widehat{q}_{opt,1}(\mathbf{V})$, $\log[e_1(\mathbf{V};\delta)/\{1-e_1(\mathbf{V};\delta)\}]$ and $\log\{t_1(\mathbf{V};\omega)\}$ were linear functions of income and income^2. [Note that when, as in this dataset, $Z = 0$ implies $D = 0$, $e_1(\mathbf{V};\delta)$ is a model for $E\{E(D|\mathbf{X},Z=1)|\mathbf{V}\}$]. The fourth doubly robust estimator, denoted with $\widehat{\beta}^{ineff}_{dr}$, used $q(\mathbf{V}) = \partial m(\mathbf{V};\beta)/\partial\beta = (1, X_4)^{\mathrm{T}}$ and the last doubly robust estimator, denoted with $\widehat{\beta}^{ineff,stable}_{dr}$ used $q(\mathbf{V}) = (1,X_4)^{\mathrm{T}}\left\{\text{expit}\left(\widehat{\zeta}_0+\widehat{\zeta}_1 X_4\right) - \text{expit}\left(\widehat{\zeta}_0+\widehat{\zeta}_1 X_4\right)^2\right\}$ where expit $\left(\widehat{\zeta}_0+\widehat{\zeta}_1 X_4\right)$ was the fitted value from a logistic regression of Z on X_4. These latter two choices of $q(\mathbf{V})$ were also used to construct the two IPW estimators, denoted with $\widehat{\beta}^{ineff}_{ipw}$ and $\widehat{\beta}^{ineff,stable}_{ipw}$ respectively.

In the calculation of the doubly robust and IPW estimators we used the propensity score model $\mathcal{P}^{k_\pi}$ which assumed that $\log[\pi(\mathbf{x};\alpha)/\{1-\pi(\mathbf{x};\alpha)\}]$ was linear in indicator variables of the combined levels of marital status and age as well as in all powers of income up to the power k_π. As in Abadie, 2003, we did not include family size because it did not significantly predict Z. Also, the outcome regression model in the calculation of the doubly robust estimators and of $\widehat{\beta}_{reg}$, denoted in the sequel with $\mathcal{H}^{k_h}_v$, assumed that $E\{H_1(\beta_0)|Z,\mathbf{X}\} = k(\mathbf{X};\nu) + \rho^T\left\{\underline{\varphi}(\mathbf{X}) - \underline{\phi}(\mathbf{V};\gamma)\right\}Z$. The function $k(\mathbf{x};\nu)$ was linear in powers of income up to power k_h and in indicators of the combined levels of age, marital status, and family size (dichotomized at its mean). The function $\underline{\varphi}(\mathbf{x})$ was a vector of indicators of combined levels of age, marital status and family size; each entry of $\underline{\phi}(\mathbf{v};\gamma)$ was a linear logistic regression model for the corresponding entry of $\underline{\varphi}(\mathbf{x})$ with covari-

ates being income, income^2, ...,income^{k_h}. The estimators $\widehat{\beta}_{dr}^{opt}$, $\widehat{\beta}_{dr}^{ineff}$ and $\widehat{\beta}_{dr}^{ineff,stable}$were computed using models $\mathcal{P}^{k_\pi}$ and $\mathcal{H}_v^{k_h}$ with $k_\pi = k_h \equiv k$. In Table 1 the first three rows report these estimators using k as indicated by the column labels. The estimator $\widehat{\beta}_{dr,\pi-fixed}^{opt}$ had k_π fixed at 4 and k_h as indicated by the column labels. Likewise the estimator $\widehat{\beta}_{dr,h-fixed}^{opt}$ had k_h fixed at 4 and k_π as indicated by the column labels. The estimators $\widehat{\beta}_{ipw}^{ineff}$ and $\widehat{\beta}_{ipw}^{ineff,stable}$ had k_π as indicated by the column labels. Finally, the estimator $\widehat{\beta}_{reg}$ had k_h as indicated by the column labels. In the dataset as well as in each bootstrap replication we first estimated the propensity scores, then threw out the data from subjects in the bottom and top one percent of the estimated values of $\pi(\mathbf{X}; \widehat{\alpha})$, and finally carried through the entire procedure for arriving at the estimators of β using the remaining data. In the dataset, this pruning did not noticeably change the values of our estimators, suggesting that the data pruning did not result in substantial bias, but it had a dramatic effect on stabilizing the bootstrap standard error estimators.

