Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2015 March ; 77(2): 373–396. doi:10.1111/rssb.12078.

Doubly robust estimation of the local average treatment effect curve

Elizabeth L. Ogburn,
Johns Hopkins University, Baltimore, USA

Andrea Rotnitzky, and
Di Tella University, Buenos Aires, Argentina and Harvard University, Boston, USA

James M. Robins
Harvard University, Boston, USA

Summary

We consider estimation of the causal effect of a binary treatment on an outcome, conditionally 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 conditionally on covariates **V** when randomization of the instrument is only true conditionally on a high dimensional vector of covariates **X**, possibly bigger than **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 conditionally on **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 by using data from the US Census Bureau's 1991 Survey of Income and Program Participation.

Keywords

Instrumental variables; Local average treatment effect; Local efficiency; Multiplicative effect

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 (IVs) 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

Address for correspondence: Elizabeth L. Ogburn, Department of Biostatistics, Johns Hopkins University, Room E3620, 719 Parke Avenue, Baltimore, MD 21205, USA. eogburn@jhsph.edu.

Imbens and Angrist (1994), non-parametric and semiparametric IV methods for estimation of local average treatment effects have received considerable attention in the literature (Angrist and Imbens, 1995; Angrist *et al.*, 1996, 2000; Abadie, 2002, 2003; Abadie *et al.*, 2002; Froelich, 2007; Tan, 2006a, 2010; Kasy, 2009, Cheng *et al.*, 2009a, b).

In this paper we consider estimation of models for the dependence of local average treatment effects on baseline covariates **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 **X** is necessary for the identifying IV assumptions to be valid. These settings are important because in practice the instrument may itself be confounded, and conditioning on covariates **X** may be required to make the key condition of instrument randomization plausible (Abadie, 2003). We extend this previous work to allow **X** to be larger than **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 **X** is the set of variables that must be conditioned on for the instrument–outcome and instrument–treatment relationships to be unconfounded within levels of covariates; however, local average treatment effects conditional on **V**, a subset of **X**, may be the relevant contrasts to help guide decision makers who, because of limited resources, will have access only to information about the subset **V** of **X**. For example, consider a study conducted in a sophisticated health maintenance organization. Suppose that the instrument is the therapy prescribed by the physician, the treatment is the therapy actually followed by the patient and **X** is a vector of measured risk factors for the outcome that were used by the health maintenance organization physician to decide on the therapy prescription. The covariates **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 on the basis of just the subset **V** of **X** that encodes the data that are available to them. Estimation of effect modification of the local average treatment effects by **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 outcomes (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 **V**.

When the dimension of the covariate vector **X** is large, as will often be required in practice for the assumption of a conditionally unconfounded instrument to hold, non-parametric estimation of LATE (Froelich, 2007), of MLATE and of parametric specifications for the dependence of these contrasts on covariates **V** is not feasible, owing to the curse of

dimensionality. When **V** is null, Tan (2006a) and Uysal (2011) derived estimators of LATE that are consistent provided that either two models for two specific conditional means given the instrument and **X** or a model for the instrument propensity score (the probability that the instrument is equal to 1 conditionally on the covariates **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 **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 **X** is correctly specified. When **V** is non-null, the conditional mean models that are required by our doubly robust estimator are guaranteed to cohere with a parametric specification for the dependence of the local average treatment effect on **V**. Extensions of the doubly robust methods that were proposed by Tan (2006a) and Uysal (2011) to the case **V** non-null do not have this property.

In Section 2 we introduce the notation, models and assumptions. We also review existing non-parametric and semiparametric methods for estimating local average treatment effects with instruments confounded by **X**. In Section 3 we describe the doubly robust estimating procedures proposed and discuss efficiency properties and estimation under incorrect specifications for the dependence of LATE or MLATE on **V**. In Section 4 we explain a surprising result that was earlier noted in the absence of covariates **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 reanalyse the data used in Póterba *et al.* (1995) and Abadie (2003) with the goal of estimating the causal effect of participating in 401(k) retirement programmes on savings by using eligibility for a 401(k) programme as a binary instrument. Section 6 concludes the paper.

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, **X** is a vector of baseline covariates and Z is a binary IV. 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 make the following assumptions:

- **a.** conditional unconfoundedness of the instrument, i.e. $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1)$ is conditionally independent of Z given **X**;
- **b.** exclusion of the instrument, i.e. $P(Y_{1d} = Y_{0d}) = 1$ for $d \in \{0, 1\}$;

c. common support of the instrument, i.e. $0 < P(Z = 1|\mathbf{X}) < 1$ with probability 1;

d. instrumentation, i.e. $P(D_1 = 1|\mathbf{V}) \neq P(D_0 = 1|\mathbf{V})$ with probability 1;

e. monotonicity, i.e. $P(D_1 \geqslant D_0) = 1$;

f. consistency, i.e. $Y = DY_1 + (1 - D)Y_0$ and $D = ZD_1 + (1 - Z)D_0$, where $Y_d \equiv Y_{1d} = Y_{0d}$ by assumption (b).

When assumptions (a)–(d) and (f) hold, Z is said to be an IV for the effect of D on Y. Assumption (a) says that, within levels of $\mathbf{X}$, Z is as good as randomly assigned. Assumption (b) 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 (c) requires that there is 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 (e) excludes the existence of defiers. Assumption (f) 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 (e), assumption (d) is the same as $P(D_1 = 1|\mathbf{V}) > P(D_0 = 1|\mathbf{V})$ which, in turn, under (a) and (f) 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 (a)–(f) are conditional versions of the assumptions that were made by Angrist *et al.* (1996), and Vytlacil (2002) noted that they are equivalent to the assumptions imposed by a non-parametric 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 (a)–(f) $E(Y_1|D_1 > D_0, \mathbf{V})$ and $E(Y_0|D_1 > D_0, \mathbf{V})$ are identified, and consequently so is

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

Under the additional assumption

(g) non-null complier mean under control, i.e. $E(Y_0 \mid D_1 > D_0, \mathbf{V}) \neq 0$ with probability 1, the contrast

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

is well defined with probability 1 and is identified.

For conciseness, we shall refer to assumptions (a)–(f) if referring to inference about LATE(·) or (a)–(g) if referring to inference about MLATE(·) as the IV assumptions.

The curves LATE(**v**) and MLATE(**v**) describe how treatment effects in the complier subpopulation vary with values **v** of **V**, the first quantifying the effects on an additive scale

and the second on a multiplicative scale. Theorem 1 of Tan (2006a) implies that under the IV assumptions LATE(**v**) is equal to the conditional version of the IV estimand,

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

and MLATE(**v**) is

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

(see also theorem 3.1 of Abadie, (2003)). The M in front of the functional MIV is a reminder that this functional identifies a multiplicative treatment effect. The functionals IV(·) and MIV(·) are the target of inference when, as we shall assume throughout, the IV assumptions are valid and interest is in estimation of LATE(·) and MLATE(·).

