*University of California, Berkeley*
U.C. Berkeley Division of Biostatistics Working Paper Series

*Year* 2016 *Paper* 350

# TMLE for Marginal Structural Models Based on an Instrument

Boriska Toth* Mark J. van der Laan†

*University of California, Berkeley, Division of Biostatistics, bori@stat.berkeley.edu

†University of California, Berkeley, Division of Biostatistics, laan@berkeley.edu



http://biostats.bepress.com/ucbbiostat/paper350

# TMLE for Marginal Structural Models Based on an Instrument

Boriska Toth and Mark J. van der Laan

**Abstract**

We consider estimation of a causal effect of a possibly continuous treatment when treatment assignment is potentially subject to unmeasured confounding, but an instrumental variable is available. Our focus is on estimating heterogeneous treatment effects, so that the treatment effect can be a function of an arbitrary subset of the observed covariates. One setting where this framework is especially useful is with clinical outcomes. Allowing the causal dose-response curve to depend on a subset of the covariates, we define our parameter of interest to be the projection of the true dose-response curve onto a user-supplied working marginal structural model. We develop a targeted minimum loss-based estimator (TMLE) of this estimand. Our TMLE can be viewed as a generalization of the two-stage regression method in the instrumental variable methodology to a semiparametric model with minimal assumptions. The asymptotic efficiency and robustness of this substitution estimator is outlined. Through detailed simulations, we demonstrate that our estimator's finite-sample performance can beat other semiparametric estimators with similar asymptotic properties. In addition, our estimator can greatly outperform standard approaches. For instance, the use of data-adaptive learning to achieve a good fit can lead to both lower bias and lower variance than for an incorrectly specified parametric estimator. Finally, we apply our estimator to a real dataset to estimate the effect of parents' education on their infant's health.

## 1. Introduction

When estimating a causal effect in an observational study, the problem of unmeasured confounding is a pervasive caveat. It is similarly problematic in inferring a causal effect of a treatment in an experiment where the treatment isn't fully randomized. A classic solution for obtaining a consistent estimate is to use an instrumental variable, assuming one exists. Informally, an instrumental variable, or instrument, affects the outcome only through its effect on the treatment, and the residual (error) term of the instrument is uncorrelated with the residual term of the outcome (Imbens and Angrist 1994, Rubins et al. 1996). Thus, the instrument produces exogenous variation in the treatment.

Instrumental variables have been used in a number of works in biometrics and biostatistics to obtain consistent estimates of a treatment effect. (See (Brookhart et al 2010) for a large collection of references.) They are a basic tool for inferring the causal effect of a clinical treatment or a medication on a health outcome, as large-scale randomization of patients is typically not feasible. In these settings, the instrumental variable is usually some attribute that is related to the health care a patient receives, but is not at the level of individual patients. Thus, the instrument is not confounded by factors affecting an individual's response to treatment. For example, (Brookhart and Schneeweiss 2007) use physician's preference for the treatment (non-steriodal anti-inflammatory medications) as the instrument in establishing the effect on gastrointestinal bleeding. (Newhouse and McClellan 1998) exploit regional variation in the availability of catheterization and revascularization procedures as their instrument in estimating the effect of these procedures on reducing mortality in heart attack patients. Another important setting for instrumental variables in health research is when the treatment is randomly assigned, but non-compliance is significant. Then the random treatment assignment serves as an ideal instrument. (van der Laan et al 2007) describe this setting.

This work goes beyond the usual estimation problem solved using an instrument, which is to estimate a single local average treatment effect. Instead, we estimate how the causal effect depends on any subset $V$ of baseline covariates $W$. Thus, we are able to estimate heterogeneous treatment effects. Specifically, we take the expected causal effect given $V$, and project that function of $V$ unto a user-supplied parametric working model. The usefulness of estimates of the treatment effect is greatly enhanced in clinical settings when the estimates can be conditional on a patient's individual characteristics and biomarkers. Furthermore, it is important that when estimating the causal effect as a function of covariates $V$, that $V$ can be a strict subset of all baseline covariates $W$. Medical data often involves a large space of covariates, and conditioning on many covariates in estimating relevant components of the data-generating distribution can be helpful in: 1) decreasing the variance of estimated conditional means, and 2) ensuring that the instrument induces exogenous variation given the covariates. However, a physician typically has a smaller set of patient variables that are available and that s/he considers reliable predictors. Thus the causal effect as a function of an arbitrary subset of baseline covariates is of great interest.

While instrumental variables are widely used to infer causal effects, the majority of studies make use of strong assumptions about the structure of the data and typically rely on parametric assumptions (Terza et al. 2008). In contrast, this work uses semiparametric modelling. Beyond the criteria that there is a valid instrument, we make use of the single structural assumption that the expected value of the outcome is linear in the treatment, conditional on the covariates. This assumption is used in virtually all similar works; however, as we discuss below, even this single assumption we make can be weakened.

We use targeted minimum loss estimation (TMLE), which is a methodology for semiparametric estimation that has very favorable theoretical properties and can be superior to other estimators in practice (van der Laan and Rubin 2006, van der Laan and Rose 2011). The

TMLE procedure targets only those components of the data-generating distribution that are relevant to the statistical parameter of interest. Initial estimates are formed of certain components, by data-adaptively learning on a library of prediction algorithms. The initial estimates are then fluctuated one or more times in a direction that removes bias and optimizes for semiparametric efficiency.

The TMLE method has a robustness guarantee: it produces consistent estimates even when the functional form is not known for all relevant components. We discuss the most common such scenario: when the conditional distribution of the outcome cannot be estimated consistently, and one only has information about the form of the distributions generating the instrument and treatment. TMLE also guarantees asymptotic efficiency when all relevant components and nuisance parameters are consistently estimated. Thus, under certain conditions, the TMLE estimator is optimal in having the asymptotically lowest variance for a consistent estimator in a general semiparametric model, thereby achieving the semiparametric Cramer-Rao lower bound (Newey 1990).

TMLE has the advantage over other semiparametric efficient estimators that it imposes constraints that ensure that the estimator matches the data well. It is a substitution estimator, meaning that the final estimate is made by evaluating the parameter of interest on the estimates of its relevant components, where these estimates respect the bounds on their parameter space. These properties have been linked to good performance in sparse data in (Gruber and van der Laan 2010), while we demonstrate performance gains over other estimators in continuous data having sharp boundaries in section 5.3.2.

In section 3, we give a general model for the setting of estimating the effect of a treatment on an outcome in the presence of an instrumental variable and both measured and unmeasured confounders. We use Pearl's model of counterfactual variables to meaningfully define the causal effect of the treatment (Pearl 2000, see also Rubin 1974). In Appendix 2, we derive

the efficient influence curve for the statistical parameter of interest in several settings, and in section 4, we give TMLE-based procedures for estimating the causal effect. Next, we establish the comparative performance of our estimator in section 5 through simulations, studying for instance: 1) how performance compares to well known approaches, including semiparametric and parametric methods; 2) the bias-variance tradeoff in using a higher variance, instrumental variable-based estimate over a biased estimate. Finally, section 6 presents an application to the (Chou et al 2010) dataset on the effect of parents' education on their infant's health.

## 2. Review of existing methods

Let $W$ be a vector of baseline covariates, and $m(W)$ denote the marginal causal effect of treatment given $W$. Most prior work on estimating the marginal causal effect of a treatment using an instrument deal with either the case where a scalar average effect $E(m(W))$ is estimated, or the entire curve $m(W)$ is estimated. In contrast, our work estimates $E(m(W)|V)$ for $V$ possibly a strict subset of $W$. (Tan 2010) is another work that lets $V$ be any subset of $W$ and gives estimators for the marginal effect of the treatment on $Y$, conditional on $V$ and level of treatment. However, their marginal effect is assumed to take a parametric form.

(Ogburn et al) is a recent work that also proposes a semiparametric estimator for the marginal causal effect given a strict subset of the covariates $V \subseteq W$[1]. They also present an estimator for the best least-squares projection of the true causal effect unto a parametric working model. Their estimators use the method estimating equations, and are efficient and double robust, but are not substitution estimators. In addition, (Ogburn et al) restrict attention to the case of a binary instrument and treatment, and make slightly stronger

[1]Ogburn et al's work was accepted for publication around the time this work was submitted for publication.

assumptions about the instrument than we do (for instance, they assume no confounding between the instrument and treatment).

(Abadie 2003) gives an estimator for the treatment effect in compliers as a function of $W$. However, the instrument propensity score $P(Z|W)$ must be estimated consistently in his approach. Both (van der Laan et al 2007) and (Robins 2004) present semiparametric, consistent, and locally efficient estimators for the effect of treatment on an outcome, as a function of covariates $W$, as motivated by the setting where $Z$ is the randomized assignment to a binary treatment, and $A$ is the binary compliance with treatment. The counterfactual outcomes are assumed to follow a parametric form $E(Y(A=0)|W,Z,A) = \tilde{m}(W,Z,A)$. The former work gives a solution for binary outcomes using the method of estimating equations, so that their estimator is double robust to misspecification of either $\Pr(Z|W)$ or $E(Y(A=0)|W,Z,A)$.

For the special case of a null $V$ where a scalar average effect is estimated, semiparametric efficient approaches abound (see for instance: Cheng et al 2009; Hong and Nekipelov 2010; Kasy 2009). (Uysal 2011) and (Tan 2006) describe doubly robust estimators, where either the propensity score $Pr(Z|W)$, or the conditional means given the instrument, must be correctly specified.

## 3. The model and causal parameter of interest

We use the notation that $P_0$ and $E_0$ refer to the true probability distribution and expectation, respectively, and $P_n$ and $E_n$ the empirical counterparts. We observe n i.i.d. copies $O_1, \ldots, O_n$ of a random variable $O = (W, Z, A, Y) \sim P_0$, where $P_0$ is its probability distribution. Here $W$ denotes the measured baseline covariates, and $Z$ denotes the subsequently (in time) realized instrument that is believed to only affect the final outcome $Y$ through the intermediate treatment variable $A$. The goal of the study is to assess a causal effect of treatment $A$ on outcome $Y$. We consider the case in which it is believed that $A$ is a function of both the

measured $W$ and also unmeasured confounders. As a consequence, methods that rely on the assumption of no unmeasured confounding will likely be biased. Figure 1 shows how the variables in our model are related; the arrows indicate the direction of causation.

[Figure 1 about here.]

Using the structural equation framework of (Pearl 2000), we assume that each variable is a function of other variables that affect it and a random term (also called error term). Let $U$ denote the error terms. Thus, we have

$W = f_W(U_W), Z = f_Z(W, U_Z), A = f_A(W, Z, U_A), Y = f_Y(W, Z, A, U_Y)$

where $U = (U_W, U_Z, U_A, U_Y) \sim P_{U,0}$ is an exogenous random variable, and $f_W$, $f_Z$, $f_A$, $f_Y$ may be unspecified or partially specified (for instance, we might know that the instrument is randomized). Further, three assumptions need to be made to guarantee that $Z$ is a valid instrument for estimating the effect of $A$ on $Y$:

**Assumptions ensuring that $Z$ is a valid instrument:**

(1) $Z$ only affects outcome $Y$ through its effect on treatment $A$. Thus, $f_Y(W, Z, A, U_Y) = f_Y(W, A, U_Y)$.

(2) Given baseline covariates $W$, the random terms $U_Z$ and $U_Y$ are conditionally independent. Equivalently, $U_Z \perp\!\!\!\perp U_Y | W$.

(3) $\text{Var}_0[E_0(A|Z, W)|W] > 0$ for all $W$.

In other words, although we don't assume that $A$ is randomized with respect to $Y$, we do assume that $Z$ is randomized with respect to $Y$, conditional on $W$ in both cases. The last assumption guarantees that for every value of covariates $W$, there is variation in the instrument, and that the instrument induces variation in the treatment. Further, we assume the following form for the marginal structural equation for outcome $Y$, where $m_0$ and $\theta_0$ are unspecified:

**Structural equation for outcome $Y$:**

$$Y = f_Y(W, A, U_Y) = Am_0(W) + \theta_0(W) + U_Y$$

Assumption 2 guarantees that $E(U_Y|Z, W) = 0$.

The linearity in $A$ of the structural equation for $Y$ is necessary for identifying the treatment effect using an instrument unless further assumptions are made. In the common case where the treatment $A$ is binary, this assumption always holds, and we have a fully general semi-parametric model that only assumes $Z$ is a valid instrument. It should also be noted that unlike many instrument-based estimators, we don't require the instrument to be randomized with respect to treatment ($U_Z \perp\!\!\!\perp U_A|\ W$ is not necessary).

We use the counterfactual framework of (Pearl 2000) to define the causal parameter of interest. Let counterfactual outcome $Y(a)$ denote the outcome given by the structural equations if the treatment variable were set to $A = a$, and all other variables, including the exogenous terms, were unchanged. We have that $Y(a) = a \cdot m_0(W) + \theta_0(W) + U_Y$ for all possible values $a \in \mathcal{A}$, where $\mathcal{A}$ denotes a support of $A$. We can now define the marginal causal effect we're interested in as $E_0(Y(a) - Y(0))$ and observe that it equals $a \cdot Em_0(W)$. Similarly, define adjusted causal effects $E_0(Y(a) - Y(0) \mid V)$ conditional on a user supplied covariate $V \subset W$. These causal effects are functions of $m_0(W)$ and the distribution of $W$.

**Causal effect of interest:**

The marginal causal effect is $E_0(Y(a) - Y(0)) = a \cdot Em_0(W)$.

The adjusted causal effect is $E_0(Y(a) - Y(0) \mid V) = a \cdot E(m_0(W) \mid V)$, given a user supplied covariate $V \subset W$.

Note that $m_0(W)$ represents the causal effect of one unit of treatment given $W$.

**Notation.** Let $\Pi_0(Z, W) \equiv E_0(A \mid Z, W)$ be the conditional mean of $A$ given $Z, W$.

Let $\mu_0(Z,W) \equiv E_0(Y \mid \Pi_0(Z,W),W)$ be the expected value of $Y$, given $W$ and $\Pi_0(Z,W)$.

The instrumental variable assumption that $E(U_Y|Z,W)=0$ implies

$$E_0(Y \mid \Pi_0(Z,W),W) = \Pi_0(Z,W)m_0(W) + \theta_0(W)$$

Thus, our structural equation model implies a semiparametric regression model for $E_0(Y \mid \Pi_0(Z,W),W)$. Note that for a pair of values $z$ and $z_1$, we have

$$E_0(Y \mid Z=z,W) - E_0(Y \mid Z=z_1,W) = \{\Pi_0(z,W) - \Pi_0(z_1,W)\}m_0(W)$$

From this equation, we get an identifiability result for $m_0$, stated below as a formal lemma.

LEMMA 1: *Let $\Pi_0(Z,W) \equiv E_0(A \mid Z,W)$. Let $d_{Z,0}$ be the conditional probability distribution of $Z$, given $W$. Let $\mathcal{W}$ be a support of the distribution $P_{W,0}$ of $W$. Let $w \in \mathcal{W}$. By assumption 3 above, $Var(\Pi_0(z,w)|W=w) > 0$, so there exists two values $(z,z_1)$ in a support of $d_{Z,0}(\cdot \mid W=w)$ for which $\Pi_0(z,w) - \Pi_0(z_1,w) \neq 0$. Thus*

$$m_0(w) = \frac{E_0(Y \mid Z=z,W=w) - E_0(Y \mid Z=z_1,W=w)}{\Pi_0(z,w) - \Pi_0(z_1,w)},$$

*which demonstrates that $m_0(w)$ is identified as a function of $P_0$.*

**Statistical model:** The above stated causal model implies the statistical model $\mathcal{M}$ consisting of all probability distributions $P$ of $O=(W,Z,A,Y)$ satisfying $E_P(Y \mid Z,W) = \Pi(P)(Z,W)m(P)(W)+\theta(P)(W)$ for some unspecified functions $m(P),\theta(P)$, and $\Pi(P)(Z,W) = E_P(A \mid Z,W)$. $\Pi(P)(Z,W)$ must satisfy $\mathrm{Var}_P[\Pi(P)(Z,W)|W] > 0$ for all $W$.