According to the theory presented in this paper, $\widehat{\beta}_{dr}^{opt}$ with $k_\pi = k_h$ sufficiently large should result in optimal inference about β. We therefore first examine the rows corresponding to $\widehat{\beta}_{dr}^{opt}$ and the columns with $k_\pi = k_h$ equal 4, 5 and 8 in Table 1. We note that the coefficient of income is roughly 330 with a standard error around 80 suggesting that 401(k) plans have more impact on the savings of families of higher income. For example, for $k_\pi = k_h = 4$, the estimated effect of 401(k) participation for an eligible person with annual income \$50,000 who chooses to participate in the program is to increase her family's net financial assets by \$14,910 whereas the increase for a person with an income of \$100,000 is \$31,310.

Unlike the slope coefficient, the intercept does not appear to be significantly different from 0; a 95% confidence interval for the intercept would include 0 as the point estimate is roughly half its standard error. For this reason, we henceforth focus attention on the behavior of the remaining estimators of the income coefficient. Since the three doubly robust estimators $\widehat{\beta}_{dr}^{opt}$, $\widehat{\beta}_{dr}^{ineff}$ and $\widehat{\beta}_{dr}^{ineff,stable}$ with $k_\pi = k_h$ greater than or equal to 4 are all approximately equal to 330, we conclude that it is likely that the linear model for $LATE(X_4)$ is approximately correct. If it were not, the estimators estimators $\widehat{\beta}_{dr}^{opt}$, $\widehat{\beta}_{dr}^{ineff}$ and $\widehat{\beta}_{dr}^{ineff,stable}$ would not be expected to exhibit similar values as they would have different probability limits because they use different functions $q(\mathbf{V})$. Therefore, in what follows, we will refer to an estimator of the slope coefficient as "unbiased" if it is roughly equal to 330. Observe that, as predicted by theory, the doubly robust estimators that use $\widehat{q}_{opt,1}(\mathbf{V})$ are more efficient than the IPW or any of the other doubly robust estimators. [In fact, these doubly robust estimators are even more efficient than the estimator $\widehat{\beta}_{reg}$; presumably this reflects the fact that the choice (16) we recommended for ease of calcula-

tion is not optimal]. Comparison of the IPW estimators with the estimator $\widehat{\beta}^{opt}_{dr,h-fixed}$ and of the outcome regression estimator with $\widehat{\beta}^{opt}_{dr,\pi-fixed}$ illustrates the advantage of doubly robust estimation over IPW and outcome regression estimation. These comparisons reveal that doubly robust estimators only require one of the two models to be nearly correct and the analyst does not need to know which one is correct. Note that whereas the IPW estimators are severely "biased" if k_π is 1 or 2, the doubly robust estimator $\widehat{\beta}^{opt}_{dr,h-fixed}$ that uses the same model for the propensity score but a model $\mathcal{H}^{k_h}_v$ with k_h equal to 4 is roughly "unbiased". Likewise, the outcome regression estimator that has k_h equal 1 or 2 is "biased" but the "bias" is corrected by the estimator $\widehat{\beta}^{opt}_{dr,\pi-fixed}$.