2.2. Review of existing estimators

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

For the special case in which **V** is null, Froelich (2007) studied the asymptotic distribution theory of estimators of the IV functional that rely on two distinct non-parametric 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 non-parametric series regression. His estimators, however, suffer from the curse of dimensionality. If the dimension of **X** is large, as will be so in many applications 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 **X** that are sufficiently close to each other to allow for the borrowing of information that is needed for the smoothing implicit in these methods. Again, for the special case in which **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 that is proposed in this paper.

Neither Froelich (2007) nor Tan (2006a) addressed the case when **V** is a non-empty, strict subset of **X**, but further difficulties arise for each of their strategies in this case. Extending Froelich's approach to estimate the functionals IV(**V**) and MIV(**V**) non-parametrically not only requires smooth estimators of the aforementioned conditional means, but also of the conditional means given **V** of the differences that are involved in the numerators and denominators of these functionals. One possible extension of Tan's (2006a) fully parametric

approach along the lines proposed there 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(·) and MLATE(·) 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 that approach involves specifying working models that may not cohere with the assumed model for LATE(·).

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 $\mathbf{V} = \mathbf{X}$ 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 that are 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 conditionally 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 $E(Y_d|D_1 > D_0, \mathbf{X} = \cdot)$, $d = 0, 1$ owing 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($\mathbf{V}$):

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

and

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

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

For the special case in which $\mathbf{V} = \mathbf{X}$, Abadie (2003) 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 to estimate LATE($\mathbf{X}$) under a model that parameterizes just this contrast rather than under separate models for each of the counterfactual means 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 that were 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 is 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 $E\{\phi(\mathbf{X})|\mathbf{V} = \cdot\}$ of a user-specified function $\phi(\mathbf{X})$ (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

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

and

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

The issue of which curves must be modeled in the doubly robust procedure, i.e. those in Tan (2007), Uysal (2011) 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, $E\{\phi(\mathbf{X})|\mathbf{V} = \cdot\}$, $E(H_1|Z = \cdot, \mathbf{X} = \cdot)$ 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, $E\{\phi(\mathbf{X})|\mathbf{V} = \cdot\}$, $E(H_2|Z = \cdot, \mathbf{X} = \cdot)$ 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 $E\{\phi(\mathbf{X})|\mathbf{V} = \cdot\}$ 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 $E\{\phi(\mathbf{X})|\mathbf{V} = \cdot\}$ 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 and H_2 respectively with IV($\cdot$) and MIV($\cdot$) is a consequence of the fact that the restrictions imposed on the law of H_1 and H_2 by the IV assumptions do not depend on the functional form of IV($\cdot$) and MIV($\cdot$)respectively. In contrast, restrictions on $E(Y|Z = \cdot, \mathbf{X} = \cdot)$ and $E(D|Z = \cdot, \mathbf{X} =$

·) or on $E(Y|Z = \cdot, D = \cdot, \mathbf{X} = \cdot)$ and $E(D|Z = \cdot, \mathbf{X} = \cdot)$ impose restrictions on IV(·) and therefore may conflict with parametric specifications for it. One could reasonably argue that this conflict is of no great importance for practice because all models are almost certainly misspecified to a larger or smaller extent. Furthermore, it may be easier to build and check models for $E(Y|Z = \cdot, D = \cdot, \mathbf{X} = \cdot)$ and $E(D|Z = \cdot, \mathbf{X} = \cdot)$ than models for $E(H_j|Z = \cdot, \mathbf{X} = \cdot)$. Although we do not disagree with these points, we nevertheless find it conceptually important to demonstrate that there is methodology that, under the assumptions stated, is internally consistent.

3.1. Estimation of LATE(·) and MLATE(·) 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(**V**) and MIV(**V**) on which our proposed estimators rely.

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

$$E\{E(H_j|Z=1,\mathbf{X}) - E(H_j|Z=0,\mathbf{X})|\mathbf{V}\}=0 \qquad \text{with probability 1} \quad (5)$$

and

$$E\{(-1)^{1-Z}\, p(Z|\mathbf{X})^{-1}\, H_j|\mathbf{V}\}=0 \qquad \text{with probability 1}, \quad (6)$$

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 result (5) from definition (1) when $j = 1$ note that the difference between the numerator on the right-hand side of definition (1) and the product of IV(**v**) with the denominator on the right-hand side of definition (1) is the same as the left-hand side of result (5). Likewise, to arrive at result (5) from (2) when $j = 2$ note that the sum of the denominator on the right-hand side of definition (2) with the product of the numerator on the right-hand side of definition (2) times MIV(**v**)$^{-1}$ is the same as the left-hand side of equation (5). Equation (6) is equivalent to equation (5) because

$$E\{(-1)^{1-Z}p(Z|\mathbf{X})^{-1}H_j|\mathbf{V}\}=E[\{Z\ p(Z|\mathbf{X})^{-1}-(1-Z)p(Z|\mathbf{X})^{-1}\}H_j|\mathbf{V}], E\{Z\ p(Z|\mathbf{X})^{-1}H_j|\mathbf{V}\}=E\{E(H_j|Z=1,\mathbf{X})|\mathbf{V}\}$$

and

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

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(**V**) and $j = 2$ if β indexes the specification (4) for MLATE(**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(**v**) defined in equation (3) and for MLATE(**v**) defined in equation (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 estimator $\hat{\beta}_{\text{ipw}}$ of β_0 under a parametric class for the instrument probabilities

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

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

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

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 (e.g. $q(\mathbf{V};\beta) = \partial m_j(\mathbf{V};\beta)/\partial\beta$), and

$$\hat{\alpha}=\arg\max_{\alpha} E_n(\log[\pi(\mathbf{X};\alpha)^Z\{1-\pi(\mathbf{X};\alpha)\}^{1-Z}]) \quad (9)$$

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), $\sqrt{n}(\hat{\beta}_{\text{ipw}} - \beta_0)$ converges in law to a mean 0 normal distribution when condition (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 consistent and asymptotically normal 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 a parametric class, we re-express the constraint (5) as the condition that, for some $r(\mathbf{X})$,

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

When $\mathbf{V} \neq \mathbf{X}$ we derive a flexible parametric specification for $E(H_j \mid Z, \mathbf{X})$ that respects constraint (5) from the following three specifications:

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

$$r(\mathbf{X}) \in \mathscr{R} = \{\rho^T \varphi(\mathbf{X}) : \rho \in R^K\} \quad (10)$$

where $\varphi(\mathbf{X}) \equiv (\phi_1(\mathbf{X}), \ldots, \phi_K(\mathbf{X}))^T$ and ϕ_s, $s \in \{1, \ldots, K\}$, are user-specified real-valued functions,

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

$$E\{\varphi(\mathbf{X})|\mathbf{V}\} \in \mathscr{M} = \{\phi(\mathbf{V};\gamma) : \gamma \in \Gamma\} \quad (11)$$

where $\varphi(\mathbf{V}; \gamma) \equiv (\varphi_1(\mathbf{V}; \gamma), \ldots, \varphi_K(\mathbf{V}; \gamma))^T$, Γ is a subset of a Euclidean space and ϕ_k, $k \in \{1, \ldots, K\}$, are user-specified real-valued functions (when $\mathbf{V}$ is null we set $\varphi(\mathbf{V}; \gamma) = \gamma$, thus leaving $\mathscr{M}$ unrestricted) and

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

$$E(H_j|Z=0,\mathbf{X}) \in \mathscr{K} = \{k(\mathbf{X};\nu) : \nu \in \Upsilon\} \quad (12)$$

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

Specifications (10)–(12) imply the following model respects the constraint (5):

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

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

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

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

where $h(\cdot; \cdot)$ is a specified function that is 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 $E(H_j | Z, \mathbf{X} = \mathbf{x})$ does not depend on Z.