**Causal parameter:** We define our causal parameter of interest to be the projection of the dose-response curve $E_0(Y(a)-Y(0) \mid V) = aE_0(m_0(W) \mid V)$ on a working model. Let $\{am_\beta(v):\beta\}$ be a working model for $E_0(Y(a)-Y(0) \mid V)$. Specifically, given some weight

function $h(A,V)$, let

$$\begin{aligned}
\beta_0 &= \arg\min_\beta E_0 \sum_a h(a,V)\{aE(m_0(W) \mid V) - am_\beta(V)\}^2 && (1)\\
&= \arg\min_\beta E_0 \sum_a h(a,V)a^2\{E(m_0(W) \mid V) - m_\beta(V)\}^2 && (2)\\
&= \arg\min_\beta E_0 \sum_a h(a,V)a^2\{m_0(W) - m_\beta(V)\}^2 && (3)\\
&\equiv \arg\min_\beta E_0\ j(V)\{m_0(W) - m_\beta(V)\}^2, && (4)
\end{aligned}$$

where we defined $j(V) \equiv \sum_a h(a,V)a^2$.

For example, if $V$ is empty, and $m_\beta(v) = \beta$, then $E_0(Y(a) - Y(0)) = \beta_0 a$. We can also select $V = W$ and $m_\beta(w) = \beta^T w$, in which case $\beta_0^T w$ is the projection of $m_0(w)$ on this linear working model $\{\beta^T W : \beta\}$.

**Statistical Target parameter:** Our target parameter is $\psi_0 = \beta_0$.

Let $\Psi : \mathcal{M} \to \mathbb{R}^d$ be the target parameter mapping so that $\Psi(P_0) = \psi_0 = \beta_0$, which exists under the identifiability assumptions stated in Lemma 1. We note that $\psi_0 = \Psi(P_0) = \Psi(m_0, P_{W,0})$ only depends on $P_0$ through $m_0$ and $P_{W,0}$, while $m_0$, as statistical parameter of $P_0$, is identified as a function of $\mu_0 = E_0(Y \mid Z, W)$ under the semiparametric regression model $\mu_0 = E_0(Y \mid Z, W) = \pi_0(Z,W)m_0(W) + \theta_0(W)$.

The statistical estimation problem is now defined. We observe $n$ i.i.d. copies of $O = (W, Z, A, Y) \sim P_0 \in \mathcal{M}$, and we want to estimate $\psi_0 = \Psi(P_0)$ defined in terms of the mapping $\Psi : \mathcal{M} \to \mathbb{R}^d$.

**Weakening the structural assumption** We briefly note that the structural assumption $Y = f_Y(W, A, U_Y) = Am(W) + \theta(W) + U_Y$ can be weakened in many cases when $Z$ is a continuous variable. For a general equation $Y = f_Y(W, A, U_Y) = q(W, A) + U_Y$, where $q(W, A)$ is any function, we can write a Taylor approximation for a k-degree polynomial in $A$ as

$$f_Y(W, A, U_Y) = A^k m_k(W) + A^{k-1} m_{k-1}(W) + ... + Am_1(W) + m_0(W) + U_Y$$

Now suppose we have $(k+1)$ values of $Z$: $(Z_k, Z_{k-1}, ...., Z_0)$. We have that $E(Y|Z_i, W) = E(A^k|Z_i, W)m_k(W) + E(A^{k-1}|Z_i, W)m_{k-1}(W) + ... + m_0(W)$. This means if the equation below is solvable (the matrix shown is not singular), then we can identify $(m_k(W), m_{k-1}(W), ..., m_0(W))$.

$$\begin{bmatrix} E(Y|Z_k, W) \\ \vdots \\ E(Y|Z_0, W) \end{bmatrix} = \begin{pmatrix} E(A^k|Z_k, W) & E(A^{k-1}|Z_k, W) & \cdots \\ \vdots & \ddots & \vdots \\ E(A^k|Z_0, W) & E(A^{k-1}|Z_0, W) & \cdots \end{pmatrix} \begin{bmatrix} m_k(W) \\ \vdots \\ m_0(W) \end{bmatrix}$$

## 4. Targeted minimum loss based estimation

### 4.1 *The efficient influence curve of* $\Psi$

The efficient influence curve for $\Psi$ is derived in Appendix 2. Recall our semiparametric model, and notation $P_{W,0}, \pi_0, (Z, W), m_0(W), \theta_0(W)$, from section 3. Let $d_0(Z, W) = Pr(Z|W)$. Also, define $h_1(V) \equiv \sum_a h(a, V)a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)$, which has the same dimension as $\beta_0$, where $h(a, V)$ is defined in section 3.

LEMMA 2: *The efficient influence curve of* $\Psi : \mathcal{M} \to \mathbb{R}^d$ *is given by*

$$\begin{aligned} &D^*(P_0) = D^*_W(P_0) \\ &+c_0^{-1} \tfrac{h_1(V)}{\sigma^2(W)} (\pi_0(Z, W) - E_0(\pi_0(Z, W) \mid W))(Y - \pi_0(Z, W)m_0(W) - \theta_0(W)) \\ &-c_0^{-1} \tfrac{h_1(V)}{\sigma^2(W)} \{(\pi_0(Z, W) - E_0(\pi_0(Z, W) \mid W))m_0(W)\} (A - \pi_0(Z, W)) \\ &\equiv D^*_W(P_0) + C_Y(Z, W)(Y - \pi_0(Z, W)m_0(W) - \theta_0(W)) \\ &\qquad\qquad -C_A(Z, W)(A - \pi_0(Z, W)) \\ &\equiv D^*_W(P_0) + D^*_Y(P_0) - D^*_A(P_0), \end{aligned} \tag{5}$$

*where*

$$c_0 \equiv E_0 \sum_a h(a, V)a^2 \left\{ \frac{d}{d\beta_0} m_{\beta_0}(V) \right\}^2,$$

*which is a* $d \times d$ *matrix, and*

$$D^*_W(P_0) \equiv c_0^{-1} \sum_a h(a, V)a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)(m_0(W) - m_{\beta_0}(V))$$

$$\begin{aligned}
\sigma^2(W) &= Var_{d_0}(\Pi_0(Z,W) \mid W)) \\
h(W) &= c_0^{-1}\frac{h_1(V)}{\sigma^2(W)} \\
C_Y(Z,W) &= h(W)(\pi_0(Z,W) - E_{d_0}(\pi_0(Z,W) \mid W)) \\
C_A(Z,W) &= C_Y(Z,W)m_0(W).
\end{aligned}$$

Note that $D^*(P_0)$ will be a vector-valued function in general.

Appendix 3 gives the derivation of the efficient influence curve in the special case of assuming a parametric form for the effect of treatment as a function of covariates, meaning that $m_0 = m_{\alpha_0}$ for some model $\{m_\alpha : \alpha\}$.

### 4.2 *The targeted minimum loss-based framework*

Targeted minimum loss-based estimation (TMLE) is a method to construct a semi-parametric substitution estimator of a target parameter $\Psi(P_0)$ of a true distribution $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a semiparametric statistical model (van der Laan and Rubin 2006, van der Laan and Rose 2011). The estimate is based on sampling $n$ i.i.d. data points $(O_1, ..., O_n)$ from $P_0$. It is consistent and asymptotically efficient under certain conditions.

(1) One first notes that the parameter of interest $\Psi(P_0)$ depends on $P_0$ only through relevant components $Q_0$ of the full distribution $P_0$, in other words, $\Psi(P_0) = \Psi(Q_0)$ [2]. TMLE *targets* these relevant components by only estimating these $Q_0$ and certain nuisance parameters $g_0$ [3] that are needed for updating the relevant components. An initial estimate $(Q_n^0, g_n)$ is formed of the relevant components and nuisance parameters. This is typically done using the Super Learner (see below) approach described in (van der Laan et al 2007), in which the best combination of learning algorithms is chosen from a library using cross-validation.

[2] We are abusing notation here for the sake of convenience by using $\Psi(\cdot)$ to denote both the mapping from the full distribution to $\mathbb{R}^d$, and from the relevant components to $\mathbb{R}^d$.

[3] The nuisance parameters are those components $g_0$ of the efficient influence curve $D^*(Q_0, g_0)$ that $\Psi(Q_0)$ does not depend on.

(2) Then the relevant components $Q_n^0$ are fluctuated, possibly in an iterative process, in an optimal direction for removing bias efficiently. (3) Finally, one evaluates the statistical target parameter on the updated relevant components $Q_n^*$, and arrives at estimate $\psi_n^* = \Psi(Q_n^*)$.

Note that the final estimate of $\psi_n^*$ is formed by evaluating the target parameter on estimates of relevant components that are consistent with a single data-generating distribution, and with the observed bounds of the data. This property of being a *substitution estimator* has been shown to be conducive to good performance in practice (Gruber and van der Laan 2010).

We use notation such as $Q_n^0$, where the subscript clarifies that an empirical estimate is being made from the sample of size $n$, while the superscript refers to the estimate being an initial one ("zeroeth" iteration). To fluctuate the initial components $Q_n^0$ to updated components $Q_n^1$, one defines a fluctuation function $\epsilon \to Q(\epsilon|g_n)$. $g_n$ is an estimate of the nuisance parameters, and the fluctuation of $Q_n^0$ can depend on $g_n$, although we sometimes drop the explicit dependency in the notation, and use $Q(\epsilon)$ to denote $Q(\epsilon|g_n)$. One also defines a loss function $L()$, where we set $Q_n^1 = Q_n^0(\epsilon_n^0|g_n)$ by solving for fluctuation $\epsilon_n^0 = \text{argmin}_\epsilon \; L(Q_n^0(\epsilon|g_n), g_n, (O_1, ..., O_n))$. We use the convention that when the fluctuation parameter $\epsilon$ is zero, $Q_n^0(\epsilon|g_n) = Q_n^0$. This procedure of updating $Q_n^{k+1} = Q_n^k(\epsilon_n^k|g_n)$ might need to be iterated to convergence. In some versions of TMLE, the nuisance parameters $g_n$ are also updated, using a fluctuation function and loss function similarly. The requirement is to choose the fluctuation and loss functions so that, upon convergence of the components to their final estimate $Q_n^*$ and $g_n^*$, the efficient influence curve equation is solved:

$$P_n \; D^*(Q_n^*, g_n^*) = 0$$



$P_n$ denotes the empirical distribution $(O_1, ..., O_n)$, and we use the shorthand notation $P_n f = \frac{1}{n} \sum_{i=1}^n f(O_i)$. The equation above is the basis for the guarantees of consistency

(under partial misspecification) and asymptotic efficiency (under correct specification of relevant components and nuisance parameters).

To give a few examples, the loss function might be the mean squared error, or the negative log likelihood function. For instance, for the estimator using iterative updating presented in section 4.5, we use fluctuation $\mu_n^1 = \mu_n^0 + \epsilon \cdot C_{Y,n}^0$, with $\mu = E(Y|W,Z)$ and $C_Y$ as defined in section 4.1. The loss function is $L(Q_n^0(\epsilon|g_n), g_n, (O_1, ..., O_n)) = \sum_{i=1}^n \left(Y[i] - \mu_n^0[i] - \epsilon \cdot C_{Y,n}^0[i]\right)^2$.

Here is the TMLE estimation procedure for our marginal structural model:

**Step 1: Forming initial estimates.**

**Components of $P_0$ that need to be estimated:** Initial estimates must be formed of *relevant components* $Q_n^0 = (m_n^0(W), P_{W,n})$, and
*nuisance parameters* $g_n^0 = (\Pi_n^0(Z,W), E_n^0(\Pi_n^0|W), \mathrm{Var}_n^0(\Pi_n^0|W), \theta_n^0(W))$.

**Super Learner.** We use the **Super Learner** approach to form initial estimates (van der Laan et al 2007), and software implementation in R (`http://cran.r-project.org/web/packages/SuperLearner/index.html`). Super Learner is a data-adaptive technique to choose the best linear combination of learning algorithms from a library. The objective that is minimized is the cross-validated empirical mean squared error. Each candidate learning algorithm is trained on all the data except for a hold-out test set, and this process is repeated over different hold-out sets so all data points are included in a test set. The linear combination of candidate learners that minimizes MSE over all test sets in chosen. This method has the very desirable guarantees that: 1) if none of the candidate learners converge at a parametric rate, Super Learner asymptotically attains the same risk as the oracle learner, which selects the true optimal combination of learners and 2) if one of the candidate learners uses a parametric model and contains the true data-generating distribution, Super Learner

converges at an almost-parametric rate.

See section 5.2 for a list of candidate learning algorithms we use for forming the initial estimates.

**Step 2: Fluctuating the relevant components $Q_n^0$.**

We present three versions of TMLE in this paper: one where the relevant components and nuisance parameters are fluctuated iteratively, and two versions of the non-iterative TMLE described below.

**Non-iterative TMLE.** Suppose we have a fluctuation function $\epsilon \to Q(\epsilon|g_n)$ so that we can solve for $\epsilon$ the equation:

$$P_n D^*(Q_n^0(\epsilon|g_n), g_n) = 0 \tag{6}$$

Then the efficient influence curve is satisfied in a single update and there is no need for iteration. This case corresponds to using the loss function $L(Q, g, (O_1, ..., O_n)) = \mid \frac{1}{n} \sum_{i=1}^n D^*(Q, g)(O_i) \mid^2$. In a single step, a solution can be found so the loss function takes its lower bound of 0.

It turns out that we can solve 6 without updating $P_W$ by setting it to its empirical distribution $P_W = P_{W,n}$ of the baseline covariates. Thus, we need to solve

$$P_n D^*(Q_n^* = \{m_n^0(\epsilon), P_{W,n}\}, g_n) = 0 \tag{7}$$

where we drop the explicit dependency of $m_n^0(\epsilon)$ on $g_n$ in the notation. Sections 4.3 and 4.4 describe versions of this non-iterative estimator that use logistic and linear fluctuations, respectively, for $m_n^0(\epsilon)$.

**Step 3: Obtain final estimate $\beta_n^* = \Psi(m_n^*, P_{W,n})$.**

**Properties of TMLE.** See Appendix 1 for sketches of proofs.

*Efficiency*

(See van der Laan and Robins 2003, and van der Laan and Rubin 2006.)

Recall that an *efficient estimator* is one that achieves the optimal asymptotic variance among semiparametric estimators. We briefly give a few relevant definitions.

An estimator is *asymptotically linear* if, informally, it is asymptotically equivalent to a sample average. Formally, we have that an estimator $\Psi_n^*$ for estimating true parameter $\Psi(P_0)$ from an iid sample $(O_1, .., O_n)$ is asymptotically linear if
$\sqrt{n}(\Psi_n^* - \Psi(P_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\Psi}_P(O_i) + o_P(1)$, where $\dot{\Psi}_P$ is a zero mean, finite variance function. $\dot{\Psi}_P$ is called the influence function.

Recall that a parameter $\Psi$ is pathwise differentiable at $P_0$ relative to a tangent space of a model $\mathcal{P}$ at $P_0$ if there exists a continuous linear map $\dot{\Psi}_{P_0}$ such that for every score function $g$ in the tangent space and submodel $t \to P_t$ with score function $g$, we have
$\frac{\Psi(P_t) - \Psi(P_0)}{t} \to \dot{\Psi}_{P_0}\ g$. By the Riesz representation theorem, we have $\dot{\Psi}_{P_0}\ g = \int \tilde{\Psi}_{P_0}\ g dP_0$ where $\tilde{\Psi}_{P_0}$ is an "influence function". The *efficient influence curve* is the unique influence function whose coordinate functions are contained in the closure of the linear span of the tangent space.

An estimator is *efficient* if it is asymptotically linear with the efficient influence curve as its influence function. Thus, we have that for an efficient estimator $\Psi_n^*$ estimating true parameter $\Psi(P_0)$ from an iid sample $(O_1, ..., O_n)$:
$\Psi_n^* - \Psi(P_0) = \frac{1}{n} \sum_{i=1}^n D^*(P_0)(O_i) + o_P(\frac{1}{\sqrt{n}})$, where $D^*$ is the efficient influence curve.

Suppose all initial estimates $(Q_n^0, g_n^0)$ are consistent, and that
$P_0\ \left(D^*(Q_n^*, g_n^*) - D^*(Q_0, g_0)\right)^2 \in o_P(1)$. Then the final estimate $\Psi(Q_n^*)$ is asymptotically efficient, with

$$\Psi(Q_n^*) - \Psi(Q_0) = \left[P_n - P_0\right]\ D^*(Q_0, g_0) + o_P(1/\sqrt{n}) \tag{8}$$

*Consistency under misspecification*

TMLE yields a consistent estimate for $\Psi^* = \beta_n^*$ under 3 scenarios of partial misspecification of components:

(1) Initial estimates $\Pi^0$ and $\Pr^0(Z|W)$ are consistent.

(2) Initial estimates $m^0$ and $\Pr^0(Z|W)$ are consistent.

(3) Initial estimates $m^0$ and $\theta^0$ are consistent.

### 4.3 *Estimator using a logistic fluctuation for scalar* $\psi$

This estimator has the advantage that it can match the bounds of the observed data in estimating $E(Y|W, \Pi(Z, W))$.