Turn now to estimation of β under a model $m(\mathbf{X};\beta)$ for $LATE(\mathbf{X})$ that assumes that $m(\mathbf{X};\beta)=\beta$. This model is presumably wrong because, as we have already seen from the previous analysis, income modifies the effect of treatment D among the compliers. Additional evidence for misspecification is presented in Figure 1. This figure displays the values of three different doubly robust estimators $\widehat{\beta}_{dr}$, denoted with $\widehat{\beta}^{opt}_{dr}$, $\widehat{\beta}^{ineff}_{dr}$ and $\widehat{\beta}^{ineff,stable}_{dr}$ which used respectively $q(\mathbf{X}) = e_1(\mathbf{X};\widehat{\delta})t_1(\mathbf{X};\widehat{\omega}), q(\mathbf{X}) = \partial m(\mathbf{X};\beta)/\partial\beta = 1$ and $q(\mathbf{X}) = \pi(\mathbf{X},\widehat{\alpha}) - \pi(\mathbf{X},\widehat{\alpha})^2$, where $\log[e_1(\mathbf{X};\delta)/\{1-e_1(\mathbf{X};\delta)\}]$ and $\log\{t_1(\mathbf{X};\omega)\}$ were linear functions of family size, income, income^2 and indicators of age and marital status. The estimators assumed model $\mathcal{P}^{k_\pi}$ for the propensity score and an outcome regression model $\mathcal{H}^{k_h}_x$ that specifies that $E\{H_1(\beta_0)|Z,\mathbf{X}\} = k(\mathbf{x};\nu)$ where $k(\mathbf{x};\nu)$ is the same function as defined earlier. [Recall that under the assumption that the model $m(\mathbf{X};\beta)$ is correct, $E\{H_1(\beta_0)|Z,\mathbf{X}\}$ does not depend on Z]. The plot displays the values of $\widehat{\beta}^{opt}_{dr}$, $\widehat{\beta}^{ineff}_{dr}$ and $\widehat{\beta}^{ineff,stable}_{dr}$ as $k_h = k_\pi \equiv k$ varies from 1 to 8. Each estimator stabilizes for k greater than or equal to 3; however each stabilizes to a different value. This is as predicted by the theory of section 3.4 according to which, when model $m(\mathbf{X};\beta)$ is incorrect and model $\mathcal{P}^{k_\pi}$ is correct each estimator converges in probability to a distinct weighted least squares approximation $\beta_{0,w}$ with a weight that depends on the choice of function $q(\mathbf{X})$. Specifically, when $\mathcal{P}^{k_\pi}$ is correct and the model $m(\mathbf{X};\beta)$ for $LATE(\mathbf{X})$ is misspecified, $\widehat{\beta}^{ineff}_{dr}, \widehat{\beta}^{ineff,stable}_{dr}$ and $\widehat{\beta}^{opt}_{dr}$ converge in probability to distinct values $\beta_{0,w_{ineff}}$, $\beta_{0,w_{ineff,stable}}$ and $\beta_{0,w_{opt}}$ where $w_{ineff}(\mathbf{X}) = 1, w_{ineff,stable}(\mathbf{X}) = \pi(\mathbf{X},\alpha_0) - \pi(\mathbf{X},\alpha_0)^2$ and $w_{opt}(\mathbf{X}) = e_1(\mathbf{X};\delta^*)t_1(\mathbf{X};\omega^*)$ with δ^* and ω^* the probability limits of $\widehat{\delta}$ and $\widehat{\omega}$.

The parameter $\beta_{0,w_{ineff}}$ is of particular interest as an easy calculation shows that $\beta_{0,w_{ineff}}$ is equal to the marginal LATE, i.e. to $\beta_{null} \equiv LATE(\mathbf{V})$ when $\mathbf{V}$ = null. Thus, the estimator $\widehat{\beta}^{ineff}_{dr}$ converges to β_{null} when the model $\mathcal{P}^{k_\pi}$ is correct. In fact, the IPW estimator $\widehat{\beta}^{ineff}_{ipw}$ that uses the same $q(\mathbf{X})$ as $\widehat{\beta}^{ineff}_{dr}$ and the same model $\mathcal{P}^{k_\pi}$ also converges