An estimator $\hat{\beta}_{\text{reg}}$ that is consistent and asymptotically normal 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 $(\hat{\beta}_{\text{reg}}, \hat{\eta})$ solving

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

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

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

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

$$l(Z,\mathbf{X};\beta,\eta,\gamma)=\begin{pmatrix} l_\eta(Z,\mathbf{X};\eta,\gamma) \\ l_\beta(Z,\mathbf{X};\beta)\end{pmatrix}=\begin{pmatrix} \partial h(Z,\mathbf{X};\eta,\gamma)/\partial\eta \\ Z\partial m(\mathbf{V};\beta)/\partial\beta\end{pmatrix}. \quad (16)$$

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

Selection of the parametric class for $E(H_j|Z, \mathbf{X})$ can be aided with the following α-level score type test of the null hypothesis $\mathbb{H}_0 : \eta_2 = 0$ where $\eta=\left(\eta_1^{\mathrm{T}}, \eta_2^{\mathrm{T}}\right)^{\mathrm{T}}$ and η_2 is of dimension, say, d_2. Let

$$R_n=E_n\left[\{\partial h(\tilde{\beta}_{\mathrm{reg}},\tilde{\eta}_1,\eta_2,\hat{\gamma})/\partial\eta_2|_{\eta_2=0}\}\varepsilon_j(\tilde{\beta}_{\mathrm{reg}},\tilde{\eta}_1,0,\hat{\gamma})\right]$$

where $(\tilde{\beta}_{\mathrm{reg}}, \tilde{\eta}_1)$ solves

$$E_n\{[\partial h(Z,\mathbf{X};\eta_1,0,\hat{\gamma})/\partial\eta_1^{\mathrm{T}}, l_\beta(Z,\mathbf{X};\beta)^{\mathrm{T}}]^{\mathrm{T}}\varepsilon_j(\beta,\eta_1,0,\hat{\gamma})\}=0.$$

Under $\mathbb{H}_0$, $\sqrt{n}R_n$ converges in law to a mean 0 d_2-variate normal distribution with variance–covariance matrix, say, J. Thus, if $\hat{\mathcal{J}}$ is a consistent estimator of J, a test that rejects $\mathbb{H}_0$ when $R_n^{\mathrm{T}}\hat{\mathcal{J}}R_n > \chi_{1-\alpha,d_2}$ where $\chi_{1-\alpha,d_2}$ is the $(1 - \alpha)$-quantile of a χ^2-distribution with d_2 degrees of freedom is an asymptotic α-level test of $\mathbb{H}_0$. A consistent variance estimator $\hat{\mathcal{J}}$ can be derived from standard Taylor expansion arguments for *M*-estimators (Stefanski and Boos, 2002).

3.2. Doubly robust estimation of LATE (·) and MLATE (·)

In this section we derive a doubly robust estimator $\hat{\beta}_{\mathrm{dr}}$ of β which satisfies that $\sqrt{n}(\hat{\beta}_{\mathrm{dr}} - \beta_0)$ converges to a mean 0 normal distribution under the IV assumptions and regularity conditions provided that one of the following two conditions (a) or (b) holds, even if both do not hold simultaneously:

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

b. specification (7) is correct.

The estimator $\hat{\beta}_{\text{dr}}$ solves the estimating equations

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

where, for each fixed β, $\hat{\eta}(\beta)$ solves $E_n\{l_\eta(Z, \mathbf{X}; \beta, \eta, \hat{\gamma})\hat{\varepsilon}_j(\beta, \eta, \hat{\gamma})\} = 0$ with l_η defined as in equation (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 $\hat{\beta}_{\text{dr}}$ is consistent for β_0 when (b) holds because

$$E\{q(\mathbf{V};\beta)(-1)^{1-Z}p(Z|\mathbf{X};\alpha_0)^{-1}a(\mathbf{X};\alpha,\eta,\gamma)\}=0$$

for all β since $E\{(-1)^{1-Z}\, p(Z|\mathbf{X}; \alpha_0)|\mathbf{X}\} = 0$.

In contrast, consistency when condition (a) holds can be seen after re-expressing equation (17) as

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

and noting that, by virtue of equality (5) of theorem 1, $E[q(\mathbf{V}; \beta)\{h(1, \mathbf{X}; \eta_0, \gamma_0) - h(0, \mathbf{X}; \eta_0, \gamma_0)\}] = 0$ and, by $E\{\varepsilon_j(\beta, \eta_0, \gamma_0)|Z, \mathbf{X}\} = 0$, $E\{b(Z, \mathbf{X})\, \varepsilon_j(\beta, \eta_0, \gamma_0)\} = 0$ 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 $\sqrt{n}(\hat{\beta}_{\text{dr}} - \beta_0)$ to a normal distribution follows after noticing that $(\hat{\beta}_{\text{dr}}, \hat{\eta}, \hat{\gamma}, \hat{\alpha})$ where $\hat{\eta} \equiv \hat{\eta}(\hat{\beta}_{\text{dr}})$ is an *M*-estimator, i.e. it solves a joint system of estimating equations. The accuracy of this asymptotic result in finite samples hinges on the strength of the instrument *Z*, i.e. on how close $\Delta(\mathbf{V}) = E\{E(D|Z = 1, \mathbf{X}) - E(D|Z = 0, \mathbf{X})|\mathbf{V}\}$ is to 0. Theoretical results exploring the asymptotic distribution of $\hat{\beta}_{\text{dr}}$ as $\Delta(\mathbf{V})$ shrinks to 0 at different rates with sample size, similarly to those in the conventional IV literature, should be explored but are beyond the scope of this paper.

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

In the special case of estimation of $\beta_0 \equiv$ LATE, i.e. when **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 they replaced $h\{Z, \mathbf{X}; \eta(\hat{\beta}), \hat{\gamma}\}$ with $\hat{E}(Y|Z, \mathbf{X}) - \beta\hat{E}(D|Z, \mathbf{X})$. Tan computed estimators $\hat{E}(Y|Z, \mathbf{X})$ and $\hat{E}(D|Z, \mathbf{X})$ under parametric models for $E(Y|D = d, Z = z, \mathbf{X} = \cdot)$ and $E(D|Z = z, \mathbf{X} = \cdot)$, $d, z = 0, 1$ whereas Uysal (2011) computed them under parametric models for $E(Y|Z = z, \mathbf{X} = \cdot)$ and $E(D|Z = z, \mathbf{X} = \cdot)$, $z = 0, 1$.