In accordance with the non-iterative TMLE procedure, we want to find $\epsilon$ such that $P_n D^*(Q_n^* = \{m_n^0(\epsilon), P_{W,n}\}, g_n) = 0$ according to 7.

A pre-processing step is done of converting $Y$-values to the range [0,1] using a linear mapping $Y \to \tilde{Y}$, where $\tilde{Y} = 0$ corresponds to $\min(Y)$ in the dataset and $\tilde{Y} = 1$ to $\max(Y)$. Thus, we can use the mapping $\tilde{Y} = (Y - \min(Y))/(\max(Y) - \min(Y))$. The equation $E(Y \mid \Pi(Z, W), W) = \Pi(Z, W)m(W) + \theta(W)$ can be written as $E(\tilde{Y} \mid \Pi(Z, W), W) = \Pi(Z, W)\tilde{m}(W) + \tilde{\theta}(W)$, where $\tilde{m}(W) = m(W)/(\max(Y) - \min(Y)) \in [-1, 1]$ and $\tilde{\theta}(W) = (\theta(W) - \min(Y))/(\max(Y) - \min(Y)) \in [0, 1]$. Now initial estimates can be formed of all relevant components and nuisance parameters using the modified data set $(W, Z, A, \tilde{Y})$.

Replacing $m_n^0(\epsilon)$ with $\tilde{m}_n^0(\epsilon)$, we use this fluctuation function in equation 7:

$$\tilde{m}_n^0(\epsilon)(W) = 2 \times \text{logistic}(\text{logit}(\frac{\tilde{m}_n^0(W) + 1}{2}) + \epsilon^T \cdot h(W)) - 1 \tag{9}$$

where logistic() denotes the function $\text{logistic}(x) = \frac{1}{1+e^{-x}}$ and logit() its inverse $\text{logit}(y) = \log \frac{y}{1-y}$. This corresponds to the mapping $f(\epsilon) = \text{logistic}(\text{logit}(f) + \epsilon \cdot h)$ where $f$ is $\tilde{m}_n^0$ scaled to be in $[0, 1]$.

Inspecting the efficient influence curve, we have that the first term $P_n\ D_W^*(Q_n^*, g_n) = 0$, because this expression is equivalent to $\beta_n^* = \arg\min_\beta P_{W,n} j(V)\{m_n^*(W) - m_\beta(V)\}^2$, which holds by definition of $\beta_n^*$. Also, we have that the
$+/-\ h(W)(\pi(Z, W) - E(\pi(Z, W) \mid W))(\pi(Z, W)m(W))$ terms cancel. Thus
$D^*(Q, g)$ reduces to $h(W)(\pi(Z, W) - E(\pi(Z, W) \mid W))(Y - A \cdot m(W) - \theta(W))$

so we need to find $\epsilon$ such that

$$P_n\ D^*(\tilde{m}_n^0(\epsilon), P_{W,n}, g_n^0) =$$

$$\frac{1}{n}\sum_{i=1}^{n}\ h_n^0(W)(\pi_n^0(Z,W) - E_n^0(\pi_n^0 \mid W))(\tilde{Y} - A \cdot \tilde{m}_n^0(\epsilon)(W) - \tilde{\theta}_n^0(W)) = 0$$

for $\tilde{m}_n^0(\epsilon)(W)$ defined in 9.

Since $E_0(\tilde{Y} - A \cdot \tilde{m}_0(W) - \tilde{\theta}_0(W)|Z,W) = 0$, the equation above has a solution $\epsilon$ for any reasonable initial estimates $(Q_n^0 = \{\tilde{m}_n^0, P_{W,n}\}, g_n^0)$. For $k = \dim(\beta)$, we have a $k$-dimensional equation in $k$-dimensional $\epsilon$. When $k = 1$ and we need a scalar $\epsilon$, we can use a bisection method as a computationally simple way to compute $\epsilon$. One first finds left and right boundaries $\epsilon_1$, $\epsilon_2$ such that

$$E_n\ h_n^0(W)(\pi_n^0(Z,W) - E_n^0(\pi_n^0 \mid W))(A \cdot \tilde{m}_n^0(\epsilon_1)(W)) \leqslant$$

$$E_n\ h_n^0(W)(\pi_n^0(Z,W) - E_n^0(\pi_n^0 \mid W))(\tilde{Y} - \tilde{\theta}_n^0(W)) \leqslant$$

$$E_n\ h_n^0(W)(\pi_n^0(Z,W) - E_n^0(\pi_n^0 \mid W))(A \cdot \tilde{m}_n^0(\epsilon_2)(W))$$

where $E_n$ denotes the empirical mean. Then one iteratively shrinks the distance between the left and right boundaries $\epsilon_1$ and $\epsilon_2$ until a suitably close approximation to the solution is found.

Once one solves for $\epsilon$ and finds $\tilde{m}_n^* = \tilde{m}_n^0(\epsilon)$, one converts back to the original scale for outcome $Y$, by setting $m_n^* = \tilde{m}_n^* \cdot (\max(Y) - \min(Y))$. Then the parameter of interest is evaluated by finding $\Psi(m_n^*, P_{W,n}) = \beta_n^*$.

When the parameter of interest $\psi$ is vector-valued, solving the efficient influence curve equation using a logistic fluctuation translates to a non-convex multi-dimensional optimization problem with no known analytical solution. Various numerical techniques and software packages are available.

One application of this estimator is to use a tighter bound for $E(Y|\Pi(Z,W),W)$ than the

bounds of the data. For instance, when $Y$ is a rare binary outcome, its conditional mean for any value of $W$ might lie in a far smaller interval than $[0, 1]$.

### 4.4 *Estimator using a linear fluctuation*

Once again, we want to find $\epsilon$ such that $P_n D^*(Q_n^* = \{m_n^0(\epsilon), P_{W,n}\}, g_n) = 0$ according to 7. A TMLE-based estimator that is especially simple to understand and implement involves using a simple linear fluctuation

$$m_n^0(\epsilon)(W) = m_n^0(W) + h(W)^T \cdot \epsilon$$

and solving for $\epsilon$ in a single non-iterative step. $h(W) = c_0^{-1} \frac{h_1(V)}{\sigma^2(W)}$ as defined in section 4.1.

As usual, we form initial estimates of all relevant components and nuisance parameters. In solving the efficient influence curve equation 7, once again we have that $P_n\ D_W^*(Q_n^*, g_n) = 0$, and we can simplify to get

$$E_n\ h_n^0(W)(\pi_n^0) - E_n^0(\pi_n^0(Z, W) \mid W))(Y - A \cdot m_n^0(W) - \theta_n^0(W))$$

$$= E_n\ h_n^0(W)(\pi_n^0 - E_n^0(\pi_n^0(Z, W) \mid W))(A \cdot h_n^0(W)^T \epsilon)$$

We can solve for (generally vector-valued) $\epsilon$ by finding the solution to a simple system of linear equations. As usual, we then set $m_n^* = m_n^0(\epsilon) = m_n^0(W) + h(W)^T \cdot \epsilon$, and evaluate the parameter of interest $\psi_n^* = \Psi(m_n^*, P_{W,n})$ by finding the projection $\beta_n^*$ of $m_n^*$ unto the working model $\{m_\beta(v) : \beta\}$.

This approach is simple and achieves the same asymptotic guarantees as any of the other formulations of TMLE. However, it has the drawback compared to the version described above using logistic fluctuation that the final estimate $\mu_n^* = \Pi_n^0 \cdot m_n^* + \theta_n^0$ is not constrained to observe the bounds of $Y$ in the data.

### 4.5 *Estimator using iterative updating*

One estimation method in the TMLE framework we developed involves iteratively updating relevant components and nuisance parameters until convergence to components $(Q_n^*, g_n^*)$ such that the efficient influence curve equation is satisfied: $P_n\ D^*(Q_n^*, g_n^*) = 0$.

As usual, initial estimates are formed of all relevant components $Q_n^0$ and the nuisance parameters $g_n^0$. We set $P_W$ to its empirical distribution $P_W = P_{W,n}$ and never update that component. Next, at each iteration until convergence, we fluctuate components as follows:

**(i)** Let $k$ denote the iteration number. For $\mu = E(Y|W,Z)$, and $C_Y(Z,W) = h(W)(\pi(Z,W) - E(\pi(Z,W) \mid W))$ as defined in section 4.1, we have:

$$\mu_n^{k+1} = \mu_n^k + \epsilon \cdot C_{Y,n}^k$$

$$\epsilon = \ \arg\min \sum_{i=1}^{n} (Y[i] - \mu_n^k[i] - \epsilon \cdot C_{Y,n}^k[i])^2$$

Note that by setting $m_n^{k+1} = m_n^k + \epsilon \cdot h_n^k$, and $\theta_n^{k+1} = \theta_n^k + \epsilon \cdot [-h_n^k E(\Pi_n^k|W)]$, where $h_n^k$ refers to $h(W) = c_0^{-1} \frac{h_1(V)}{\sigma^2(W)}$, we have that $\mu_n^{k+1} = m_n^{k+1} \cdot \Pi_n^k + \theta_n^{k+1}$ and thus remains in our marginal structural model.

**(ii)** Given $\mu_n^{k+1}$, we update $C_{A,n}^{k+1} = C_{Y,n}^k m_n^{k+1}$ and then fluctuate $\Pi_n^k(Z,W) = E_n^k(A|Z,W)$ as follows. If $A$ is continuous, we first replace $A$ with linear transformation $A' \in [0,1]$, where $A' = (A - \min(A))/(\text{range}(A))$, and apply the inverse transformation to get the final $\Pi(Z,W)$.

$$\Pi_n^{k+1} = \Pi_n^k(\epsilon) = \text{logistic}(\text{logit}(\Pi_n^k) + \epsilon \cdot C_{A,n}^{k+1})$$

$$\epsilon = \arg\min \sum_{i=1}^{n} \Big[ - A[i] \cdot \log(\Pi_n^k(\epsilon)[i]) - (1 - A[i]) \cdot \log(1 - \Pi_n^k(\epsilon)[i]) \Big]$$

where the logistic function is $\frac{1}{1+e^{-x}}$ and the logit its inverse. The optimization above is

solved using standard logistic regression software, even though the independent variable can be continuous in $[0, 1]$ here. We then update

$C_{Y,n}^{k+1} = h_n^k \cdot ( \Pi_n^{k+1} - E\big[\Pi_n^{k+1}(Z, W|W)\big] )$.

**(iii)** Finally, we update the $E(\Pi(Z, W)|W)$-component to be $E(\Pi_n^{k+1}(Z, W)|W)$, and $\sigma^2(W) = \text{Var}(\Pi(Z, W)|W)$ as $\text{Var}(\Pi_n^{k+1}(Z, W)|W)$, using the initial estimates for the relevant parts of $\Pr(Z|W)$.

This algorithm converges to components $(Q_n^*, g_n^*)$. In each step of updating $\mu_n^k$, $\Pi_n^k$, we are solving for $\epsilon^k$ to minimize $\sum_{i=1}^n L(P_n^k(\epsilon))(O_i)$, for some loss function $L$ and parametric submodel $P(\epsilon)$. Thus we have $\frac{d}{d\epsilon}\sum_{i=1}^n L(P_n^k(\epsilon))(O_i)|_{\epsilon=\epsilon^k} = 0$. As the algorithm converges, we have that the objective $\sum_{i=1}^n L(P_n^*(\epsilon))(O_i)$ is minimized with $\epsilon = 0$; in other words, the components $(Q_n^*, g_n^*)$ are already optimal for the loss function and do not get fluctuated. Thus, we have $\frac{d}{d\epsilon}\sum_{i=1}^n L(P_n^*(\epsilon))(O_i)|_{\epsilon=0} = 0$.

It is easy to check that for the loss function used to update $\mu_n^k$, we have $\frac{d}{d\epsilon}L(P(\epsilon))|_{\epsilon=0} = C_Y \cdot (Y - \mu) = D_Y^*$, so we have $P_n\ D_Y^* = 0$ upon convergence. Similarly, for the loss function used to update $\Pi_n^k$, we have $\frac{d}{d\epsilon}L(P(\epsilon))|_{\epsilon=0} = C_A \cdot (A - \Pi) = D_A^*$, so we have $P_n\ D_A^* = 0$ upon convergence. We have that the first term $P_n\ D_W^* = 0$, because this expression is equivalent to $\beta_n^* = \arg\min_\beta P_{W,n} j(V)\{m_n^*(W) - m_\beta(V)\}^2$, which holds by definition of $\beta_n^*$. Thus, $P_n D^*(Q_n^*, g_n^*) = 0$ and we have a valid TMLE procedure.

## 5. Simulation results

We show results from a number of simulations. We compare all three versions of a TMLE-based estimator proposed above to several standard methods: 1) a likewise semiparametric, locally efficient estimator based on the method of estimating equations; 2) two-stage least squares, which is a standard parametric approach; 3) a biased estimate of the causal effect of $A$ on $Y$ ignoring the confounding.

There are two cases we use for the parameter of interest: *Scalar.* We estimate a constant mean causal effect $E(Y(1) - Y(0)) = E(m(W)) = \beta = m_\beta(v)$. *Vector-valued linear.* We use a linear working model $m_\beta(w) = \beta^T \binom{1}{w}$ for $E(Y(1) - Y(0)|W) = m(W)$.

## 5.1 ***Standard approaches for comparison.***

5.1.1 *The method of estimating equations.* (van der Laan and Robins 2003) presents background on the method of estimating equations. When the efficient influence curve is an explicit function of the parameter of interest $\Psi_0$, under regularity conditions, one can solve for $\Psi$ using the equation

$$P_n D^*(P) = P_n \; D^*(P_W, \Pi, \mathrm{E}(\Pi|W), \mathrm{Var}(\Pi|W)m, \theta, \Psi) = 0$$

The components of $D^*$ are estimated using Super Learner, just as with TMLE. Estimating equations has the same properties of local efficiency and robustness to misspecification as the TMLE-based estimators: when all relevant components and nuisance parameters are estimated consistently, the estimate is asymptotically efficient, and as long as $(P_W, \Pi_n^0, \mathrm{E}_n^0(\Pi|W), \mathrm{Var}_n^0(\Pi|W))$ are estimated consistently, the estimate for the parameter of interest $\Psi = \beta$ is consistent.

In the scalar case, our estimating equation is

$$E_n \left[ c_0^{-1} j(V)(m(W) - \beta) + D_Y^*(P)(Y, Z, W) - D_A^*(P)(A, Z, W) \right] = 0$$

where the $D_Y^*$, $D_A^*$ terms to do not depend on $\beta$.

For the case of a linear working model, the estimating equation is

$$E_n \left[ c_0^{-1} j(V) \begin{pmatrix} 1 \\ W \end{pmatrix} (m(W) - \beta' \begin{pmatrix} 1 \\ W \end{pmatrix}) + D_Y^*(P)(Y, Z, W) - D_A^*(P)(A, Z, W) \right] = 0$$

which can also be solved as a linear equation of $\beta$. The terms $D_Y^*$, $D_A^*$ do not depend on $\beta$ and are vector-valued here.



5.1.2 *Two-stage least squares.* The most widely used solution to estimating the effect of a treatment on an outcome in the presence of a confounder and valid instrument is to

use a linear model for both the "first-stage" equation $A = \alpha_Z Z + \alpha_W W + \alpha_1 1 + \epsilon_A$ and the "second stage": $Y = \beta_A A + \beta_W W + \beta_1 1 + \epsilon_Y$. When there is a single instrumental variable and treatment, which is the case we study, a solution for scalar $\hat{\beta}$ that is consistent and asymptotically optimal among linear models is $\hat{\beta} = ((Z, W, 1)'(A, W, 1))^{-1}((Z, W, 1)'Y)$. This estimate corresponds to the two-stage least squares solution where one estimates $A^* = E(A|Z, W)$ using a linear model, and then estimates the effect of $(A^*, W)$ on $Y$ using a linear model again (having exogenous variation).

When estimating a vector-valued causal effect, we find $A^* = E(A|Z, W)$ and then do linear regression of $Y$ on cross terms $A^* \times (1, W)$ and covariates $(1, W)$, thus finding a linear treatment effect modifier function $m(W)$ and a linear additive effect function $\theta(W)$. 2SLS is a parametric model and is in general not consistent for estimating our causal parameter of interest.