to β_{null} when model $\mathcal{P}^{k_\pi}$ is correct. This is so because $\widehat{\beta}_{dr}$ and $\widehat{\beta}_{ipw}$ have the same probability limits when they use the same correctly specified propensity score model regardless of whether or not the parametric specification for $LATE(\cdot)$ is correct. These theoretical results are confirmed in Figure 2. The figure displays the estimators $\widehat{\beta}_{ipw}^{ineff}$ and $\widehat{\beta}_{dr}^{ineff}$ computed under model $\mathcal{P}^{k_\pi}$ and model $\mathcal{H}_x^{k_h}$ with $k_h = k_\pi = k$. In addition, the figure displays the doubly robust estimator $\widehat{\beta}_{null,dr}$ of β_{null}, i.e. of the marginal LATE. This estimator is computed under model $\mathcal{P}^{k_\pi}$ and a model $\mathcal{H}_{null}^{k_h}$ that assumes that $E\{H_1(\beta_{null})|Z,\mathbf{X}\} = k(\mathbf{X};\nu) + \rho^T\left[\underline{\varphi}(\mathbf{X}) - E\left\{\underline{\varphi}(\mathbf{X})\right\}\right]Z$ with $k(\mathbf{x};\nu)$ as defined earlier and $\underline{\varphi}(\mathbf{x})$ a vector function of indicators of the combined levels of age, marital status, family size (dichotomized at its mean) and powers of income up to power k_h. Note that in Figure 2 $\widehat{\beta}_{ipw}^{ineff}$ and $\widehat{\beta}_{dr}^{ineff}$ are both close to $\widehat{\beta}_{null,dr}$ for k_π greater than or equal 4.

If model $\mathcal{P}^{k_\pi}$ is wrong and $m(\mathbf{X};\beta) = \beta$ is an incorrect specification for $LATE(\mathbf{X})$ both $\widehat{\beta}_{ipw}^{ineff}$ and $\widehat{\beta}_{dr}^{ineff}$ are inconsistent for $\beta_{0,w_{ineff}} = \beta_{null}$. This occurs because, as discussed in section 3.4, $\widehat{\beta}_{dr}^{ineff}$ is not doubly robust for $\beta_{0,w_{ineff}}$ under incorrect specification of the model for the $LATE(\cdot)$ curve. In contrast, $\widehat{\beta}_{null,dr}$ is double robust for β_{null}, i.e. it is consistent either if model $\mathcal{P}^{k_\pi}$ is correct or if model $\mathcal{H}_{null}^{k_h}$ is correct. In fact, $\widehat{\beta}_{null,dr}$ is a member of the class of estimators $\widehat{\widehat{\beta}}_{dr}$ described in section 3.4; it is algebraically equal to the estimator $\widehat{\widehat{\beta}}_{dr}$ that uses $q_w(\mathbf{V}) = 1$ with $\mathbf{V} = \mathbf{X}$. Recall that, unlike $\widehat{\beta}_{dr}$, the estimator $\widehat{\widehat{\beta}}_{dr}$ that uses a given $q_w(\mathbf{V})$ is doubly robust for $\beta_{0,w}$. Table 2 illustrates these points. The row labeled "Model $\mathcal{P}^{k_\pi}$" lists estimators computed under model $\mathcal{P}^{k_\pi}$ with $k_\pi = 4$. The row labeled "Model $\mathcal{P}^{wrong}$" lists estimators computed under the model $\mathcal{P}^{wrong}$ that incorrectly sets $P(Z=1|\mathbf{X})$ to be equal to the constant 1/2. For estimators $\widehat{\beta}_{null,dr}$ and $\widehat{\beta}_{dr}^{ineff}$, k_h was chosen to be 4. All the estimators in the first row are approximately equal. However, a column by column comparison of the two rows reveals that of the three estimators only $\widehat{\beta}_{null,dr}$ remains approximately unchanged when it is computed under $\mathcal{P}^{wrong}$. This is as predicted by theory (provided that the model $\mathcal{H}_{null}^{k_h}$ with k_h =4 is approximately correct). To confirm that these findings were unlikely due to chance , we computed for each column the ratio $\widehat{T} = \widehat{\Delta}/\widehat{SE}$ where $\widehat{\Delta}$ is the difference between the first and second row, and $\widehat{SE}$ is the bootstrap standard error of $\widehat{\Delta}$. Under the null hypothesis that the probability limits of the estimators in the two rows are the same, T should approximately have a standard normal distribution. For $\widehat{\beta}_{null,dr}$, $\widehat{T}$ was 0.51 whereas for $\widehat{\beta}_{dr}^{ineff}$ and $\widehat{\beta}_{ipw}^{ineff}$, $\widehat{T}$ was -1.91 and -3.14 respectively.