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

In addition to $\hat{\beta}_{\text{ipw}}$ and $\hat{\beta}_{\text{dr}}$, 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 $\hat{\beta}_{\text{S}}$ solving

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

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

$$s_{\text{opt},j}(\mathbf{X})=\{1 - \pi(\mathbf{X})\}E(H_j|Z=1, \mathbf{X})+\pi(\mathbf{X})E(H_j|Z=0, \mathbf{X}).$$

In the supplementary Web appendix we also show that, when the specifications (11), (13) and (7) hold if **V** is not equal to **X** or when the specifications (14) and (7) hold if **V** = **X**, the limiting distribution of $\sqrt{n}(\hat{\beta}_{\text{dr}} - \beta_0)$ has variance precisely equal to the bound $\Sigma_{q,s_{\text{opt},j}}$. The estimator $\hat{\beta}_{\text{dr}}$, however, may have asymptotic variance even larger than that of $\hat{\beta}_{\text{ipw}}$ if specification (11) and/or (13) is incorrect when $\mathbf{V} \neq \mathbf{X}$ or if specification (14) is incorrect when **V** = **X**. Using ideas similar to those in Tan (2006b, 2010) we can construct another doubly robust estimator $\tilde{\beta}_{\text{dr}}$ that remedies this flaw. The estimator $\tilde{\beta}_{\text{dr}}$ is computed by solving

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

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

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

with

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

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

A further result, which is derived in the supplementary Web appendix, establishes that for $j \in \{1, 2\}$ the optimal function $q_{\mathrm{opt},j}(\cdot)$, in the sense that $\Sigma_{q,s_{\mathrm{opt},j}} - \Sigma_{q_{\mathrm{opt},j},s_{\mathrm{opt},j}} \geqslant 0$ for any $q(\cdot)$, is

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

where

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

The optimal function $q_{\mathrm{opt},j}(\cdot)$ 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,

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

and

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

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 $\hat{\delta}$ and $\hat{\omega}$ by regressing $(-1)^{1-Z}p(Z|\mathbf{X};\hat{\alpha})^{-1}DY^{j-1}$ and $p(Z|\mathbf{X};\hat{\alpha})^{-2}[H_j(\hat{\beta}_{\mathrm{dr}})-a\{\mathbf{X};\hat{\alpha},\hat{\eta}(\hat{\beta}_{\mathrm{dr}}),\hat{\gamma}\}]^2$ on **V** under models (19) and (20) respectively, where $\hat{\beta}_{\mathrm{dr}}$ is a preliminary doubly robust estimator of β computed using an arbitrary $q(\mathbf{V};\beta)$. We then estimate $q_{\mathrm{opt},j}(\mathbf{V};\beta)$ with

$$\hat{q}_{\mathrm{opt},j}(\mathbf{V};\beta)\equiv-\{\partial m(\mathbf{V};\beta)/\partial\beta\}m_j(\mathbf{V};\beta)^{2(1-j)}e_j(\mathbf{V};\hat{\delta})t_j(\mathbf{V};\hat{\omega})^{-1}.$$

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

a. $\sqrt{n}(\hat{\beta}_{dr} - \beta_0)$, $\sqrt{n}(\tilde{\beta}_{dr} - \beta_0)$ and $\sqrt{n}(\tilde{\beta}_C - \beta_0)$ converge to mean 0 normal distributions with variances Σ_{dr}, $\Sigma_{better.dr}$ and Σ_C respectively. Furthermore, $\Sigma_{better.dr} - \Sigma_C \leq 0$ and $\Sigma_{better.dr} - \Sigma_{dr} \leq 0$.

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

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

A slight modification of the procedure for computing $\hat{\beta}_{dr}$ and $\tilde{\beta}_{dr}$ 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(·) curve is

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

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

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

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 (b) of Section 3.2 holds (i.e. when the propensity score specification (7) is correct), the estimators $\hat{\beta}_{dr}$ and $\tilde{\beta}_{dr}$ that use $q(\mathbf{V}; \beta) = q_w(\mathbf{V}; \beta) \equiv w(\mathbf{V})\partial m_1(\mathbf{V}; \beta)/\partial\beta$ converge in probability to $\beta_{w,0}$ even if the specification (3) is incorrect.

However, unfortunately, $\hat{\beta}_{dr}$ and $\tilde{\beta}_{dr}$ need not converge to $\beta_{w,0}$ for any w when the propensity score model is incorrect even if condition (a) of Section 3.2 holds. This happens essentially because equation (22) is equivalent to

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

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 consistent and asymptotically normal 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 $E\{H_1(\beta_{w,0})|Z, \mathbf{X}\} \in \mathscr{H}$ where $\mathscr{H}$ is defined in specification (13) with $j = 1$ since this specification is necessarily wrong if the model (11) is correct. This happens because $\mathscr{H}$ respects the constraint (5) but $E\{H_1(\beta_{w,0})|Z,$

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 $H_1 = Y - \mathrm{IV}(\mathbf{V})D$ satisfies the constraint (5) as this constraint identifies the IV(·) curve.

A slight modification to the class $\mathscr{H}$ 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 formulate a correctly specified model for $E\{H_1(\beta_{w,0})|Z, \mathbf{X}\}$. Specifically, the parametric specification

$$E\{H_1(\beta_{w,0})|Z=z,\mathbf{X}=\mathbf{x}\} \in \mathscr{H}_w=\{k(\mathbf{x};\nu)+\lambda^{\mathrm{T}}\{\varphi(\mathbf{x})-\theta q_w(\mathbf{v})\}z:\lambda \in R^K, \nu \in \Upsilon] \quad (24)$$

where $\phi(\cdot)$ and $k(\cdot;\cdot)$ are user-chosen functions as defined in Section 3.1 and

$$\theta=E\{\varphi(\mathbf{X})q_w(\mathbf{V})^{\mathrm{T}}\}E\{q_w(\mathbf{V})q_w(\mathbf{V})^{\mathrm{T}}\}^{-1} \quad (25)$$

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

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

- **a.** $\sqrt{n}(\hat{\hat{\beta}}_{\mathrm{dr}} - \beta_0)$ converges to a normal distribution if the parametric specification (3) for LATE(·) is correct and either condition (a) or condition (b) of Section 3.2 holds, and
- **b.** $\sqrt{n}(\hat{\hat{\beta}}_{\mathrm{dr}} - \beta_{w,0})$ converges to a normal distribution if the parametric specification (3) for LATE(·) is incorrect but either condition (b) of Section 3.2 or the parametric specification (24) holds.

Consider first the case $\mathbf{V} \neq \mathbf{X}$. The estimator $\hat{\hat{\beta}}_{\mathrm{dr}}$ solves equation (17) with $q_w(\mathbf{V};\beta)$ instead of $q(\mathbf{V};\beta)$, and with $a\{\mathbf{X};\alpha,\hat{\eta}(\beta),\gamma\}$ 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 $\eta = (\nu, \rho, \lambda)$,

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

$\hat{\eta}(\beta)$ solves

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

with

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

$\hat{\gamma}$ solves

$$E_n[\{\partial\phi(\mathbf{V};\gamma)^{\mathrm{T}}/\partial\gamma\}\{\phi(\mathbf{X})-\phi(\mathbf{V};\gamma)\}]=0$$

and

$$\hat{\theta}(\beta)\equiv E_n\{\varphi(\mathbf{X})q_w(\mathbf{V})^{\mathrm{T}}\}E_n\{q_w(\mathbf{V})q_w(\mathbf{V})^{\mathrm{T}}\}^{-1}.$$

When $\mathbf{V}=\mathbf{X}$, $\hat{\beta}_{\mathrm{dr}}$ is computed analogously except that ρ is set to 0 and γ is absent.