5.1.3 *Ignoring the confounding.* We include a "confounded" estimator in each table that ignores the unmeasured confounding between the treatment and outcome, and does not use an instrument. We use a correctly specified parametric model for $m(W)$, $\theta(W)$, and estimate their parameters using $E(Y|W, A) = A \cdot m(W) + \theta(W)$, which will give a biased estimate for $m(W)$ by ignoring the confounding between $A$ and the residual term. The correctly specified model for $m(W)$ converges at a parametric rate, and for large $n$, we isolate the effect of the bias arising from not using an instrument.

## 5.2 ***Initial estimates.***

For the semiparametric approaches (our three estimators based on TMLE, and estimating equations), initial estimates are formed as follows. We use the empirical distribution of $W$ for $P_W$ and never update this component. For $\text{Var}_n^0(\Pi_n^0(Z, W)|W)$ and $E_n^0(\Pi_n^0(Z, W)|W)$, noting that our instrument $Z$ is binary in the simulations below, we estimate $P(Z = 1|W) = E(Z|W)$ and find the expectation and variance of $\Pi_n^0(Z, W)$ from $P(Z = 1|W)$,

instead of directly estimating them as a function of $Z$, $W$. Thus we need initial estimates for $E(Z|W), \Pi(Z,W), \theta(W), m(W)$ from the data.

For $\Pi(Z,W)$ in cases where $A$ is binary, and for $E(Z|W)$, we use as candidate learners the following R packages (see the corresponding function specifications in Super Learner): **glm, step, knn, DSA.2, svm, randomForest** (Sinisi and van der Laan 2004). For glm (generalized linear models), step (stepwise model selection using AIC), and svm (support vector machines), we use both linear and second-order terms. In addition, we use cross-validation to find the highest degree of polynomial terms in glm that results in the lowest prediction error, thus using terms of degree higher than two with glm. For $\Pi(Z,W)$ in cases where $A$ is continuous, we use candidate learners **glm, step, svm, randomForest, nnet** and **polymars**.

For $m(W)$ and $\theta(W)$ which involve continuous outcomes, we use candidate learners **glm, step, svm,** and **polymars**. We need to predict $m(W)$ and $\theta(W)$ so that $\mu(Z,W) = \pi(Z,W) \cdot m(W) + \theta(W)$ retains the structural form. We include $\Pi \times m(W)$ cross-terms as well as $\theta(W)$ terms, having various functional forms for parameterizing $m(W)$, $\theta(W)$.



5.3 ***Results.***

In the simulations that follow, we use the following general format for generating data. In accordance with R's notation, the right-hand side of the formulas specify the regressors but leave the link function unspecified. $\epsilon_{AY}$ is a confounding term, while the treatment effect modifier function $m_W$ can be highly non-linear.

$$
\begin{aligned}
W &\sim N(\mu, \Sigma) \\
Z &\sim Binom(p(W)) \\
A &\sim W + Z + \epsilon_{AY} \\
Y &\sim A \cdot m(W) + \theta(W) + \epsilon_{AY}
\end{aligned}
$$

5.3.1 *Nonlinear design 1.* We test our estimators in the case of highly nonlinear treatment effect modification $m(W) \sim e^W$ in tables 1-2. As we show, 2SLS can be extremely biased in recovering the correct projection of $m(W)$ unto a linear working model. We use $W \sim N(3, 1)$, $p = .5$ for $Z$, and a continuous treatment generated as a linear function of its regressor terms.

**Scalar parameter.** (Table 1.) The true effect is 33.23, sample size of $n = 1000$ is relatively small for using an instrumental variable, and 10,000 repetitions are made. The "initial substitution" estimator is formed by substituting the estimates of relevant components into the parameter of interest, which is just $\beta_n^0 = \Psi(Q_n^0) = E_{W,n} m_n^0(W)$ here, or the estimated mean treatment effect. When consistent initial estimates are formed of all components of $D^*$ using Super Learner, we observed a bias of just .0038, and variance of .6990 for the initial substitution estimator. The three new methods all performed very similarly, achieving lower bias than the initial substitution estimator, as well as slightly lower variance. Since all relevant components are consistently specified, the TMLE-based estimators are asymptotically guaranteed to have the lowest possible variance within the class of consistent estimators in our semiparametric model. The same asymptotic guarantees hold for the estimating equations estimator, which achieves similar magnitude bias and slightly higher variance than the TMLE-based estimators. The two-stage least squares (2SLS) estimator, in contrast, achieves not only much higher bias but vastly higher variance than the semiparametric estimators,

even though it is a parametric estimator. The highly misspecified linear model that 2SLS fits for the conditional outcome brings about the bias and large finite-sample variance. Finally, the estimate that ignores confounding has a bias of about 21.

In table 2, we use an inconsistent initial estimate of $Q(W, R)$, namely, we fit an incorrect linear model $m(W) = b'\binom{1}{W}$. Thus, the substitution estimator essentially functions like 2SLS. The confounded and 2SLS estimators are unchanged. The TMLE-based estimators often show bias removal at the expense of some increase in variance as compared to the unfluctuated initial substitution estimator in the case of misspecification. However we don't see that here with the modest sample size (n=1000), for which the initial substitution estimator has fairly large variance in this simulation. Also, in this case of a scalar parameter, the bias of the initial estimator was quite small (less than 2%). Performing the TMLE fluctuation step causes neither an improvement nor substantial decline in performance here.

**Vector-valued parameter.** For the projection of $m_0(W)$ unto a linear working model, the true two-dimensional parameter of interest is $[-64.2, 32.3]$.

2SLS solves the following optimization in the second stage:
$\arg\min_{\beta_1,\beta_2} \sum_{i=1}^n (Y - \Pi(Z, W)\beta_1^T\binom{1}{W} - \beta_2^T\binom{1}{W})^2$. $\beta_1$ is output as the parameter of interest. It is easy to check that this can give a very different solution than a semiparametric approach which estimates a function $m(W)$ that can take a variety of functional forms, and then solves $\beta = \arg\min \sum_{i=1}^n (m(W) - \beta^T\binom{1}{W})^2$. Specifically, let $\epsilon_\beta(W) = m(W) - \beta^T\binom{1}{W}$ denote the vector of residuals in approximating $m(W)$ by $\beta^T\binom{1}{W}$. Then in the case of a linear $\theta(W)$, 2SLS solves $\arg\min_\beta \sum_{i=1}^n (\Pi(Z, W)\epsilon_\beta(W))^2$, while the semiparametric approach solves $\arg\min_\beta \sum_{i=1}^n (\epsilon_\beta(W))^2$.

We see in table 2 that 2SLS has a mean absolute bias of around 136. A typical value for its estimate is $[-224, 90]$. It is useless for estimating our parameter of interest without knowing the functional form for $m(W)$ a priori. The confounded estimator that is fully

correctly specified in its functional forms but ignores confounding has a bias of roughly 10. All the semiparametric approaches achieve very low bias when initial estimates are consistent. Furthermore, they all achieve similar and low variance for a large sample size, as the $n = 10000$ column shows. For the sample sizes in our simulation, the 2SLS estimator is not only extremely biased, it also has larger variance than the semiparametric estimators, due to the large mismatch between the second-stage linear model it fits and the data-generating process.

The right-hand side of table 2 shows an incorrect linear fit for $m(W)$ to form an inconsistent initial estimate of $\mu(Z, W)$. The initial substitution estimator works essentially like two-stage least squares in this case. We deliberately start with this enormously biased initial estimator to see if the semiparametric estimators can remove bias sufficiently. Indeed, we see very low finite-sample bias for the three semiparametric consistent estimators. The iterative TMLE-based approach performs best here, with mean absolute bias around just .25 at $n = 10000$ (compared to a mean absolute effect around 48). Furthermore, while the variance of the semiparametric consistent estimators can be an order of magnitude higher than for the initial substitution estimator when $n = 1000$, the variances are at a comparable scale for $n = 10000$.

5.3.2 *Scalar effect, nonlinear design 2.* In table 3, we generate a continuous outcome such that $E(Y|Z, W)$ lies within sharp boundaries covering a much smaller range than $Y$. TMLE using the logistic fluctuation has been shown to be especially effective with similarly generated data, where the data or conditional outcome falls within sharp cutoffs (Gruber and van der Laan 2010).

We use a 3-dimensional $W \sim N(1, 1)$, $p = .5$ for $Z$, a binary treatment generated using the binomial link function. The confounding term is $\epsilon_{AY} \sim N(0, 5)$. $m(W)$ and $\theta(W)$ are continuous, and they each have the form $a \cdot plogis(\beta W) + b$, for some constants $a, b$. Thus,

$m(W)$ and $\theta(W)$ fall within some bounds $[b, a+b]$. Furthermore, the parameters are set so that many values for each function are close to the boundaries.

The true effect is 1.00, and we use $n = 1000$. We see that without using an instrument, the estimate is confounded by more than 50%. For the case of consistently specifying all initial estimates, we include the correct parametric form for $E(Y|Z,W)$ in Super Learner's library. In this case the initial substitution estimator has both lowest bias and lowest variance. The logistic fluctuation and estimating equations estimators also do well with relatively low bias and variance, followed by the iterative and linear fluctuation TMLE, and finally, 2SLS has the highest MSE of the unconfounded estimators. In the right hand of table 3, we misspecify the initial estimate for $E(Y|Z,W)$ as a second-order polynomial. In this case, TMLE using logistic fluctuation is the clear winner. It achieves an MSE (dominated by the variance) of .34, compared to roughly .45 for the other semiparametric approaches. It also achieves a large reduction in bias for minimal gain in variance compared to the initial substitution estimator.

5.3.3 *Vector-valued effect, linear model.* In table 4, we use a linear model for $m(W)$, so that two-stage least squares with the correctly specified cross terms $\Pi(Z,W) \times W$ estimates $\mu(Z,W)$ consistently. Here we use a 3-dimensional covariate $W \sim N(2,1)$, $Z$ is binary and of the form $E(Z|W) = \text{plogis}(\alpha' W + \alpha_0)$. Treatment $A$ is also binary and uses the logit link function; $m(W) = \beta^T \binom{1}{W}$.

We see that although 2SLS uses the correct second-stage specification for $E(Y|W,Z)$, it remains slightly biased for all $n$, with .2 mean absolute bias (about 17%), since $E(A|W,Z)$ uses a nonlinear link function. The confounded estimate has (mean absolute) bias of .34. The semiparametric consistent estimators have much lower bias than 2SLS even for $n = 1000$, with linear fluctuation and estimating equations achieving lower bias than the initial substitution estimator. The table reflects the roughly $\sqrt{n}$ decrease in bias of the consistent estimators and decrease in SD of all estimators. The initial substitution estimator has just

slightly higher SD than 2SLS, as the former chooses the correct linear model from a library of methods.

When we use an inconsistent initial estimate for $\mu(Z,W)$: one of the coefficients in $\beta$ is fixed to an incorrect value and then a linear model is fit (Super Learner is only used for estimating $Pr(Z|W),\ \Pi(Z,W)$). This makes for a mean absolute bias of roughly 1.5 in the initial substitution estimator (corresponding to an error of 100%). The three semiparametric consistent estimates successfully remove bias; the two TMLE-based approaches have particularly low bias (about 94% of the bias is removed for $n = 10000$). The semiparametric estimates have mean SD's of only around .3 for n=10,000 where mean absolute effect is 1.5. The linear fluctuation TMLE-based estimator performs the best overall, with lowest bias and variance for large samples.

5.3.4 *Confidence intervals.* Table 5 shows 95% confidence intervals corresponding to tables 2,4. These are calculated separately for each component of the vector-valued parameter of interest. For the semiparametric estimators, as proved in (van der Laan and Rubin 2006), the following equation holds:

$$\Psi(Q_n^*) - \Psi(Q_0) = \big[P_n - P_0\big]\ D^*(Q_n^*, g_n^*) - \big[P_n - P_0\big]\mathrm{Proj}(D^*(Q_n^*, g_n^*)|\mathrm{Tang}(g_0)) + o_P(\frac{1}{\sqrt{n}})$$

Here $\mathrm{Proj}(D^*(Q_n^*, g_n^*)|\mathrm{Tang}(g_0))$ denotes the projection of the efficient influence curve $D^*$ unto the tangent space of nuisance parameters, $T(g_0)$. It thus follows that a conservative estimator for the variance of $\beta_n^* = \Psi(Q_n^*)$ is the variance of $D^*(Q^*, g^*)$. Note that when all its components are consistently estimated, under regularity conditions, $[P_n - P_0]\ D^*(Q^*, g^*) = [P_n - P_0]\ D^*(Q_0, g_0) + o_P(\frac{1}{\sqrt{n}})$, and thus the semiparametric efficiency bound is achieved. For the three semiparametric consistent estimators, shown at the top of the list in table 5, we use the estimated variance of the efficient influence curve $D^*(Q^*, g^*)$ to calculate confidence interval width. For the other three estimators, we simply use the empirical variance. For

these cases, we demonstrate that even when "cheating" by accurately knowing the correct width of the confidence intervals, coverage is still very poor due to the bias of the estimators.

We see that for all three semiparametric estimators, the coverage is generally overestimated, as the theory suggests, but is usually not too far from 95%. For the case of consistent initial estimates, coverage is around 96% when estimating a linear treatment effect and closer to 97% when estimating a nonlinear effect. Similar results holds when using misspecified initial estimates; however, estimating equations has poor coverage (in the 80's) due to finite-sample bias. The initial substitution estimator is consistent when the initial estimates of components are; however, it has coverage slightly below 95% even when using the empirical variance to estimate the variance. This could be due to its not being normally distributed. When the initial estimates of components is not consistent, the initial substitution estimator can be heavily biased, and we see 0 coverage for most columns, even using an accurate variance. Likewise the large bias of the confounded and 2SLS estimators for the case of the nonlinear treatment effect causes 0 coverage. When a linear treatment effect is estimated, both the confounded and 2SLS estimators exhibit poor coverage that deteriorates with $n$. In the case of 2SLS, the bias is due to the mismatch between the linear model and the nonlinear distribution of the conditional treatment $\Pi(Z, W)$.



[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

## 6. Application to a dataset: estimating the effect of parents' education on infant health

We apply our TMLE-based estimators in the context of a program that expanded schooling in Taiwan. In 1968, Taiwan expanded mandatory schooling from 6 years to 9 years, and more than 150 new junior high schools were opened in 1968-1973 to accommodate this program. Prior to this expansion of schools, enrollment in junior high was based on a competitive process in which only part of the population of 12-14 year-old children was accepted. There is significant variation in how much the schooling expansion program affected an individual's access to education based on the individual's birth cohort and county of residence. In counties where there were previously relatively few educated people and spots in school beyond grade 6, many new junior high schools were opened per child. Thus, program intensity as a function of birth cohort and county serves as an instrumental variable that causes exogenous variation in people's educational attainment. This lets one make a consistent estimate of the effect of parents' education on their child's health.

The school expansion program caused junior high enrollment to jump from 62% to 75% within a year in 1968, before leveling off around 84% in 1973.

We use the same dataset as (Chou et al 2010). The treatment variable is either the mother's or the father's education in years (starting from first grade). There are four outcomes we study: low birth weight ($< 2500$g), neonatal mortality (in the first 27 days after birth), postneonatal mortality (between day 28 and 365), and infant mortality (either neonatal or postneonatal). The instrument is the cumulative number of new junior highs opened in a county by the time a birth cohort reaches junior high, per 1000 children age 12-14 in that year. This serves as a proxy for the intensity of the school expansion program for a particular birth-county cohort. The data is taken by checking every birth certificate for children born in Taiwan between 1978 and 1999. The birth certificates list for both parents their ages,

number of years of education, and county of birth (which we use as a usually correct guess for the county in which the parent went to school), as well as the incidence of low birth weight. Birth certificates are matched to death certificates from a similar period using a unique identification number issued for each person born to ascertain if an infant death has taken place. The previous study done on this dataset (Chou et al 2010) used standard OLS and 2SLS, which are sensitive to highly collinear regressors, and as a result separately estimated the effect of father's and mother's education on infant's health. To ease comparison with prior results, we do the same here. Only datapoints where the father was born in [1943-1968], or the mother in [1948-1968] were included in the study. Those points where the parent was at most 12 years old in 1968 constitute the treatment group, and the rest the control group where the instrument $Z$ is 0 (for those who were unaffected by the school reform). This resulted in a sample size of about 6.5 million, of which roughly 4 million were in the treatment group, for either case of parent.

We reestimate (Chou et al 2010)'s scalar effect estimates using our TMLE-based approach. We also give previously unpublished estimates of treatment effect heterogeneity as a function of the parent's and children's birth cohorts.