Table 1. Estimators of (β_0, β_1) and their bootstrap standard errors under model $LATE(income) = \beta_0 + \beta_1 income$.

		Power k of income in the outcome regression and propensity score models					
		1	2	3	4	5	8
Intercept	$\widehat{\beta}_{dr}^{opt}$	-4640 (2940)	-1845 (3220)	-1888 (2940)	-1490 (2900)	-1623 (2907)	-1566 (2896)
	$\widehat{\beta}_{dr}^{ineff}$	1774 (5720)	-12860 (10720)	-3846 (5797)	-14201 (11244)	-3877 (7061)	-1578 (7009)
	$\widehat{\beta}_{dr}^{ineff,\,stable}$	-418 (4827)	-4958 (5547)	-2049 (4385)	-1814 (4527)	-2448 (4465)	-1590 (4543)
	$\widehat{\beta}_{dr,h-fixed}^{opt}$	-1572 (3292)	-1411 (3146)	-1285 (2873)	-1490 (2900)	-1592 (2989)	-1674 (2914)
	$\widehat{\beta}_{dr,\pi-fixed}^{opt}$	-2093 (2961)	-1421 (2947)	-1911 (2816)	-1490 (2900)	-1650 (2826)	-1517 (2920)
	$\widehat{\beta}_{ipw}^{ineff}$	17075 (7870)	-18515 (11587)	-4905 (6487)	-858 (6841)	-1980 (6732)	-593 (7655)
	$\widehat{\beta}_{ipw}^{ineff,\,stable}$	12331 (6076)	-3489 (5632)	-2775 (4101)	-1478 (4019)	-1537 (4202)	-1179 (4409)
	$\widehat{\beta}_{reg}$	-6992 (7019)	1929 (7665)	-2652 (6886)	-1266 (6796)	-1721 (6702)	-1494 (7004)
Income	$\widehat{\beta}_{dr}^{opt}$	382 (88)	337 (92)	338 (83)	328 (82)	330 (83)	328 (83)
	$\widehat{\beta}_{dr}^{ineff}$	205 (171)	634 (290)	390 (165)	351 (197)	392 (197)	329 (196)
	$\widehat{\beta}_{dr}^{ineff,\,stable}$	272 (128)	425 (149)	345 (115)	340 (123)	354 (122)	331 (120)
	$\widehat{\beta}_{dr,h-fixed}^{opt}$	319 (96)	323 (90)	326 (80)	328 (82)	329 (82)	332 (84)
	$\widehat{\beta}_{dr,\pi-fixed}^{opt}$	342 (84)	328 (82)	340 (84)	328 (82)	332 (82)	328 (79)
	$\widehat{\beta}_{ipw}^{ineff}$	-139 (218)	785 (306)	425 (178)	320 (181)	347 (181)	311 (201)
	$\widehat{\beta}_{ipw}^{ineff,\,stable}$	14 (161)	385 (154)	368 (119)	339 (117)	336 (114)	329 (123)
	$\widehat{\beta}_{reg}$	510 (187)	272 (210)	361 (181)	345 (183)	357 (180)	353 (194)

Table 2. Estimation of the marginal LATE effect.