The desired properties (a) and (b) of the estimator $\hat{\beta}_{\mathrm{dr}}$ are deduced from the following considerations. When condition (b) holds, the estimator $\hat{\beta}_{\mathrm{dr}}$ is consistent and asymptotically normal for $\beta_{w,0}$ regardless of whether or not specification (3) holds because $E_n[q_w(\mathbf{V};\beta)(-1)^{1-Z}\times p(Z|\mathbf{X};\hat{\alpha})^{-1}b\{\mathbf{X};\beta,\hat{\alpha},\hat{\eta}(\beta),\hat{\gamma},\hat{\theta}(\beta)\}]$ converges to 0 in probability for all β. In contrast, the convergence of $\hat{\beta}_{\mathrm{dr}}$ to β_0 when specification (3) and condition (a) hold, and the convergence of $\hat{\beta}_{\mathrm{dr}}$ to $\beta_{w,0}$ when specification (3) is incorrect but condition (24) holds follows arguing as in Section 3.2 for the convergence of $\hat{\beta}_{\mathrm{dr}}$ to β_0 when condition (a) holds, after noting that the class

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

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

An estimator $\tilde{\beta}_{\mathrm{dr}}$ satisfying properties (a) and (b) and additionally guaranteed to be at least as efficient asymptotically as $\hat{\beta}_{\mathrm{ipw}}$ is constructed just like $\tilde{\beta}_{\mathrm{dr}}$ in Section 3.2 but replacing $a\{\mathbf{X};\hat{\alpha},\hat{\eta}(\beta),\hat{\gamma}\}$ with $b\{\mathbf{X};\beta,\hat{\alpha},\hat{\eta}(\beta),\hat{\gamma},\hat{\theta}(\beta)\}$, $q(\mathbf{V};\beta)$ with $q_w(\mathbf{V};\beta)$ and $h\{Z,\mathbf{X};\hat{\eta}(\beta),\hat{\gamma}\}$ with $h_w\{Z,\mathbf{X};\beta,\eta,\gamma,\hat{\theta}(\beta)\}$. In the supplementary Web appendix we also describe an estimator $\hat{\beta}_{\mathrm{opt,dr}}$ which satisfies property (a) and has limiting normal distribution with variance equal to $\Sigma_{q_{\mathrm{opt},1},s_{\mathrm{opt},1}}$ when conditions (a) and (b) of Section 3.2 hold and yet converges to a weighted least squares approximation when the specification (3) for LATE(**V**) is wrong.

For estimation of the MLATE(·) curve in the supplementary Web appendix we show that the estimator $\hat{\beta}_{\mathrm{dr}}$ that is computed by using $H_2(\beta)$ instead of $H_1(\beta)$ and with $q_w(\mathbf{V};\beta)$ redefined

as $m_2(\mathbf{V};\beta)\{\partial m_2(\mathbf{V};\beta)/\partial\beta\}w(\mathbf{V})$ satisfies properties (a) and (b) where in the statements of these properties, specifications (3) and (24) are replaced with specification (4) and the specification that $E\{H_2(\beta_{w,0})|Z = z, \mathbf{X} = \mathbf{x}\} \in \mathcal{H}_w$ respectively, and $\beta_{w,0}$ is redefined as

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

with $e_0(\mathbf{v}) \equiv E(Y_0|D_1 > D_0, \mathbf{V} = \mathbf{v})$. Note that, unlike 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(·) 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

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

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 (a)–(d) and (f) in Section 2.1 and specific restrictions on ATT(·, ·). In particular, Robins (1994) showed that, when $\mathbf{V} = \mathbf{X}$, ATT(z, $\mathbf{v}$) is identified under assumptions (a)–(d) and (f), and the no additive treatment–instrument interaction on the treated: ATT(z, $\mathbf{v}$) = ATT($\mathbf{v}$) does not depend on z (assumption v-ATT).

Remarkably, Robins showed that under these assumptions ATT($\mathbf{v}$) = 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 (a)–(d) and (f) in Section 2.1, the structural interpretation of the observed data functional IV($\mathbf{v}$) depends on which of the assumptions (e) 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 a 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 that is addressed in this paper. For short, refer to the model defined by assumptions (a)–(f) in Section 2.1 as ‘our additive model’ and to the model defined by assumptions (a)–(d), (f) 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 (a)–(f) in Section 2.1 and v-ATT, i.e. the model that makes simultaneously the assumptions of our additive model and of the Robins–Tan additive model, LATE(**v**) and ATT(**v**) are indeed identical causal effect contrasts;
- **b.** our model is statistically indistinguishable from the intersection model, i.e., 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 that are 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(**v**) = ATT(**v**) under the intersection model if and only if it is equal to LATE(**v**) under our additive model. If this were not so, there would be some observed data law functional equal to LATE(**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 a case, there would be a restriction, specifically the restriction that sets the new functional equal to LATE(**v**), that would be satisfied under the intersection model but not under our additive model, thus contradicting result (b).

Result (c) implies that a functional of the observed data law is equal to ATT(**v**) under the intersection model if and only if it is equal to ATT(**v**) under the Robins–Tan additive model. If this were not so, the intersection model would satisfy an equality constraint that is not satisfied by the Robins–Tan additive model, namely the constraint that sets a new functional of the observed data law equal to ATT(**v**), thus contradicting result (c).

Results (a)–(c) then imply that any functional of the observed data law that is equal to ATT(**v**) under the Robins–Tan model must be equal to LATE(**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(**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(**v**) under our additive model.

A further result (result (d) stated below) implies that IV(**v**) is indeed the only functional of the observed data law that is equal to LATE(**v**) under our additive model and, consequently, the only observed data functional equal to ATT(**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(**v**) can be equal to LATE(**v**) under our additive model. If this were not so, then the observed data law would satisfy an equality constraint under our model, namely the equality that sets IV(**v**) equal to the other functional that agrees with LATE(**v**), thus contradicting result (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 (a)–(f) in Section 2.1 and the additional restriction

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

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

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

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

Result (a) follows because restriction (27) implies that ATT(**v**) $\equiv E(Y_1 - Y_0|T \in \{\text{co, at}\}, \mathbf{V} = \mathbf{v}) = E(Y_1 - Y_0|T = \text{co}, \mathbf{V} = \mathbf{v}) \equiv$ LATE(**v**), so under the intersection model, LATE(**v**) is indeed equal to ATT(**v**). Result (b) follows because, under assumptions (a)–(f), a test of the intersection model is a test that restriction (27) holds. No test can be constructed with power to detect departures from equation (27) because $E(Y_0|T = \text{at}, \mathbf{V})$ 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 $y < y'$,

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

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

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

Lemma 2. The only restrictions on the observed data law that are 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,

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

and under our additive model,

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

Abadie (2003) has previously derived result (32) in the special case **V** = **X** under our additive model. Observe that only under the Robins–Tan additive model and only for the special case **V** = **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*, **X**) = ATT(**X**) when **V** = **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 paper.