The usefulness of the semiparametric approaches depend on the $\sigma^2(W) = \mathrm{Var}(\Pi(Z,W)|W)$ term being large (recall $\Pi(Z,W) = E(A|Z,W)$). This term captures the strength of the instrument in predicting the treatment given $W$, and the variance of the instrument-based estimators blow up when $\sigma^2$ is small. Our instrument only depends on the parent's birth cohort and the county, so $\sigma^2$ would be 0 if we include both these variables in $W$. Since most variation in $Z$ is by county (of parent's birth), we do not include county in $W$, and use as covariates $W$ only parent's and child's birth cohort, coding these as dummy variables. In addition, we remove datapoints where $\sigma^2(W) = 0$, which corresponds to including only

points where the parent was born after or in 1956. People born earlier were unaffected by the schooling reform.

We need to check that county (parent's county of birth) does not serve as a confounder causing $U_Z$, $U_Y$ to be correlated. Using modified outcome $Y'$ (the log-odds ratio for a binary health outcome, see below), we compare the between-county vs the within-county variation. We see that, for any of the 4 health outcomes, and using either mother's or father's county, fixing $W$, at most 1.1% of the variation in $Y'$ is between-county, but on average only .5%. Thus, we can rule out that confounding from county will effect our estimates. The IV-assumptions in section 3 are satisfied.

Table 8 shows summary statistics. Note that for the outcome of postneonatal mortality, we only include datapoints where the child survived the neonatal period.

[Table 6 about here.]

We perform our TMLE-based estimates using the noniterative, linear-fluctuation estimator, as this was found to perform well across multiple simulations, and had low bootstrap variance on the data, suggesting a good fit. We use the same library of initial estimates described above in section 5.4, and the empirical distribution for the probability of a county given the birth cohorts, $\Pr(Z|W)$. Since our outcomes are binary with relatively few positives, and the covariates are indicator variables that divide the dataset into cells, we modify our dataset $(W, Z, A, Y) \longrightarrow (W, Z, \bar{A}, Y')$ when forming initial estimates $\Pi(Z, W),\ m(W),\ \theta(W)$. $\bar{A}_i$ is the average value of education $A$ in the $i^{th}$ cell given by the parent and child's birth cohorts and the county (thus, fixing $W$ and $Z$). $Y'$ is the log-odds ratio given by Cox's modified logistic transformation: $Y'_i = \log \frac{N_i + .5}{D_i - N_i + .5}$, where there are $D_i$ total points in the $i$-th cell, and $N_i$ of these are 1, for one of the four outcomes of interest.

Table 9 gives estimates of the scalar treatment effects. For the OLS and 2SLS estimates,

we include the parent and child's birth cohorts as covariates, with heteroscedasticity-robust standard errors (White's method as implemented in R's sandwich package).

We use the final semiparametric model of the components that TMLE fits ($P_W, \Pi(Z,W)$, etc...), as well as a linear 2SLS model to estimate the number of adverse infant health outcomes prevented by schooling reform. Using our modified log-odds outcome $Y_i' = \log \frac{N_i+.5}{D_i-N_i+.5}$, where $i$ indexes a cell, we estimate the counterfactuals for $Y'$ without the schooling reform, denoted $Y'(Z=0)$. We have $Y'(Z=0) = m(W)\Pi(Z=0,W) + \theta(W)$, where $\Pi(Z=0,W)$ estimates the counterfactual $E(A(Z=0)|W)$. Then we convert from $Y_i'(Z=0)$ to $N_i(Z=0)$, which is the (counterfactual) number of adverse outcomes in a cell. $\Delta N = \sum_{cells} N_i(Z=0) - N_i$ gives the estimated total reduction in an adverse outcome from the schooling reform. We also show the linear 2SLS model's estimate. In this case, $Y'(Z=0)$ simplifies to $(1,(1,0,W)'\beta_1, W)^T(\beta_2)$, where $\beta_1,\ \beta_2$ are the first- and second-stage coefficients, indexing $(1,Z,W)^T$, and $(1,A,W)^T$, respectively.

As table 9 shows, estimates of the scalar effect of (a parent's) education on the log-odds ratio of (infant's) health outcome range from -.2 to -1.0. The estimated percent reduction in adverse outcomes ranges from 1.5% for low birthweight (father's education is treatment, TMLE is the estimator) to 16.7% for neonatal mortality (with mother's education, TMLE estimator). The results imply a significant human benefit from the schooling reform regarding health: our TMLE estimator estimates roughly 1850 infant deaths were spared as an indirect effect of schooling reform.[4] The TMLE estimator finds a significantly greater reduction in adverse outcomes than 2SLS when the outcome is neonatal mortality and mother's education is the treatment, and for infant mortality when father's education is the treatment. TMLE and 2SLS yield similar estimates for the effect for low-birthweight/mother's-education and postneonatal-mortality/father's-education, while TMLE gives a somewhat lower estimate

[4]This estimate is made using semiparametric, TMLE-based estimates of the effect of father's education on reducing infant mortality in the treated population.

than 2SLS in the remaining two cases. The beneficial effect of father's education on infant and postneonatal mortality was highly significant for either estimation method, while the effect of mother's education on neonatal mortality was highly significant only for the TMLE estimator.

The use of a library of learners (TMLE) instead of a linear parametric model (2SLS) is reflected in the better fit and higher cross-validated $R^2$ value achieved for both stages. The "first stage" of the method of instruments refers to fitting $\Pi(Z, W)$ in our semiparametric model, and the "second stage" to fitting $\mu(Z, W)$. Especially for the second stage of father's education, there is a large gain in $R^2$ of .2-.3 from using data-adaptive learning. Super-Learner chooses a least-squares linear model with largest weight in every case; however, our semiparametric model for $E(Y|Z, W) = \Pi(Z, W)m(W) + \theta(W)$ even when $m(W)$, $\theta(W)$ are set to be linear in $W$ is more flexible than the standard linear 2SLS model $E(Y|Z, W) = \beta_A E(A|Z, W) + \beta_W^T \binom{1}{W}$, and we include quadratic terms in $W$. Support vector machine is also chosen with large weight for both stages, and Random Forest for the first stage. The instrument is slightly stronger for predicting mother's education than father's, which might explain the higher first-stage $R^2$ values for mother's education.

As expected, our semiparametric estimator typically has higher variance than 2SLS; however, this is not always true, as TMLE achieves a better fit, which can make for a lower variance despite the added complexity of choosing from various learners.

We had expected that OLS would be biased from unmeasured confounding between a parent's education and his/her infant's health. One would expect that confounding factors would increase parents' education and decrease adverse health effects, or vice versa, biasing the OLS estimates to overestimate the beneficial effects of education. Surprisingly, we saw that the OLS estimates were smaller in magnitude that either of the instrument-based estimates for several columns in our table. One possible explanation is there might not have

been significant unmeasured confounding. Indeed, the Hausman-Wu (F-) test for exogeneity gives low evidence in support of confounding.

In tables 10-11, we estimate the treatment effect modification, where the parent's or child's membership in a particular birth cohort is a modifier (given by the dummy variables $W$). As before, we estimate a vector-valued parameter $\beta$ using a linear working model $\{m_\beta(W) = \beta^T \binom{1}{W} | \beta\}$. Since all covariates in $W$ are binary indicators for birth cohorts, the coefficients in $\beta$ can be directly compared to one another to reflect treatment effect modification.

The six effect modifiers that are largest in magnitude for each case are shown. The child being born in the late 70's or 80's often corresponded to a substantial increase in the beneficial effects of parent's education. The father being born in 1965, 1967, or 1968 corresponded to increased beneficial effects of his education on his infant's mortality. However, the effect of mother's education on her child's good health was found be diminished for babies born in 1998 or 1999. Virtually all the treatment effect modifiers shown are highly significant for mothers, as well as for fathers when postneonatal mortality is the outcome. The largest magnitude effect modifiers were not necessarily the most statistically significant ones, so the treatment effect modifiers are summarized for each case both as original and as standardized values (effect modifier ÷ SE). There were roughly 33 total effect modifiers. We see that for some cases, a significant fraction of the effect modifiers had a coefficient of statistically significant magnitude (neonatal mortality for mothers, and low birthweight and postneonatal mortality for fathers).

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

## 7. Discussion

We consider the problem of estimating the causal effect of a treatment on an outcome in the presence of unmeasured confounding and a valid instrument. Assuming the treatment effect $(Y(A=a)-Y(A=0))$ is a function of covariates $W$, we are interested in the average effect of treatment given an arbitrary subset $V$ of $W$. Our causal parameter of interest is the projection of this treatment effect $E(Y(A=a)-Y(A=0)|V)$, as modulated by variables $V$, unto a user-supplied parametric model. We derive our solution in a highly general framework compared to prior work. We allow a binary or continuous instrument and treatment, and use a fully semiparametric model that invokes minimal assumptions to ensure identification in the instrumental variables setting.

Our solution is based on the *targeted minimum loss-based estimation* (TMLE) methodology. A first step is to find the efficient influence curve of the parameter of interest. We do so both for the general semiparametric case, as well as for the case when the treatment effect has a parametric form. The TMLE procedure is to construct initial estimates of certain components of the data-generating distribution, then fluctuate some of the components in a direction that optimizes efficiency while removing bias. We describe three different implementations of the TMLE procedure for this problem, and demonstrate in simulations that each of these implementations has its advantages. Our estimators have a number of desirable properties both theoretically and empirically.

Our simulations reflect that even compared to a parametric estimator for the scalar effect of interest, such as two-stage least squares, the semiparametric efficient estimates can have both lower bias and far lower variance due to the better fit with relevant components of the data-generating distribution. We also showed that two-stage least squares can be enormously biased when estimating a vector-valued parameter, while TMLE is very effective at removing bias with only a moderate gain in variance in finite samples. Using TMLE with a logistic

fluctuation can give the best performance when the conditional mean of the outcome follows sharp cutoffs, and each of the three TMLE-based estimators we describe has datasets on which is it the strongest performer. Finally, using the (estimated) variance of the efficient influence curve to estimate the standard error gives confidence intervals that are just slightly conservative. The confidence intervals based on TMLE can perform better than those based on a conventional semiparametric estimator.

We performed an extensive data analysis estimating the effect of parents' education on their infant's health in the context of a schooling reform in Taiwan. We identified a number of birth cohorts, pertaining to either the parent or the infant, that significantly increased, or decreased, the beneficial effect of education on health.

Several avenues for future work are of interest. One is to work with instrumental variables in the context of more complex causal models, such as when there are multiple instruments and treatments. This may for example occur in the setting of longitudinal data where each time point has an instrument, or in the context that a multivariate instrument is used to control for a multivariate confounded treatment. A number of extensions are of interest along empirical lines as well. For instance, future work could apply our methods to data having a very high-dimensional covariate space $W$, where $V$ is a tiny subset of $W$, in finding the effect of the treatment given $V$.

The authors gratefully acknowledge the support of the National Institutes of Health, through NIAID grant number 5R01AI074345.



## References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics, Vol. 113, p.213-263* .

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using

instrumental variables. *Journal of the American Statistical Association, Vol. 91, p. 444-471* .

Brookhart, M. A., Rassen, J. A., and Schneeweiss, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety, Vol. 19, Issue 6, p. 537-554* .

Brookhart, M. A. and Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects. *International Journal of Biostatistics: Vol.3(1), p.1-14* .

Cheng, J., Small, D., Tan, Z., and Have, T. T. (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika: No. 96, p.1-9* .

Chou, S.-Y., Liu, J.-T., Grossman, M., and Joyce, T. (2010). Parental education and child health: Evidence from a natural experiment in taiwan. *American Economic Journal: Applied Economics, Volume 2, Issue 1, p. 33-61* .

Gruber, S. and van der Laan, M. J. (2010). A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics: Vol. 6, Issue 1, Article 26* .

Hong, H. and Nekipelov, D. (2010). Semiparametric efficiency in nonlinear late models. *Quantitative Economics, Vol. 1, Issue 2* .

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica, Vol. 62, p. 467-475* .

Kasy, M. (2009). Semiparametrically efficient estimation of conditional instrumental variable parameters. *International Journal of Biostatistics, Vol. 5, Issue 1, Article 2* .

Newey, W. (1990). Semiparametric efficiency bounds. *Journal of the Applied Econometrics, Vol. 5, No. 2, p. 99-135* .

Newhouse, J. P. and McClellan, M. (1998). Econometrics in outcomes research: the use of

instrumental variables. *Annual Review of Public Health, Vol. 19, p. 17-34* .

Ogburn, E., Rotnizky, A., and Robins, J. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society, Series B (in press)* .

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Robins, J. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics: Vol. 23, p.2379-2412* .

Robins, J. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, Vol. 11, p.550-560* .

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology: Vol. 66, No. 6, p.689* .

Sinisi, S. and van der Laan, M. (2004). Loss-based cross-validation deletion/substitution/addition algorithms in estimation. *Technical report, UC Berkeley Division of Biostatistics Working Paper Series No. 143* .

Stock, J., Wright, J., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of the American Statistical Association, Vol. 20, Issue 4, p. 518-529* .

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

Tan, Z. (2010). Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association, Vol. 105, p.157-169* .

Terza, J., Bradford, D., and Dismuke, C. E. (2008). The use of linear instrumental variables methods in health services research and health economics: A cautionary note. *Health Services Research: Vol. 43, Issue 3, p.1102-1120* .

Uysal, S. (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 .).

van der Laan, M., Hubbard, A., and Jewell, N. (2007). Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 69.* .

van der Laan, M. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. New York, Springer.*

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology: Vol. 6, Issue 1, Article 25* .

van der Laan, M. J. and Robins, J. M. (2003). Unified methods for censored longitudinal data and causality. springer verlag: New york. Technical report.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics: Vol. 2, Issue 1, Article 11* .

## Appendix 1: Proofs of properties of the TMLE-based estimators

***Consistency*** *under partial misspecification.*

TMLE is constructed so that the efficient influence curve equation holds. We can explicitly write this as a function of the final estimate $\Psi^* = \beta^*$ using the definition of $\beta$. Thus we have $P_n D^*(Q^*, g^*, \beta^*) = 0$ (we drop the $n$-subscript notation here). Since $P_n D^*(Q^*, g^*, \beta^*)$ converges to $P_0 D^*(Q', g', \beta')$, where $\{Q', g', \beta'\}$ are the components in the limiting distribution, when the true parameter of interest $\beta' = \beta_0$ solves $P_0 D^*(Q', g', \beta') = 0$, for some case of consistent specification of some of $\{Q', g'\}$, then we have that $\beta^* \longrightarrow \beta_0$ for our TMLE estimators.

Simplifying slightly, we get that $P_0 D^*(Q, g, \beta) = 0$ reduces to

$$P_0 \; c_0^{-1} h_1 (m - m_\beta) \tag{A.1}$$

$$+ P_0 \; c_0^{-1} \frac{h_1}{\sigma^2} (\Pi - E(\Pi))((m_0 - m)\Pi_0 - (\theta_0 - \theta)) \tag{A.2}$$

$$= 0 \tag{A.3}$$

TMLE yields a consistent estimate for $\Psi^* = \beta^*$ under 3 scenarios of partial misspecification of components given below, with the reasoning sketched. Note that for the non-iterative versions of TMLE, only $m^0$ is updated, and the initial estimates are the same as the final estimates for the other components. For iterative TMLE, it is easy to check that when an initial estimate for a component is consistent, so is the final estimate (i.e. at every step $k$, $\epsilon_A^k \to 0$ when $\Pi^0$ is consistent ).

(1) Initial estimates $\Pi^0$ and $\Pr^0(Z|W)$ are consistent.

We have $E_0(\ (\Pi_0(Z,W) - E_0(\Pi_0)|W)\Pi_0(Z,W)\ |W) = \sigma^2(W)$ since $\Pi, E\Pi|W$ are correctly specified. Also, since $E_0(\ (\Pi_0(Z,W) - E_0(\Pi_0)|W)\ |W) = 0$, the term involving $(\theta_0(W) - \theta(W))$ is 0 in expectation. Thus, A.2 reduces to

$P_0\ c_0^{-1}h_1(m_0 - m)$, so $P_0 D^*(Q,g,\beta) = 0$ becomes $P_0\ c_0^{-1}h_1(m_0 - m_\beta) = 0$, and this is solved by $\beta = \beta_0$ by definition of $\beta$.

(2) Initial estimates $m^0$ and $\Pr^0(Z|W)$ are consistent.

The term in A.2 involving $(m_0 - m)$ is 0 by the consistency of $m$, and the term involving $(\theta_0 - \theta)$ is also 0 since $E_0(\ \Pi(Z,W) - E_0(\Pi(Z,W)|W)\ |W) = 0$. Thus, we have $P_0\ c_0^{-1}h_1(m_0 - m_\beta) = 0$, which is solved by $\beta = \beta_0$ by definition of $\beta$.