	Point estimators*		
	$\widehat{\beta}_{null,dr} = \widehat{\widehat{\beta}}_{dr}$	$\widehat{\beta}_{dr}^{ineff}$	$\widehat{\beta}_{ipw}^{ineff}$
Model $\mathcal{P}^{k_\pi=4}$	12213	12179	12434
Model $\mathcal{P}^{wrong}$	11859	13140	17651
	Test statistic**		
	0.51	-1.91	-3.14

* $\widehat{\widehat{\beta}}_{dr}$ is the estimator of section 3.4 that uses $q_w(\boldsymbol{V}) = 1$.
** Test statistic is the difference of the estimators in the first and second rows divided by the bootstrap standard error of the difference.

Figure 1: Estimation of the marginal LATE based on incorrectly assuming that $LATE(X) = LATE$.

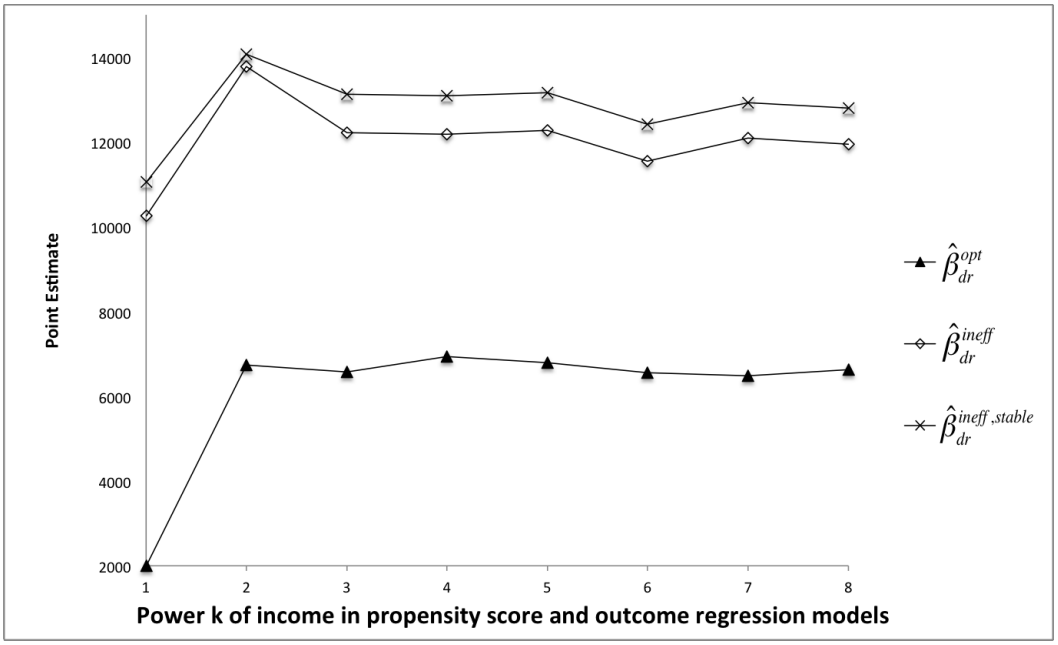
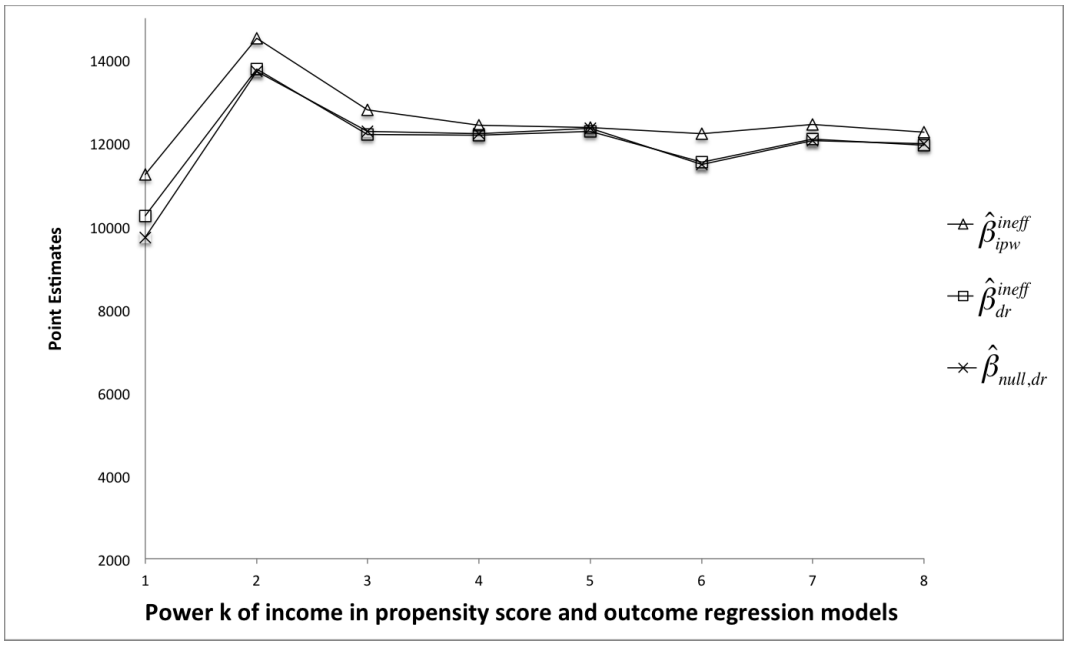