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

5. Data Analysis

We apply the procedures that are 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 USA. 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 to be 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) analysed data from the US Census Bureau's 1991 Survey of Income and Program Participation to test whether participation in 401(k) plans increases household savings. Here we reanalyze the data that were analyzed by Abadie

(2003), consisting of a sample of 9725 household reference subjects aged 25–64 years and their spouses, with annual income between $10000 and $200000. 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 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 thousands of dollars).

In this example, the instrumentation assumption (d) and monotonicity assumption (e) 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 (b) 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 behaviour operates only through the employee's choice to participate or not in the programme (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 these survey 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 shall 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 behaviour 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, probably 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 these data in Abadie (2003).

Table 1 reports the estimators of β with their bootstrap standard errors in parentheses in the case $\mathbf{V} = X_4$ under the specification $m(X_4;\beta) = \beta_0 + \beta_1 X_4$. Table 1 reports results for eight estimators: five doubly robust estimators $\hat{\beta}_{\mathrm{dr}}$, two IPW estimators $\hat{\beta}_{\mathrm{ipw}}$ and one outcome regression estimator $\hat{\beta}_{\mathrm{reg}}$. The estimator $\hat{\beta}_{\mathrm{reg}}$ was computed by using the function $l(Z, \mathbf{X}; \beta, \eta, \gamma)$ given in expression (16). Three of the doubly robust estimators, denoted by $\hat{\beta}^{\mathrm{opt}}_{\mathrm{dr}}$, $\hat{\beta}^{\mathrm{opt}}_{\mathrm{dr},\pi-\mathrm{fixed}}$, $\hat{\beta}^{\mathrm{opt}}_{\mathrm{dr},h-\mathrm{fixed}}$, used $q(\mathbf{V})$ equal to $\hat{q}_{\mathrm{opt},1}(\mathbf{V})$ as defined in Section 3.3. In the calculation of $\hat{q}_{\mathrm{opt},1}(\mathbf{V})$, $\log[e_1(\mathbf{V};\delta)/\{1 - e_1(\mathbf{V};\delta)\}]$ and $\log\{t_1(\mathbf{V};\omega)\}$ were linear functions of income and income^2. (When, as in this data set, $Z = 0$ 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, which is denoted with $\hat{\beta}^{\mathrm{ineff}}_{\mathrm{dr}}$, used $q(\mathbf{V}) = \partial m(\mathbf{V};\beta)/\partial\beta = (1, X_4)^{\mathrm{T}}$ and the last doubly robust estimator, which is denoted by $\hat{\beta}^{\mathrm{ineff,stable}}_{\mathrm{dr}}$ used $q(\mathbf{V}) = (1, X_4)^{\mathrm{T}}\{\mathrm{expit}(\hat{\zeta}_0 + \hat{\zeta}_1 X_4) - \mathrm{expit}(\hat{\zeta}_0 + \hat{\zeta}_1 X_4)^2\}$ where $\mathrm{expit}(\hat{\zeta}_0 + \hat{\zeta}_1 X_4)^{\mathrm{dr}}$ was the fitted value from a logistic regression of Z on X_4. These last two

choices of $q(\mathbf{V})$ were also used to construct the two IPW estimators, which are denoted by $\hat{\beta}_{\text{ipw}}^{\text{ineff}}$ and $\hat{\beta}_{\text{ipw}}^{\text{ineff,stable}}$ 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 $\hat{\beta}_{\text{reg}}$, which is denoted in what follows by $\mathcal{H}_\upsilon^{k_h}$, assumed that $E\{H_1(\beta_0)|Z,\mathbf{X}\} = k(\mathbf{X};\nu) + \rho^{\mathrm{T}}\{\phi(\mathbf{X}) - \phi(\mathbf{V};\gamma)\}Z$. 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 $\phi(\mathbf{x})$ was a vector of indicators of combined levels of age, marital status and family size; each entry of $\phi(\mathbf{v};\gamma)$ was a linear logistic regression model for the corresponding entry of $\phi(\mathbf{x})$ with covariates being income, income^2, …, income^{kh}. The estimators $\hat{\beta}_{\text{dr}}^{\text{opt}}$, $\hat{\beta}_{\text{dr}}^{\text{ineff}}$ and $\hat{\beta}_{\text{dr}}^{\text{ineff,stable}}$ were computed by using models ($\mathcal{P}^{k_\pi}$ and $\mathcal{H}_\upsilon^{k_h}$ with $k_\pi = k_h \equiv k$. In Table 1 the first three rows report these estimators by using k as indicated by the column labels. The estimator $\hat{\beta}_{\text{dr},\pi-\text{fixed}}^{\text{opt}}$ had k_π fixed at 4 and k_h as indicated by the column labels. Likewise the estimator $\hat{\beta}_{\text{dr},h-\text{fixed}}^{\text{opt}}$ had k_h fixed at 4 and k_π as indicated by the column labels. The estimators $\hat{\beta}_{\text{ipw}}^{\text{ineff}}$ and $\hat{\beta}_{\text{ipw}}^{\text{ineff,stable}}$ had k_π as indicated by the column labels. Finally, the estimator $\hat{\beta}_{\text{reg}}$ had k_h as indicated by the column labels. In the data set 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 1% of the estimated values of $\pi(\mathbf{X};\hat{\alpha})$, and finally carried through the entire procedure for arriving at the estimators of β using the remaining data. In the data set, 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 that is presented in this paper, $\hat{\beta}_{\text{dr}}^{\text{opt}}$ with $k_\pi = k_h$ sufficiently large should result in optimal inference about β. We therefore first examine the rows corresponding to $\hat{\beta}_{\text{dr}}^{\text{opt}}$ 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 effect 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 \$50000 who chooses to participate in the programme is to increase her family's net financial assets by \$14910 whereas the increase for a person with an income of \$100000 is \$31310.

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 behaviour of the remaining estimators of the income coefficient. Since the three doubly

robust estimators $\hat{\beta}_{\mathrm{dr}}^{\mathrm{opt}}$, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff}}$ and $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff,stable}}$ with $k_\pi = k_h \geqslant 4$ 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 $\hat{\beta}_{\mathrm{dr}}^{\mathrm{opt}}$, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff}}$ and $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff,stable}}$ 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 shall 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 $\hat{q}_{\mathrm{opt},1}(\mathbf{V})$ 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 $\hat{\beta}_{\mathrm{reg}}$; presumably this reflects the fact that the choice (16) that we recommended for ease of calculation is not optimal). A comparison of the IPW estimators with the estimator $\hat{\beta}_{\mathrm{dr},h-\mathrm{fixed}}^{\mathrm{opt}}$ and of the outcome regression estimator with $\hat{\beta}_{\mathrm{dr},\pi-\mathrm{fixed}}^{\mathrm{opt}}$ illustrates the advantage of doubly robust estimation over IPW and outcome regression estimation. These comparisons reveal that doubly robust estimators require only 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 $\hat{\beta}_{\mathrm{dr},h-\mathrm{fixed}}^{\mathrm{opt}}$ that uses the same model for the propensity score but a model $\mathscr{H}_v^{k_h}$ with $k_h = 4$ is roughly 'unbiased'. Likewise, the outcome regression estimator that has k_h equal to 1 or 2 is biased but the 'bias' is corrected by the estimator $\hat{\beta}_{\mathrm{dr},\pi-\mathrm{fixed}}^{\mathrm{opt}}$.