(3) Initial estimates $m^0$ and $\theta^0$ are consistent.

A.2 goes to 0 because both $m_0 - m = 0$, $\theta_0 - \theta = 0$. The rest of the reasoning is the same as above.

***Efficiency*** *under correct specification of all relevant components and nuisance parameters.* (See van der Laan and Robins 2003, and van der Laan and Rubin 2006.)

Suppose all initial estimates $(Q_n^0, g_n^0)$ are consistent, and that $\mathrm{Var}(D^*(Q_n^*, g_n^*) - D^*(Q_0, g_0)) \in o(1)$. Then the final estimate $\Psi(Q_n^*)$ is asymptotically efficient, with

$$\Psi(Q_n^*) - \Psi(Q_0) = \left[P_n - P_0\right]\ D^*(Q_0, g_0) + o_p(1/\sqrt{n}) \tag{A.4}$$

Sketch of proof: Note that when all initial estimates are consistent, then so are all final estimates $(Q_n^*, g_n^*)$. In the non-iterative case, only $m_n^0(W)$ is updated and $m_n^* \to m_n^0$ when the other components are consistent (see Consistency proof above). Using the definition of the canonical gradient $D^*$ at $(Q_n^*, g_0)$ and taking a Taylor expansion (see van der Laan and Robins 2003), we have

$$\Psi(Q_n^*) - \Psi(Q_0) = -P_0\ D^*(Q_n^*, g_0) + o_p(1/\sqrt{n}) \tag{A.5}$$

We can expand the first term on rhs into

$$-P_0\ D^*(Q_n^*, g_0) = -P_0\ D^*(Q_n^*, g_n^*) + \Big[P_0\ D^*(Q_n^*, g_n^*) - P_0\ D^*(Q_n^*, g_0)\Big] \tag{A.6}$$

The expression in brackets is equal to an empirical process-like expression involving the projection unto the tangent space of $g_0$:

$$\big[P_n - P_0\big]\ (\text{Proj}(D^*(Q_n^*, g_n^*))|\text{Tang}(g_0)) + o_P(1/\sqrt(n)) \tag{A.7}$$

Rewriting equation A.5, using A.6, A.7 and the key property of TMLE that $P_n D^*(Q_n^*, g_n^*) = 0$, we get

$$\begin{aligned}
\Psi(Q_n^*) - \Psi(Q_0) &= \big[P_n - P_0\big]\ D^*(Q_n^*, g_n^*) + \big[P_n - P_0\big]\ (\text{Proj}(D^*(Q_n^*, g_n^*))|\text{Tang}(g_0)) + o_p(1/\sqrt{n}) \\
&= \big[P_n - P_0\big]\ D^*(Q_0, g_0) + \big[P_n - P_0\big](D^*(Q_n^*, g_n^*) - D^*(Q_0, g_0)) + o_p(1/\sqrt{n}) \\
&= \big[P_n - P_0\big]\ D^*(Q_0, g_0) + o_p(1/\sqrt{n})
\end{aligned}$$

## Appendix 2: Efficient influence curve of target parameter

We determine the efficient influence curve of $\Psi : \mathcal{M} \to \mathbb{R}^d$ in a two step process. Firstly, we determine the efficient influence curve in the model in which $\Pi_0$ is assumed to be known. Subsequently, we compute the correction term that yields the efficient influence curve in our model of interest in which $\Pi_0$ is unspecified.

*Efficient influence curve in model in which $\Pi_0$ is known.*

First, we consider the statistical model $\mathcal{M}(\pi_0) \subset \mathcal{M}$ in which $\Pi_0(Z,W) = E_0(A \mid Z,W)$ is known. For the sake of the derivation of the canonical gradient, let $W \in \mathbb{R}^N$ be discrete with support $\mathcal{W}$ so that we can view our model as a high dimensional parametric model, allowing us to re-use previously established results. That is, we represent the semiparametric regression model as $E_0(Y \mid Z,W) = \Pi_0(Z,W)\sum_w m_0(w)I(W=w)+\theta_0(W)$ so that it corresponds with a linear regression $f_{m_0}(Z,W) = \Pi_0(Z,W)\sum_w m_0(w)I(W=w)$ in which $m_0$ represents the coefficient vector. Define the $N$-dimensional vector $h(\Pi_0)(Z,W) = d/dm_0 f_{m_0}(Z,W) = (\Pi_0(Z,W)I(W=w) : w \in \mathcal{W})$. By previous results on the semiparametric regression model, a gradient for the $N$-dimensional parameter $m(P)$ at $P = P_0 \in \mathcal{M}(\pi_0)$ is given by

$$D^*_{m,\Pi_0}(P_0) = C(\pi_0)^{-1}(h(\Pi_0)(Z,W) - E(h(\Pi_0)(Z,W) \mid W))(Y - f_{m_0}(Z,W) - \theta_0(W)),$$

where $C(\pi_0)$ is a $N \times N$ matrix defined as

$$\begin{aligned} C(\pi_0) &= E_0\{d/dm_0 f_{m_0}(Z,W) - E_0(d/dm_0 f_{m_0}(Z,W) \mid W)\}^2 \\ &= E_0\{(I(W=w)\{\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W\} : w\}^2 \\ &= \text{Diag}\left(E_0\{I(W=w)\{\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W=w)\}^2\} : w\right) \\ &= \text{Diag}\left(P_{W,0}(w)E_0\left(\{\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W)\}^2 \mid W=w\right) : w\right). \end{aligned}$$

For notational convenience, given a vector $X$, we used notation $X^2$ for the matrix $XX^\top$. We also used the notation $\text{Diag}(x)$ for the $N \times N$ diagonal matrix with diagonal elements defined by vector $x$. Thus, the inverse of $C(\pi_0)$ exists in closed form and is given by:

$$C(\pi_0)^{-1} = \text{Diag}\left(\frac{1}{P_{W,0}(w)E_0(\{\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W)\}^2 \mid W=w)} : w\right).$$

This yields the following formula for the efficient influence curve of $m_0$ in model $\mathcal{M}(\pi_0)$:

$$D^*_{m,\Pi_0,w}(P_0) = \frac{1}{P_{W,0}(w)E_0(\{\Pi_0(Z,W) - E_0(\Pi_0(Z,W)|W)\}^2|W=w)}$$
$$I(W=w)(\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W))(Y - \Pi_0(Z,W)m_0(W) - \theta_0(W)),$$

where $D^*_{m,\Pi_0}(P_0)$ is $N \times 1$ vector with components $D^*_{m_0,\Pi_0,w}(P_0)$ indexed by $w \in \mathcal{W}$. We can further simplify this as follows:

$$D^*_{m,\Pi_0,w}(P_0)(W,Z,Y) = \frac{1}{P_{W,0}(w)E_0(\{\Pi_0(Z,W)-E_0(\Pi_0(Z,W)|W)\}^2|W=w)}$$
$$I(W=w)(\Pi_0(Z,w) - E_0(\Pi_0(Z,W) \mid W=w))(Y - \Pi_0(Z,w)m_0(w) - \theta_0(w)).$$

This gradient equals the canonical gradient of $m_0$ in this model $\mathcal{M}(\pi_0)$, if $E_0((Y - E_0(Y \mid \Pi_0, W))^2 \mid Z, W)$ is only a function of $W$. For example, this would hold if $E(U_Y^2 \mid Z, W) = E_0(U_Y^2 \mid W)$. This might be a reasonable assumption for an instrumental variable $Z$. For the sake of presentation, we work with this gradient due to its relative simplicity. and the fact that it still equals the actual canonical gradient under this assumption.

We have that $\psi_0 = \phi(m_0, P_{W,0})$ for a mapping

$$\phi(m_0.P_{W,0}) = \arg\min_\beta E_0 \sum_a h(a,V)a^2 \left(m_0(W) - m_\beta(V)\right)^2,$$

defined by working model $\{m_\beta : \beta\}$. Let $d\phi(m_0, P_{W,0})(h_m, h_W) = \frac{d}{dm_0}\phi(m_0, P_{W,0})(h_m) + \frac{d}{dP_{W,0}}\phi(m_0, P_{W,0})(h_W)$ be the directional derivative in direction $(h_m, h_W)$. The gradient of $\Psi : \mathcal{M}(\Pi_0) \to \mathbb{R}^d$ is given by $D^*_{\psi,\Pi_0}(P_0) = \frac{d}{dm_0}\phi(m_0, P_{W,0})D^*_{m,\Pi_0}(P_0) + \frac{d}{dP_{W,0}}\phi(m_0, P_{W,0})IC_W$, where $IC_W(O) = (I(W = w) - P_{W,0}(w) : w)$. We note that $\beta_0 = \phi(m_0, P_{W,0})$ solves the following $d \times 1$ equation

$$U(\beta_0, m_0, P_{W,0}) \equiv E_0 \sum_a h(a,V)a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)(m_0(W) - m_{\beta_0}(V)) = 0.$$

By the implicit function theorem, the directional derivative of $\beta_0 = \phi(m_0, P_{W,0})$ is given by

$$d\phi(m_0, P_{W,0})(h_m, h_W) = -\left\{\frac{d}{d\beta_0}U(\beta_0, m_0, P_{W,0})\right\}^{-1}$$
$$\left\{\frac{d}{dm_0}U(\beta_0, m_0, P_{W,0})(h_m) + \frac{d}{dP_{W,0}}U(\beta_0, m_0, P_{W,0})(h_W)\right\}.$$

We need to apply this directional derivative to $(h_m, h_W) = (D^*_{m,\Pi_0}(P_0), IC_W)$. Recall we assumed that $m_\beta$ is linear in $\beta$. We have

$$c_0 \equiv -\frac{d}{d\beta_0}U(\beta_0, m_0) = E_0 \sum_a h(a,V)a^2 \left\{\frac{d}{d\beta_0} m_{\beta_0}(V)\right\}^2,$$

which is a $d \times d$ matrix. Note that if $m_\beta(V) = \sum_j \beta_j V_j$, then this reduces to

$$c_0 = E_0 \sum_a h(a,V)a^2 \vec{V}\vec{V}^\top,$$

where $\vec{V} = (V_1, \ldots, V_d)$. We have

$$\frac{d}{dP_{W,0}} U(\beta_0, m_0, P_{W,0})(h_W) = \sum_w h_W(w) \sum_a h(a,v) a^2 \frac{d}{d\beta_0} m_{\beta_0}(v)(m_0(w) - m_{\beta_0}(v)).$$

Thus, the latter expression applied to $IC_W(O)$ yields $c_0^{-1} D^*_W(P_0)$, where

$$D^*_W(P_0) \equiv \sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)(m_0(W) - m_{\beta_0}(V)).$$

In addition, the directional derivative $\frac{d}{d\epsilon} U(\beta_0, m_0 + \epsilon h_m, P_{W,0})|_{\epsilon=0}$ in the direction of the function $h_m$ is given by

$$\frac{d}{dm_0} U(\beta_0, m_0, P_{W,0})(h_m) = E_0 \sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) h_m(W).$$

We conclude that

$$d\phi(m_0, P_{W,0})(h_m, h_W) = D^*_W(P_0) + c_0^{-1} \left\{ E_0 \sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) D^*_{m,W}(P_0) \right\}.$$

We conclude that the canonical gradient of $\Psi : \mathcal{M}(\Pi_0) \to \mathbb{R}^d$ is given by

$$\begin{aligned}
&D^*_{\psi,\Pi_0}(P_0)(O) = D^*_W(P_0)(O) \\
&+c_0^{-1} E_0 \textstyle\sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) D^*_{m,W}(P_0) \\
&= c_0^{-1} \textstyle\sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)(m_0(W) - m_{\beta_0}(V)) \\
&+c_0^{-1} \textstyle\sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) \frac{1}{E_0(\{\Pi_0(Z,W) - E(\Pi_0(Z,W)|W)\}^2|W)} \\
&(\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W))(Y - \Pi_0(Z,W) m_0(W) - \theta_0(W)).
\end{aligned}$$

We state this result in the following lemma and also state a robustness result for this efficient influence curve.

LEMMA 3: *The efficient influence curve of* $\Psi : \mathcal{M}(\Pi_0) \to \mathbb{R}^d$ *is given by*

$$\begin{aligned}
&D^*_{\psi,\Pi_0}(P_0) = c_0^{-1} \textstyle\sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)(m_0(W) - m_{\beta_0}(V)) \\
&+c_0^{-1} \textstyle\sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) \frac{1}{E_0(\{\Pi_0(Z,W) - E(\Pi_0(Z,W)|W)\}^2|W)} \\
&(\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W))(Y - \Pi_0(Z,W) m_0(W) - \theta_0(W)).
\end{aligned}$$

*Assume the linear working model* $m_\beta(V) = \beta \vec{V}$. *Let* $h_1(V) = \sum_a h(a,V) a^2 \vec{V}$. *We have that for all* $\theta$, *(*$d_0$ *below refers to* $\Pr(Z|W)$*):*

$$P_0 D^*_{\psi,\Pi_0}(g_0, m, \theta) = 0 \text{ if } E_0 h_1(V)(m - m_0)(W) = 0,$$

*or, equivalently, if* $\psi \equiv \Psi(m, P_{W,0}) = \Psi(m_0, P_{W,0}) = \psi_0$.

*Efficient influence curve in model in which* $\Pi_0$ *is unknown*

We will now derive the efficient influence curve in model $\mathcal{M}$ in which $\Pi_0$ is unknown, which is obtained by adding a correction term $D_\pi(P_0)$ to the above derived $D^*_{\psi,\Pi_0}(P_0)$. The correction term $D_\pi(P_0)$ that needs to be added to $D^*_{\psi,\Pi_0}$ is the influence curve of $P_0\{D^*_{\psi,\Pi_0}(\pi_n) - D^*_{\psi,\Pi_0}(\pi_0)\}$, where $D^*_{\psi,\Pi_0}(\pi) = D^*_{\psi,\Pi_0}(\beta_0, \theta_0, m_0, d_0, \pi)$ is the efficient influence curve in model $\mathcal{M}(\pi_0)$, as derived above with $\pi_0$ replaced by $\pi$, and $\pi_n$ is the nonparametric NPMLE of $\pi_0$. Let $h_1(V) \equiv \sum_a h(a,v)a^2\frac{d}{d\beta_0}m_{\beta_0}(v)$. Let $\pi(\epsilon) = \pi + \epsilon\eta$. We plug in for $\eta$ the influence curve of the NPMLE $\Pi_n(z,w)$, which is given by

$$\eta(z,w) = \frac{I(Z=z, W=w)}{P_0(z,w)}(A - \Pi(Z,W)).$$

We have

$$\begin{aligned}
D_\pi(P_0) &= \tfrac{d}{d\epsilon}P_0 D^*_\psi(\pi(\epsilon))\big|_{\epsilon=0} \\
&= P_0 c_0^{-1} h_1(V)\left\{-2\tfrac{E_0((\pi-E(\pi|W))(\eta-E(\eta|W))|W)}{E_0((\pi-E(\pi|W))^2|W)}\right. \\
&\qquad\qquad \left.(\pi - E(\pi \mid W)(Y - \pi m_0 - \theta_0)\right\} \\
&+P_0 c_0^{-1} h_1(V)\left\{\tfrac{(\eta-E(\eta|W))(Y-\pi m_0-\theta_0)}{E_0((\pi-E(\pi|W))^2|W)}\right\} \\
&-P_0 c_0^{-1} h_1(V)\left\{\tfrac{(\pi-E(\pi|W))\eta m_0}{E_0((\pi-E(\pi|W))^2|W)}\right\}.
\end{aligned}$$

By writing the expectation w.r.t. $P_0$ as an expectation of a conditional expectation, given $Z, W$, and noting that $E(Y - \pi_0 m_0 - \theta_0 \mid Z, W) = 0$, it follows that the first two terms equal zero. Thus,

$$D_\pi(P_0) = -P_0 c_0^{-1} h_1(V)\left\{\tfrac{(\pi-E_0(\pi|W))\eta m_0}{E_0((\pi-E_0(\pi|W))^2|W)}\right\}.$$

This yields as correction term:

$$\begin{aligned}
D_\pi(P_0) &= -(A - \Pi_0(Z,W))\textstyle\int_{z,w} P_0(z,w)c_0^{-1}h_1(V)\left\{\tfrac{(\pi-E(\pi|W))\frac{I(Z=z,W=w)}{P_0(z,w)}m_0}{E_0((\pi-E(\pi|W))^2|W)}\right\} \\
&= -(A - \Pi_0(Z,W))c_0^{-1}h_1(V)\left\{\tfrac{(\pi(Z,W)-E(\pi(Z,W)|W))m_0(W)}{E_0((\pi(Z,W)-E_0(\pi(Z,W)|W))^2|W)}\right\}.
\end{aligned}$$

This proves the following lemma.