Figure 2: Doubly robust estimation of the marginal LATE vs estimation based on incorrectly assuming that $LATE(X) = LATE$.



6. Conclusion

In this paper we introduced a new class of estimators for parametric forms for additive and multiplicative local average treatment effect curves as functions of covariates $\mathbf{V}$, where $\mathbf{V}$ may be a subset of the covariates $\mathbf{X}$ required for the candidate instrument to be a valid instrumental variable. Our estimators are doubly robust, i.e. they are consistent and asymptotically normal if either one of two dimension reducing models is correctly specified. Unlike other proposals, these dimension reducing models are always compatible with the assumed parametric functional form for the local average treatment effect on the additive scale if Y has unbounded support, and with the assumed parametric functional form for the effect on the multiplicative scale if Y has support in the positive real line and is unbounded. We discussed the connection between our model for the local average treatment effects and the Robins-Tan model for the effect of treatment on the treated, and argued that the correspondence between the two models is unsurprising because the restrictions on the observed data law imposed by the two models differ only in inequality constraints, and because under an untestable assumption about the distribution of the counterfactual outcomes the two estimands are identified by the same functional of the observed data.

Future work is needed to explore the performance of our estimators for weak instruments in finite samples. Another potential topic for future work arises from the fact that, when Y is binary, the outcome regression model and the model for $MLATE(\cdot)$ are not variation independent. Thus, the model $m_2(\cdot;\beta)$ could conflict with a proposed model for $E\left(H_2|Z,\mathbf{X}\right)$. If the propensity score model is correctly specified the resulting estimator of β_0 will still be consistent, however this variation dependence implies that we may not have two independent opportunities for valid inference about β_0. In forthcoming work, we reparameterize the model for MLATE when Y is binary to recover doubly robustness.

Acknowledgements

The authors are grateful to Alberto Abadie for his helpful comments on an earlier draft. Elizabeth Ogburn was supported by by a training grant from the National Institutes of Health (5T32 AI 7358-22) from the National Institutes of Health. Andrea Rotnitzky and James Robins were partially supported by grant R01-AI051164 from the National Institutes of Health.

References

Abadie, A. (2002) Bootstrap tests for distributional treatment effects in instrumental variable models. *J. Am. Statist. Ass.*, **97**, 284–292.