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 $\boldsymbol{D}$ among the compliers. Additional evidence for misspecification is presented in Fig. 1, which displays the values of three different doubly robust estimators $\hat{\beta}_{\mathrm{dr}}$, denoted with $\hat{\beta}_{\mathrm{dr}}^{\mathrm{opt}}$, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff}}$ and $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff,stable}}$ which used respectively $q(\mathbf{X}) = e_1(\mathbf{X}; \delta)\hat{t}_1(\mathbf{X}; \hat{\omega})$, $q(\mathbf{X}) = \partial m(\mathbf{X}; \beta)/\partial\beta = 1$ and $q(\mathbf{X}) = \pi(\mathbf{X}, \hat{\alpha}) - \pi(\mathbf{X}, \hat{\alpha})^2$, 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 $\mathscr{P}^{k_\pi}$ for the propensity score and an outcome regression model $\mathscr{H}_x^{k_h}$ 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 $\hat{\beta}_{\mathrm{dr}}^{\mathrm{opt}}$, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff}}$ and $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff,stable}}$ as $k_h = k_\pi \equiv k$ varies from 1 to 8. Each estimator stabilizes for $k \geqslant 3$; 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 $\mathscr{P}^{k_\pi}$ 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 $\mathscr{P}^{k_\pi}$ is correct and the model $m(\mathbf{X}; \beta)$ for LATE($\mathbf{X}$) is misspecified, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff}}$, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff,stable}}$ and $\hat{\beta}_{\mathrm{dr}}^{\mathrm{opt}}$ converge in probability to distinct values $\beta_{0,w_{\mathrm{ineff}}}$, $\beta_{0,w_{\mathrm{ineff,stable}}}$ and $\beta_{0,w_{\mathrm{opt}}}$ where $w_{\mathrm{ineff}}(\mathbf{X}) = 1$, $w_{\mathrm{ineff,stable}}(\mathbf{X}) = \pi(\mathbf{X}, \alpha_0) - \pi(\mathbf{X}, \alpha_0)^2$ and $w_{\mathrm{opt}}(\mathbf{X}) = e_1(\mathbf{X}; \delta^*)t_1(\mathbf{X}; \omega^*)$ with δ^* and ω^* the probability limits of $\hat{\delta}$ and $\hat{\omega}$.

The parameter $\beta_{0,w_{\text{ineff}}}$ is of particular interest as an easy calculation shows that $\beta_{0,w_{\text{ineff}}}$ is equal to the marginal LATE, i.e. to $\beta_{\text{null}} \equiv \text{LATE}(\mathbf{V})$ when $\mathbf{V}$ = null. Thus, the estimator $\hat{\beta}_{\text{dr}}^{\text{ineff}}$ converges to β_{null} when the model $\mathcal{P}^{k_\pi}$ is correct. In fact, the IPW estimator $\hat{\beta}_{\text{ipw}}^{\text{ineff}}$ that uses the same $q(\mathbf{X})$ as $\hat{\beta}_{\text{dr}}^{\text{ineff}}$ 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 $\hat{\beta_{\text{dr}}}$ and $\hat{\beta_{\text{ipw}}}$ 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(·) is correct. These theoretical results are confirmed in Fig. 2, which displays the estimators $\hat{\beta}_{\text{ipw}}^{\text{ineff}}$ and $\hat{\beta}_{\text{dr}}^{\text{ineff}}$ computed under model $\mathcal{P}^{k_\pi}$ and model $\mathcal{H}_x^{k_h}$ with $k_h = k_\pi = k$. In addition, Fig. 2 displays the doubly robust estimator $\hat{\beta_{\text{null,dr}}}$ of β_{null}, i.e. of the marginal LATE. This estimator is computed under model $\mathcal{P}^{k_\pi}$ and a model $\mathcal{H}_{\text{null}}^{k_h}$ that assumes that $E\{H_1(\beta_{\text{null}})|Z, \mathbf{X}\} = k(\mathbf{X};\ \nu) + \rho^{\text{T}}[\phi(\mathbf{X}) - E\{\phi(\mathbf{X})\}]Z$ with $k(\mathbf{x};\ \nu)$ as defined earlier and $\phi(\mathbf{x})$ 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 Fig. 2 $\hat{\beta}_{\text{ipw}}^{\text{ineff}}$ and $\hat{\beta}_{\text{dr}}^{\text{ineff}}$ are both close to $\hat{\beta_{\text{null,dr}}}$ for $k_\pi \geqslant 4$.

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

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 **V**, where **V** may be a subset of the covariates **X** required for the candidate instrument to be a valid IV 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 we argued that the correspondence between the two models is unsurprising because the restrictions on the observed data law that is 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(·) are not variation independent. Thus, the model $m_2(\cdot; \beta)$ could conflict with a proposed model for $E(H_2|Z, \mathbf{X})$. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

References

Abadie A. Bootstrap tests for distributional treatment effects in instrumental variable models. J Am Statist Ass. 2002; 97:284–292.

Abadie A. Semiparametric Instrumental Variable Estimation of Treatment Response Models. J Econometrics. 2003; 113:213–263.

Abadie A, Angrist JD, Imbens GW. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica. 2002; 70:91–117.

Angrist JD, Graddy K, Imbens GW. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Rev Econ Stud. 2000; 67:499–527.

Angrist JD, Imbens GW. Average causal response with variable treatment intensity. J Am Statist Ass. 1995; 90:431–442.

Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental Variables (with discussion). J Am Statist Ass. 1996; 91:444–471.

Cheng J, Qin J, Zhang B. Semiparametric estimation and inference for distributional and general treatment effects. J R Statist Soc B. 2009a; 71:881–904.

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

Clarke P, Windmeijer F. Identification of causal effects on binary outcomes using structural mean models. Biostatistics. 2010; 11:756–770. [PubMed: 20522728]

Froelich M. Nonparametric IV estimation of local average treatment effects with covariates. J Econometrics. 2007; 139:35–75.

Gill RD. Non- and semi-parametric maximum likelihood estimators and the Von Mises method (Part 1). Scand J Statist. 1989; 16:97–128.

Heckman J. The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. Ann Econ Soc Measmnt. 1976; 5:475–492.

Hirano K, Imbens GW, Rubin DB, Zhou XH. Assessing the effect of an Influenza vaccine in an encouragement design. Biostatistics. 2000; 1:69–88. [PubMed: 12933526]

Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. Econometrica. 1994; 62:467–475.

Kasy M. Semiparametrically efficient estimation of conditional instrumental variables parameters. Int J Biostatist. 2009; 5 Article 22.

Little RJ, Yau LHY. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. Psychol Meth. 1998; 3:147–159.

Newey W. The asymptotic variance of semiparametric estimators. Econometrica. 1994; 62:1349–1382.