LEMMA 4: *The efficient influence curve of* $\Psi : \mathcal{M} \to \mathbb{R}^d$ *is given by*

$$\begin{aligned}
&D^*(P_0) = D^*_W(P_0)\\
&+c_0^{-1}\tfrac{h_1(V)}{\sigma^2(d_0,\pi_0)(W)}(\pi_0(Z,W) - E_0(\pi_0(Z,W) \mid W))(Y - \pi_0(Z,W)m_0(W) - \theta_0(W))\\
&-c_0^{-1}\tfrac{h_1(V)}{\sigma^2(d_0,\pi_0)(W)}\left\{(\pi_0(Z,W) - E_0(\pi_0(Z,W) \mid W))m_0(W)\right\}(A - \pi_0(Z,W))\\
&\equiv D^*_W(P_0) + C_Y(d_0,\pi_0)(Z,W)(Y - \pi_0(Z,W)m_0(W) - \theta_0(W))\\
&\qquad\qquad -C_A(d_0,\pi_0,m_0)(A - \pi_0(Z,W))\\
&\equiv D^*_W(P_0) + D^*_Y(P_0) - D^*_A(P_0),
\end{aligned}$$

*where*

$$\begin{aligned}
\sigma^2(d_0,\pi_0)(W) &= E_0(\{\Pi_0(Z,W) - E(\Pi_0(Z,W) \mid W)\}^2 \mid W)\\
h(d_0,\pi_0)(W) &= c_0^{-1}\frac{h_1(V)}{\sigma^2(d_0,\pi_0)(W)}\\
C_Y(d_0,\pi_0)(Z,W) &= h(d_0,\pi_0)(W)(\pi_0(Z,W) - E_{d_0}(\pi_0(Z,W) \mid W))\\
C_A(d_0,\pi_0,m_0)(Z,W) &= C_Y(d_0,\pi_0)(Z,W)m_0(W).
\end{aligned}$$

**Double robustness of efficient influence curve:** We already showed $P_0D^*(\pi_0,d_0,m,\theta) = 0$ if $\phi(m,P_{W,0}) = \phi(m_0,P_{W,0})$. If $\phi(m,P_{W,0}) = \phi(m_0,P_{W,0})$ (i.e., $\psi = \psi_0$), then,

$$P_0D^*(\pi,d_0,m,\theta) = P_0\frac{h_1}{\sigma^2(d_0,\pi)}(\pi - P_{d_0}\pi)(\pi_0 - \pi)(m_0 - m),$$

where we used notation $P_{d_0}h = E_{d_0}(h(Z,W) \mid W)$ for the conditional expectation operator over $Z$, given $W$. This is thus second order in $(m - m_0)(\pi - \pi_0)$. In particular, it equals zero if $m = m_0$ or $\pi = \pi_0$. We can thus also state the following double robustness result: if $m = m_0$, then $P_0D^*(\pi,d,m_0,\theta) = 0$ if $d = d_0$ or if $\pi = \pi_0$.

## APPENDIX 3: EFFICIENT INFLUENCE CURVE OF TARGET PARAMETER WHEN ASSUMING A PARAMETRIC FORM FOR EFFECT OF TREATMENT AS FUNCTION OF COVARIATES



We now assume $m_0 = m_{\alpha_0}$ for some model $\{m_\alpha : \alpha\}$, which implies the semiparametric regression model $E_0(Y \mid Z,W) = \Pi_0(Z,W)m_{\beta_0}(W) + \theta_0(W)$. Let $f_\beta(Z,W) = \Pi_0(Z,W)m_\beta(W)$.

Let $m_\alpha(W) = \alpha^\top W^*$, where $W^*$ is $k$-dimensional vector of functions of $W$. Note that $\alpha$ is $d$-dimensional and $\frac{d}{d\alpha} m_\alpha(W) = W^*$.

*Efficient influence curve in model in which $\Pi_0$ is known.*

First, we consider the statistical model $\mathcal{M}(\pi_0) \subset \mathcal{M}$ in which $\Pi_0(Z,W) = E_0(A \mid Z,W)$ is known. Define the $k$-dimensional vector

$$h(\Pi_0)(Z,W) = d/\alpha_0 m_{\alpha_0}(Z,W) = \Pi_0(Z,W) d/d\alpha_0 m_{\alpha_0}(W) = \Pi_0(Z,W) W^*.$$

By previous results on the semiparametric regression model, a gradient for the $k$-dimensional parameter $\alpha(P)$ at $P = P_0 \in \mathcal{M}(\pi_0)$ is given by

$$D^*_{\alpha,\Pi_0}(P_0) = C(\pi_0)^{-1}(h(\Pi_0)(Z,W) - E(h(\Pi_0)(Z,W) \mid W))(Y - f_{\alpha_0}(Z,W) - \theta_0(W)),$$

where $C(\pi_0)$ is a $k \times k$ matrix defined as

$$\begin{aligned} C(\pi_0) &= E_0\{d/d\alpha_0 f_{\alpha_0}(Z,W) - E_0(d/d\alpha_0 f_{\alpha_0}(Z,W) \mid W)\}^2 \\ &= E_0\{(W^* W^{*\top}\{\Pi_0(Z,W) - E_0(\Pi_0(Z,W) \mid W\}^2\}. \end{aligned}$$

Let $C(\pi_0)^{-1}$ be the inverse of $C(\pi_0)$.

This gradient equals the canonical gradient of $\alpha_0$ in this model $\mathcal{M}(\pi_0)$, if $E_0((Y - E_0(Y \mid \Pi_0, W))^2 \mid Z, W)$ is only a function of $W$. For example, this would hold if $E(U_Y^2 \mid Z,W) = E_0(U_Y^2 \mid W)$. This might be a reasonable assumption for an instrumental variable $Z$. For the sake of presentation, we work with this gradient due to its relative simplicity. and the fact that it still equals the actual canonical gradient under this assumption.

We have that $\psi_0 = \phi(\alpha_0, P_{W,0})$ for a mapping

$$\phi(\alpha_0.P_{W,0}) = \arg\min_\beta E_0 \sum_a h(a,V) a^2 \left(m_{\alpha_0}(W) - m_\beta(V)\right)^2,$$

defined by working model $\{m_\beta : \beta\}$. Let $d\phi(\alpha_0, P_{W,0})(h_\alpha, h_W) = \frac{d}{d\alpha_0}\phi(\alpha_0, P_{W,0})(h_\alpha) + \frac{d}{dP_{W,0}}\phi(\alpha_0, P_{W,0})(h_W)$ be the directional derivative in direction $(h_\beta, h_W)$. The gradient of $\Psi : \mathcal{M}(\Pi_0) \to \mathbb{R}^d$ is given by $D^*_{\alpha,\Pi_0}(P_0) = \frac{d}{d\alpha_0}\phi(\alpha_0, P_{W,0}) D^*_{\alpha,\Pi_0}(P_0) + \frac{d}{dP_{W,0}}\phi(\alpha_0, P_{W,0}) IC_W$, where $IC_W(O) = (I(W = w) - P_{W,0}(w) : w)$ is the influence curve of the empirical

distribution of $W$. We note that $\beta_0 = \phi(\alpha_0, P_{W,0})$ solves the following $d \times 1$ equation

$$U(\beta_0, \alpha_0, P_{W,0}) \equiv E_0 \sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)(m_{\alpha_0}(W) - m_{\beta_0}(V)) = 0.$$

By the implicit function theorem, the directional derivative of $\beta_0 = \phi(\alpha_0, P_{W,0})$ is given by

$$\begin{aligned} d\phi(\alpha_0, P_{W,0})(h_\alpha, h_W) = &- \left\{ \tfrac{d}{d\beta_0} U(\beta_0, \alpha_0, P_{W,0}) \right\}^{-1} \\ &\left\{ \tfrac{d}{d\alpha_0} U(\beta_0, \alpha_0, P_{W,0})(h_\alpha) + \tfrac{d}{dP_{W,0}} U(\beta_0, \alpha_0, P_{W,0})(h_W) \right\}. \end{aligned}$$

We need to apply this directional derivative to $(h_\alpha, h_W) = (D^*_{\alpha,\Pi_0}(P_0), IC_W)$. Recall we assumed that $m_\beta$ is linear in $\beta$. We have

$$c_0 \equiv -\frac{d}{d\beta_0} U(\beta_0, \alpha_0, P_{W,0}) = E_0 \sum_a h(a,V) a^2 \left\{ \frac{d}{d\beta_0} m_{\beta_0}(V) \right\}^2,$$

which is a $d \times d$ matrix. Note that if $m_\beta(V) = \sum_j \beta_j V_j$, then this reduces to

$$c_0 = E_0 \sum_a h(a,V) a^2 \vec{V} \vec{V}^\top,$$

where $\vec{V} = (V_1, \ldots, V_d)$. We have

$$\frac{d}{dP_{W,0}} U(\beta_0, \alpha_0, P_{W,0})(h_W) = \sum_w h_W(w) \sum_a h(a,v) a^2 \frac{d}{d\beta_0} m_{\beta_0}(v)(m_{\alpha_0}(w) - m_{\beta_0}(v)).$$

Thus, the latter expression applied to $IC_W(O)$ yields the contribution $c_0^{-1} D^*_W(P_0)$, where

$$D^*_W(P_0) \equiv \sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V)(m_{\alpha_0}(W) - m_{\beta_0}(V)).$$

In addition,

$$\frac{d}{d\alpha_0} U(\beta_0, \alpha_0, P_{W,0}) = E_0 \sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) \frac{d}{d\alpha_0} m_{\alpha_0}(W).$$

We conclude that

$$\begin{aligned} &d\phi(\alpha_0, P_{W,0})(h_\alpha, h_W) = \\ &D^*_W(P_0) + c_0^{-1} \left\{ E_0 \textstyle\sum_a h(a,V) a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) \frac{d}{d\alpha_0} m_{\alpha_0}(W) D^*_{\alpha,\Pi_0}(P_0) \right\}. \end{aligned}$$

We conclude that the canonical gradient of $\Psi : \mathcal{M}(\Pi_0) \to \mathbb{R}^d$ is given by

$$\begin{aligned} D^*_{\psi,\Pi_0}(P_0) &= D^*_W(P_0)(O) \\ &\quad + c_0^{-1}\left\{E_0 \sum_a h(a,V)a^2 \frac{d}{d\beta_0} m_{\beta_0}(V) \frac{d}{d\alpha_0} m_{\alpha_0}(W)\right\} D^*_{\alpha,\Pi_0}(P_0)(O) \\ &= D^*_W(P_0)(O) + \\ &\quad c_0^{-1}\left\{E_0 h_1(V)\vec{V}\vec{W}^{*\top}\right\} C(\pi_0)^{-1}(h(\Pi_0)(Z,W) - E(h(\Pi_0)(Z,W) \mid W)) \times \\ &\quad (Y - f_{\alpha_0}(Z,W) - \theta_0(W)). \end{aligned}$$

We state this result in the following lemma and also state a robustness result for this efficient influence curve.

LEMMA 5: *Let $h_1(V) = \sum_a h(a,V)a^2\vec{V}$. The efficient influence curve of $\Psi : \mathcal{M}(\Pi_0) \to \mathbb{R}^d$ is given by*

$$\begin{aligned} &D^*_{\psi,\Pi_0}(P_0) = c_0^{-1} h_1(V) \tfrac{d}{d\beta_0} m_{\beta_0}(V)(m_{\alpha_0}(W) - m_{\beta_0}(V)) \\ &+ c_0^{-1}\left\{E_0 h_1(V)\vec{V}\vec{W}^{*\top}\right\} C(\pi_0)^{-1}(h(\Pi_0)(Z,W) - E(h(\Pi_0)(Z,W) \mid W)) \times \\ &(Y - f_{\alpha_0}(Z,W) - \theta_0(W)). \end{aligned}$$

*We have that*

$$P_0 D^*_{\psi,\Pi_0}(d, m_{\alpha_0}, \theta) = 0, \text{ if either } d = d_0 \text{ or } \theta = \theta_0.$$

*Efficient influence curve in model in which $\Pi_0$ is unknown*

We will now derive the efficient influence curve in model $\mathcal{M}$ in which $\Pi_0$ is unknown, which is obtained by adding a correction term $D_\pi(P_0)$ to the above derived $D^*_{\psi,\Pi_0}(P_0)$. The correction term $D_\pi(P_0)$ that needs to be added to $D^*_{\psi,\Pi_0}$ is the influence curve of $P_0\{D^*_{\psi,\Pi_0}(\pi_n) - D^*_{\psi,\Pi_0}(\pi_0)\}$, where $D^*_{\psi,\Pi_0}(\pi) = D^*_{\psi,\Pi_0}(\beta_0, \theta_0, \alpha_0, d_0, \pi)$ is the efficient influence curve in model $\mathcal{M}(\pi_0)$, as derived above with $\pi_0$ replaced by $\pi$, and $\pi_n$ is the nonparametric NPMLE of $\pi_0$. Let $h_1(V) \equiv \sum_a h(a,v)a^2 \frac{d}{d\beta_0} m_{\beta_0}(v)$. Let $\pi(\epsilon) = \pi + \epsilon\eta$. We plug in for $\eta$ the influence curve

of the NPMLE $\Pi_n(z,w)$, which is given by

$$\eta(z,w) = \frac{I(Z=z,W=w)}{P_0(z,w)}(A - \Pi(Z,W)).$$

We have

$$\begin{aligned} D_\pi(P_0) &= \frac{d}{d\epsilon} P_0 D^*_\psi(\pi(\epsilon))\Big|_{\epsilon=0} \\ &= -\left\{P_0 c_0^{-1} h_1(V)\vec{V}W^{*\top}\right\} C(\pi_0)^{-1} P_0 \left\{W^* W^{*\top}(\pi_0 - E(\pi_0 \mid W))\eta(Z,W)\right\}. \end{aligned}$$

This yields as correction term:

$$\begin{aligned} &D_\pi(P_0)(O) = -(A - \Pi_0(Z,W)) \\ &\left\{P_0 c_0^{-1} h_1(V)\vec{V}W^{*\top}\right\} C(\pi_0)^{-1} \left\{W^* W^{*\top}(\pi_0(Z,W) - E(\pi_0 \mid W))\right\}. \end{aligned}$$

This proves the following lemma.

LEMMA 6: *The efficient influence curve of $\Psi : \mathcal{M} \to \mathbb{R}^d$ is given by*

$$\begin{aligned} &D^*(P_0) = D^*_W(P_0) \\ &+c_0^{-1}\left\{E_0 h_1(V)\vec{V}\vec{W}^{*\top}\right\} C(\pi_0)^{-1} W^*(\Pi_0 - E(\Pi_0(Z,W) \mid W))(Y - f_{\alpha_0}(Z,W) - \theta_0(W)) \\ &-\left\{P_0 c_0^{-1} h_1(V)\vec{V}W^{*\top}\right\} C(\pi_0)^{-1} \left\{W^* W^{*\top}(\pi_0(Z,W) - E(\pi_0 \mid W))\right\}(A - \Pi_0(Z,W)) \\ &\equiv D^*_W(P_0) + C_Y(d_0,\pi_0)(Z,W)(Y - \pi_0(Z,W)m_{\alpha_0}(W) - \theta_0(W)) \\ &\qquad\qquad -C_A(d_0,\pi_0,m_0)(A - \pi_0(Z,W)) \\ &\equiv D^*_W(P_0) + D^*_Y(P_0) - D^*_A(P_0), \end{aligned}$$

*where*

$$\begin{aligned} C_Y(d_0,\pi_0)(Z,W) &= c_0^{-1}\left\{E_0 \textstyle\sum_a h(a,V)a^2\vec{V}\vec{W}^{*\top}\right\} \times \\ &\qquad C(\pi_0)^{-1}(h(\Pi_0)(Z,W) - E(h(\Pi_0)(Z,W) \mid W)) \\ C_A(d_0,\pi_0,m_0)(Z,W) &= \left\{P_0 c_0^{-1} h_1(V)\vec{V}W^{*\top}\right\} C(\pi_0)^{-1}\left\{W^* W^{*\top}(\pi_0(Z,W) - E(\pi_0 \mid W))\right\}. \end{aligned}$$

**Double robustness of efficient influence curve:** We already showed $P_0 D^*(\pi_0, d, \alpha_0, \theta) = 0$ if $d = d_0$ or $\theta = \theta_0$. We also have that $P_0 D^*(\pi, d_0, \alpha_0, \theta) = 0$ for all $\theta$ and $\pi$.