——— (2003) Semiparametric Instrumental Variable Estimation of Treatment Response Models. *J. Econometrics*, **113**, 213–263.

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

Angrist, J. D., Graddy, K., and Imbens, G. W. (2000) The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Rev. Econ. Stud.*, **67**, 499–527.

Angrist, J. D. and Imbens, G. W. (1995) Average causal response with variable treatment intensity. *J. Am. Statist. Ass.*, **90**, 431–442.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996) Identification of Causal Effects Using Instrumental Variables (with discussion). *J. Am. Statist. Ass.*, **91**, 444–471.

Clarke, P. and Windmeijer, F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics.*, **11**, 756–770.

Cheng, J, Small, D., Tan, Z. and Ten Have, T. (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika.*, **96**, 1–9.

Cheng, J, Qin, J., Zhang, B. (2009). Semiparametric estimation and inference for distributional and general treatment effects. *Journal of the Royal Statistical Society, Series B.*, **71**, 881–904.

Froelich, M. (2007) Nonparametric IV estimation of local average treatment effects with covariates. *J. Econometrics*, **139**, 35–75.

Gill, R. D. (1989) Non- and Semi-Parametric Maximum Likelihood Estimators and the Von Mises Method (Part 1). *Scand. J. Statist.*, **16**, 97–128.

Heckman, J. (1976) The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models. *Ann. Econ. Soc. Meas.*, **5**, 475–492.

Hirano. K., Imbens, G. W., Rubin, D. B., and Zhou, X. H. (2000) Assessing the Effect of an Influenza Vaccine in an Encouragement Design. *Biostatistics*, **1**, 69–88.

Imbens, G. W. and Angrist, J. D. (1994) Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**, 467–475.

Kasy, M. (2009) Semiparametrically Efficient Estimation of Conditional Instrumental Variables Parameters. *Int. J. of Biostat.*, **5**, Article 22.

Little, R. J. and Yau, L. H. Y. (1998) Statistical Techniques for Analyzing Data From Prevention Trials: Treatment of No-Shows Using Rubin's Causal Model. *Psychol. Methods*, **3**, 147–159.

Newey, W. (1994) The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, **62**, 1349–1382.

Poterba, J. M., Venti, S. F. and Wise, D. A. (1994) 401(k) Plans and Tax-Deferred Savings. In *Studies in the Economics of Aging* (ed. D. Wise), pp. 105–138. Chicago: University of Chicago Press.

——— (1995) Do 401(k) Contributions Crowd Out Other Personal Saving? *J. Public Econ.*, **58**, 1–32.

Robins, J. M. (1994) Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models. *Commun. Statist. A—Theor.*, **23**, 2379–2412.

Robins, J. M. and Hernan, M. A. (2006) Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, **17**, 360–372.

Robins, J. M. and Rotnitzky, A. (1992) Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In *Aids Epidemiology: Methodological Issues* (eds. N. Jewell, K. Dietz, and V. Farewell), pp. 297–331. Boston: Birkhauser.

Stefanski, L. A. and Boos, D. D. (2002) The Calculus of M-Estimation. *Am. Stat.*, **56**, 29–38.

Tan, Z. (2006a) Regression and Weighting Methods for Causal Inference Using Instrumental Variables. *J. Am. Statist. Ass.*, **101**, 1607–1618.

——— (2006b) A Distributional Approach for Causal Inference Using Propensity Scores. *J. Am. Statist. Ass.*, **101**, 1619–1637.

———(2010) Marginal and Nested Structural Models Using Instrumental Variables. *J. Am. Statist. Ass.*, **105**, 157–169.

Uysal, S. D. (2011) Doubly Robust IV Estimation of the Local Average Treatment Effects. (Available from http://www.ihs.ac.at/vienna/resources/Economics/Papers/Uysal_paper.pdf.)

Vytlacil, E. J. (2002) Independence, monotonicity, and latent index models: an equivalency result. *Econometrica,* **70**, 331–341.