Poterba, JM.; Venti, SF.; Wise, DA. 401(k) Plans and Tax-Deferred Savings. In: Wise, D., editor. Studies in the Economics of Aging. Chicago: University of Chicago Press; 1994. p. 105-138.

Poterba JM, Venti SF, Wise DA. Do 401(k) Contributions Crowd Out Other Personal Saving? J Public Econ. 1995; 58:1–32.

Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Communs Statist Theor Meth. 1994; 23:2379–2412.

Robins JM, Hernan MA. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006; 17:360–372. [PubMed: 16755261]

Robins, JM.; Rotnitzky, A. Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In: Jewell, N.; Dietz, K.; Farewell, V., editors. Aids Epidemiology: Methodological Issues. Boston: Birkhäuser; 1992. p. 297-331.

Stefanski LA, Boos DD. The Calculus of M-Estimation. Am Statistn. 2002; 56:29–38.

Tan Z. Regression and Weighting Methods for Causal Inference Using Instrumental Variables. J Am Statist Ass. 2006a; 101:1607–1618.

Tan Z. A Distributional approach for causal inference using propensity scores. J Am Statist Ass. 2006b; 101:1619–1637.

Tan Z. Marginal and Nested Structural Models Using Instrumental Variables. J Am Statist Ass. 2010; 105:157–169.

Uysal, SD. Department of Economics, University of Konstanz, Konstanz; 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 EJ. Independence, monotonicity, and latent index models: an equivalency result. Econometrica. 2002; 70:331–341.

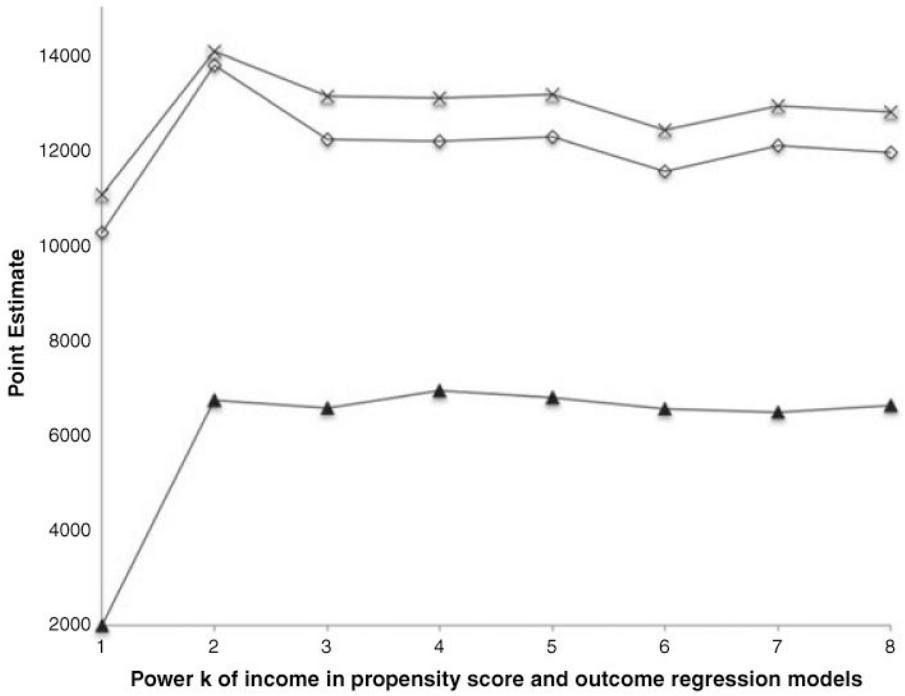

Fig. 1.

Estimation of the marginal LATE based on incorrectly assuming that $\mathrm{LATE}(\mathbf{X})=\mathrm{LATE}$âŰš, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{opt}}$;◇, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff}}$;×, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff,stable}}$

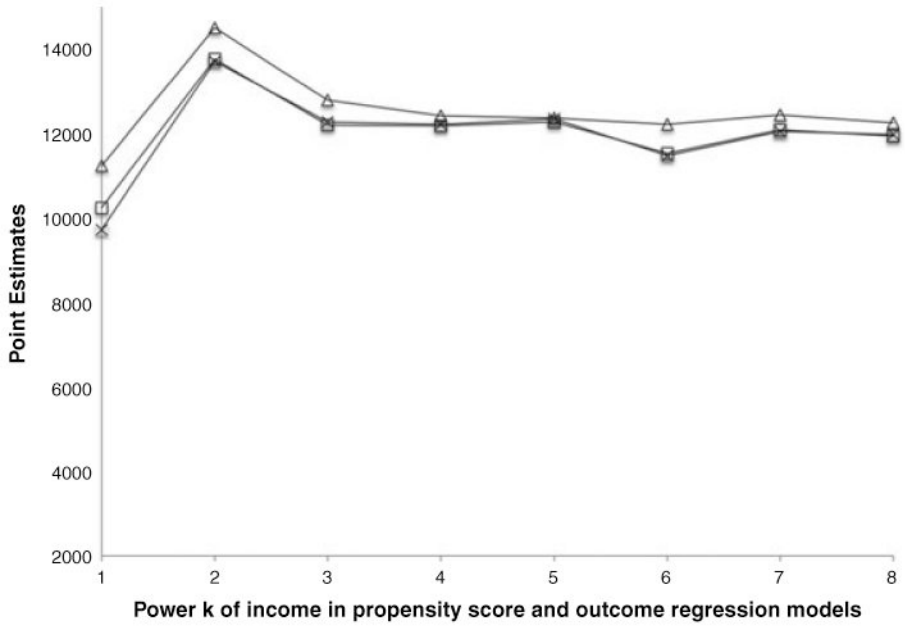

Fig. 2.

Doubly robust estimation of the marginal LATE *versus* estimation based on incorrectly assuming that $\mathrm{LATE}(\mathbf{X})=\mathrm{LATE}\Delta$, $\hat{\beta}_{\mathrm{ipw}}^{\mathrm{ineff}};\square$, $\hat{\beta}_{\mathrm{dr}}^{\mathrm{ineff}};\times$, $\hat{\beta}_{\mathrm{null,dr}}$

Table 1

Estimators of (β_0,β_1) and their bootstrap standard errors under model LATE(income) =β_0 + β_1 income

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

Parameter	Results for the following powers k of income in the outcome regression and propensity score models:					
	1	**2**	**3**	**4**	**5**	**8**
$\hat{\beta}_{\text{ipw}}^{\text{ineff}}$	−139 (218)	785 (306)	425 (178)	320 (181)	347 (181)	311 (201)
$\hat{\beta}_{\text{ipw}}^{\text{ineff,stable}}$	14 (161)	385 (154)	368 (119)	339 (117)	336 (114)	329 (123)
$\hat{\beta_{\text{reg}}}$	510 (187)	272 (210)	361 (181)	345 (183)	357 (180)	353 (194)

Table 2

Estimation of the marginal LATE effect

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

† $\hat{\beta}_{\text{dr}}$ is the estimator of Section 3.4 that uses $q_W(\mathbf{V})=1$.

‡ Test statistic is the difference of the estimators in the first and second rows divided by the bootstrap standard error of the difference.