The TMLE is analogue to the TMLE presented for the nonparametric model for $m_0(W)$.

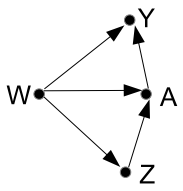


**Figure 1.** Causal diagram

**Table 1**

*Performance of estimators in estimating a* ***scalar causal effect, nonlinear design 1****. The initial estimator for* $E(Y|Z,W)$ *is either consistently specified or misspecified, and all other nuisance parameters are consistenly specified. Sample size is 1000, and 10,000 repetitions were made. The true effect is* 33.23.

CONSISTENTLY SPECIFIED

| | **Estimator** | **Bias** | **Var** | **MSE** |
|---|---|---|---|---|
| *New methods* | Iterative | .0016 | .6103 | .6103 |
| | Linear fluctuation | .0015 | .6189 | .6189 |
| | Logistic fluctuation | .0015 | .6189 | .6189 |
| *Non-parametric* | Estimating equations | −.0016 | .7834 | .7834 |
| | Initial substitution estimator | .0038 | .6990 | .6990 |
| | Confounded | 20.97 | 0.000 | 439.7 |
| | Two-stage least squares | −.3904 | 52.74 | 52.89 |

$E(Y|W,Z)$ IS MISSPECIFIED

| | **Estimator** | **Bias** | **Var** | **MSE** |
|---|---|---|---|---|
| *New methods* | Iterative | .3157 | 117.7 | 117.8 |
| | Linear fluctuation | .6214 | 78.27 | 78.65 |
| | Logistic fluctuation | .8193 | 82.99 | 83.66 |
| *Non-parametric* | Estimating equations | −.2088 | 35.14 | 35.18 |
| | Initial substitution estimator | −.3941 | 54.07 | 54.22 |
| | Confounded | 20.97 | 0.000 | 439.7 |
| | Two-stage least squares | −.3904 | 52.74 | 52.89 |

**Table 2**

*Performance of estimators in estimating* ***vector-valued*** *causal effect, when the* ***treatment effect is nonlinear (design 1)****. Causal parameter $\beta$ to estimate is projection of effect unto linear function of covariates $\{m_\beta(W) = \beta^T W|\beta\}$. The true effect is* $[-64.2, 32.3]$.

| | MEAN ABSOLUTE BIAS OF ESTIMATORS | | | | | |
|---|---|---|---|---|---|---|
| | Consistent specification | | | $E(Y\|Z,W)$ is misspecified | | |
| | n=1000 | n=3000 | n=10000 | n=1000 | n=3000 | n=10000 |
| Iterative | .0683 | .0219 | .0191 | 1.350 | .6043 | .2552 |
| Linear fluctuation | .3025 | .0247 | .0056 | 8.773 | 2.589 | .6587 |
| Estimating equations | .0128 | .0084 | .0119 | 1.521 | 1.013 | .4110 |
| Initial substitution estimator | .6478 | .0595 | .0473 | 136.0 | 136.3 | 136.1 |
| Two-stage least squares | 136.6 | 136.4 | 136.6 | 136.6 | 136.4 | 136.6 |
| Confounded | 10.93 | 10.15 | 10.72 | 10.93 | 10.15 | 10.72 |

| | MEAN ABSOLUTE STD DEV OF ESTIMATORS | | | | | |
|---|---|---|---|---|---|---|
| | Consistent specification | | | $E(Y\|Z,W)$ is misspecified | | |
| | n=1000 | n=3000 | n=10000 | n=1000 | n=3000 | n=10000 |
| Iterative | 11.34 | 3.782 | 1.860 | 34.59 | 12.00 | 4.851 |
| Linear fluctuation | 8.192 | 3.016 | 1.494 | 86.78 | 17.34 | 5.212 |
| Estimating equations | 5.029 | 2.741 | 1.492 | 19.03 | 10.15 | 5.954 |
| Initial substitution estimator | 4.861 | 2.709 | 1.565 | 11.37 | 6.743 | 3.789 |
| Two-stage least squares | 11.12 | 6.235 | 3.694 | 11.12 | 6.235 | 3.694 |
| Confounded | .0021 | .0009 | .0005 | .0021 | .0009 | .0005 |

**Table 3**

*Performance of estimators in estimating a* ***scalar causal effect, nonlinear design 2****, where* $E(Y|W,Z)$ *follows* ***sharp cutoffs****. The initial estimator for* $E(Y|Z,W)$ *is either consistently specified or misspecified, and all other nuisance parameters are consistenly specified. Sample size is 1000, and 10,000 repetitions were made. The true effect is* 1.00*.*

CONSISTENTLY SPECIFIED

| | **Estimator** | **Bias** | **Var** | **MSE** |
|---|---|---|---|---|
| *New methods* | Iterative | −.0853 | .2226 | .2299 |
| | Linear fluctuation | −.0827 | .2198 | .2266 |
| | Logistic fluctuation | .0307 | .1645 | .1654 |
| *Non-parametric* | Estimating equations | −.0643 | .1508 | .1549 |
| | Initial substitution estimator | .0202 | .1196 | .1200 |
| | Confounded | .5735 | .0170 | .3459 |
| | Two-stage least squares | .0926 | .2792 | .2878 |

$E(Y|W,Z)$ IS MISSPECIFIED

| | **Estimator** | **Bias** | **Var** | **MSE** |
|---|---|---|---|---|
| *New methods* | Iterative | −.0703 | .4498 | .4547 |
| | Linear fluctuation | −.0414 | .4561 | .4578 |
| | Logistic fluctuation | .0487 | .3396 | .3420 |
| *Non-parametric* | Estimating equations | −.0636 | .4492 | .4532 |
| | Initial substitution estimator | .0865 | .3870 | .3945 |
| | Confounded | .5735 | .0170 | .3459 |
| | Two-stage least squares | .0926 | .2792 | .2878 |

**Table 4**

*Performance of estimators in estimating* ***vector-valued*** *causal effect, when the* ***treatment effect is linear****. The true effect is* $[0, 1, 2, 3]$*. Causal parameter* $\beta$ *to estimate is coefficient* $\alpha$*, where the treatment effect is* $m(W) = \alpha^T W$*.*

| | MEAN ABSOLUTE BIAS OF ESTIMATORS | | | | | |
|---|---|---|---|---|---|---|
| | Consistent specification | | | $E(Y\|Z,W)$ is misspecified | | |
| | n=1000 | n=3000 | n=10000 | n=1000 | n=3000 | n=10000 |
| Iterative | .0705 | .0417 | .0071 | .1540 | .1302 | .0980 |
| Linear fluctuation | .0049 | .0034 | .0015 | .1101 | .1134 | .0861 |
| Estimating equations | .0062 | .0027 | .0020 | .4194 | .3803 | .2374 |
| Initial substitution estimator | .0090 | .0117 | .0029 | 1.546 | 1.499 | 1.503 |
| Two-stage least squares | .2446 | .2324 | .2443 | .2446 | .2324 | .2443 |
| Confounded | .3484 | .3432 | .3430 | .3484 | .3432 | .3430 |

| | MEAN ABSOLUTE STD DEV OF ESTIMATORS | | | | | |
|---|---|---|---|---|---|---|
| | Consistent specification | | | $E(Y\|Z,W)$ is misspecified | | |
| | n=1000 | n=3000 | n=10000 | n=1000 | n=3000 | n=10000 |
| Iterative | 1.044 | .5967 | .1927 | 1.038 | .6305 | .3421 |
| Linear fluctuation | .5746 | .3067 | .1799 | .5528 | .3410 | .2268 |
| Estimating equations | .6356 | .3944 | .1618 | .5371 | .3549 | .2989 |
| Initial substitution estimator | .5413 | .3140 | .1713 | .4296 | .2514 | .1345 |
| Two-stage least squares | .5104 | .2906 | .1580 | .5104 | .2906 | .1580 |
| Confounded | .1188 | .0657 | .0359 | .1188 | .0657 | .0359 |

**Table 5**

*Mean coverage of **95% confidence intervals**. The coverage is calculated for each dimension of the parameter of interest and the average taken. For the top three estimators in each table, the empirical variance of the efficient influence curve $Var(D^*(Q_n^*, g_n^*))$ is used to calculate the standard error. For the other estimators, we give the unfair advantage of using the accurate variance in calculating the confidence intervals*

| | LINEAR TREATMENT EFFECT | | | | | |
|---|---|---|---|---|---|---|
| | Consistent specification | | | $E(Y\|Z,W)$ is misspecified | | |
| | n=1000 | n=3000 | n=10000 | n=1000 | n=3000 | n=10000 |
| Iterative | 96.8 | 96.4 | 96.2 | 96.1 | 96.0 | 95.2 |
| Linear fluctuation | 96.6 | 96.1 | 96.3 | 95.9 | 95.1 | 94.7 |
| Estimating equations | 96.4 | 96.0 | 96.3 | 89.4 | 88.8 | 90.3 |
| Initial substitution estimator | 94.6 | 94.8 | 94.8 | 5.98 | 0 | 0 |
| Two-stage least squares | 92.6 | 87.5 | 67.7 | 92.6 | 87.5 | 67.7 |
| Confounded | 19.8 | 0.22 | 0 | 19.8 | 0.22 | 0 |

| | NONLINEAR TREATMENT EFFECT | | | | | |
|---|---|---|---|---|---|---|
| | Consistent specification | | | $E(Y\|Z,W)$ is misspecified | | |
| | n=1000 | n=3000 | n=10000 | n=1000 | n=3000 | n=10000 |
| Iterative | 97.5 | 97.1 | 96.7 | 97.3 | 97.2 | 96.9 |
| Linear fluctuation | 97.2 | 96.5 | 96.8 | 96.9 | 95.9 | 96.6 |
| Estimating equations | 96.4 | 95.7 | 96.2 | 96.8 | 96.3 | 97.0 |
| Initial substitution estimator | 94.2 | 94.6 | 94.3 | 0 | 0 | 0 |
| Two-stage least squares | 0 | 0 | 0 | 0 | 0 | 0 |
| Confounded | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 6**
*Means and SDs of variables.*

| | Mothers | Fathers |
|---|---|---|
| Sample size | 4,101,825 | 4,001,970 |
| Program intensity (R) | 0.22 | 0.22 |
| | (0.11) | (0.11) |
| Parent's years of schooling | 9.93 | 10.67 |
| | (1.46) | (1.15) |
| Percentage of low-birthweight births | 4.50 | 4.80 |
| | (1.24) | (1.25) |
| Neonatal mortality (deaths per thousand births) | 2.32 | 2.33 |
| | (2.38) | (2.38) |
| Postneonatal mortality (deaths per thousand neonatal survivors) | 3.50 | 3.38 |
| | (2.56) | (2.71) |
| Infant mortality (deaths per thousand births) | 5.81 | 5.71 |
| | (3.58) | (3.67) |

Note: the SDs for the binary outcomes (low birth weight, and mortality) are the SD's for the average rates within each cell (in which county, and parent and child's birth cohorts are fixed). Each cell is weighted by its sample size for the relevant outcome (for example, the total number of births in a cell for infant mortality) in finding the SD.

**Table 7**

*Estimates of effect of parents' education on infant's health, where the latter is given by the log-odds ratio of a binary outcome. Significant effects are marked with (∗) and highly significant effects with (∗∗).*

| | Mother's education | | | Father's education | | |
|---|---|---|---|---|---|---|
| | Low birthweight | Neonatal mortality | Postneonatal mortality | Low birthweight | Infant mortality | Postneonatal mortality |
| **TMLE, linear fluctuation** | | | | | | |
| Mean effect | -.266 (.163)* | -1.04 (.193)** | -.358 (.255) | -.126 (.179) | -.632 (.105)** | -.569 (.242)** |
| First stage CV-$R^2$ | .814 | .805 | .798 | .801 | .784 | .751 |
| Second stage CV-$R^2$ | .539 | .561 | .453 | .616 | .622 | .626 |
| Change in outcome | -.041% | -.387 | -.222 | -.001 | -.342 | -.166 |
| Percent change in outcome | -7.07% | -16.7% | -6.54% | -1.54% | -5.99% | -4.75% |
| **Two-stage least squares** | | | | | | |
| Mean effect | -.212 (.175) | -.265 (.124)* | -.529 (.229)* | -.298 (.160)* | -.480 (.182)** | -.602 (.254)** |
| First stage CV-$R^2$ | .778 | .747 | .760 | .749 | .752 | .716 |
| Second stage CV-$R^2$ | .426 | .531 | .376 | .318 | .378 | .395 |
| Change in outcome | -.027% | -.078 | -.255 | -.001 | -.161 | -.178 |
| Percent change in outcome | -4.58% | -3.42% | -7.14% | -2.03% | -2.70% | -5.29% |
| F-test for weak IV | 15.5 | 12.1 | 13.7 | 11.1 | 9.22 | 9.69 |
| Hausman-Wu test | 1.64 | .738 | .684 | .892 | .186 | .432 |
| **OLS** | | | | | | |
| Mean effect | -.177 (.009) | -.381 (.011) | -.434 (.010) | -.223 (.006) | -.345 (.010) | -.419 (.011) |
| $R^2$ | .428 | .538 | .424 | .345 | .400 | .415 |

**Table 8**

*Additive treatment effect modifiers for the effect of* ***mother's*** *education on infant's health. Significant effects are marked with* $(*)$ *and highly significant effects with* $(**)$.

| | Low birthweight | Neonatal mortality | Postneonatal mortality |
|---|---|---|---|
| Strongest additive treatment effect modifiers | $YOB1982$: -1.24 (.472)**<br>$YOB1985$: -.608 (.162)**<br>$YOB1988$: -.506 (.167)**<br>$YOB1989$: -.497 (.168)**<br>$YOB1987$: -.494 (.166)**<br>$YOB1998$: +.494 (.188)** | $YOB1982$: -3.10 (.933)**<br>$YOB1979$: -2.64 (1.31)*<br>$YOB1980$: -2.33 (.955)*<br>$YOB1985$: -2.25 (.346)**<br>$YOB1988$: -1.84 (.349)**<br>$YOB1986$: -1.62 (.346)** | $YOB1980$ +1.22 (.430)**<br>$YOB1982$ -.916 (.317)**<br>$YOB1989$ -.886 (.227)**<br>$YOB1998$ +.795 (.233)**<br>$YOB1999$ +.667 (.342)<br>$YOB1987$ -.664 (.202)** |
| Summary statistics | Q1 Q2 Mean Q3<br>-.237 -.072 -.148 -.010 | Q1 Q2 Mean Q3<br>-1.06 -.587 -.738 -.009 | Q1 Q2 Mean Q3<br>-.358 -.104 -.111 .047 |
| Summary statisics of z-scores | Q1 Q2 Mean Q3<br>-1.78 -.511 -.833 -.056 | Q1 Q2 Mean Q3<br>-2.82 -1.64 -1.64 -.106 | Q1 Q2 Mean Q3<br>-1.75 -.721 -.628 .372 |

**Table 9**

*Additive treatment effect modifiers for the effect of **father's** education on infant's health. Significant effects are marked with (∗) and highly significant effects with (∗∗).*

| | Low birthweight | Infant mortality | Postneonatal mortality |
|---|---|---|---|
| Strongest additive treatment effect modifiers | $YOB$1980: -3.98 (8.93)<br>$YOB$1979: +3.54 (1.73)*<br>$YOB$1978: -2.39 (.962)*<br>$YOB$1981: -1.43 (.437)**<br>$YOB$1985: -1.13 (.239)**<br>$YOB$1982: -.880 (.385)* | $YOB$1980: -4.24 (.419)**<br>$YOB$1979: -3.77 (.835)**<br>$YOB$1978: -1.52 (.614)*<br>$COH$1968 -.849 (.464)<br>$COH$1967 -.791 (.464)<br>$COH$1965 -.704 (.466) | $YOB$1979 -5.103 (.297)**<br>$YOB$1978 -3.27 (1.14)**<br>$YOB$1981 -1.34 (.241)**<br>$COH$1968: -1.29 (.427)**<br>$COH$1967: -1.26 (.429)**<br>$YOB$1983: -1.11 (.778) |
| Summary statistics | Q1 Q2 Mean Q3<br>-.792 -.351 -.504 -.307 | Q1 Q2 Mean Q3<br>-.561 -.356 -.442 .119 | Q1 Q2 Mean Q3<br>-.931 -.419 -.636 -.055 |
| Summary statisics of z-scores | Q1 Q2 Mean Q3<br>-3.16 -.634 -1.59 -.276 | Q1 Q2 Mean Q3<br>-1.22 -.617 -.129 1.30 | Q1 Q2 Mean Q3<br>-2.17 -1.39 -1.19 -.198 |