
DOUBLE ROBUST VARIANCE ESTIMATION

Bonnie E. Shook-Sa^{1,*}, Paul N. Zivich², Chanhwa Lee¹, Keyi Xue¹,
Rachael K. Ross³, Jessie K. Edwards², Jeffrey S. A. Stringer⁴, and Stephen R. Cole²

¹Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

²Department of Epidemiology, University of North Carolina at Chapel Hill, U.S.A.

³Department of Epidemiology, Mailman School of Public Health, Columbia University, New York City, New York, U.S.A.

⁴Department of Obstetrics and Gynecology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

*bshooksa@email.unc.edu

April 26, 2024

ABSTRACT

Doubly robust estimators have gained popularity in the field of causal inference due to their ability to provide consistent point estimates when either an outcome or exposure model is correctly specified. However, the influence function based variance estimator frequently used with doubly robust estimators is only consistent when both the outcome and exposure models are correctly specified. Here, use of M-estimation and the empirical sandwich variance estimator for doubly robust point *and* variance estimation is demonstrated. Simulation studies illustrate the properties of the influence function based and empirical sandwich variance estimators. Estimators are applied to data from the Improving Pregnancy Outcomes with Progesterone (IPOP) trial to estimate the effect of maternal anemia on birth weight among women with HIV. In the example, birth weights if all women had anemia were estimated to be lower than birth weights if no women had anemia, though estimates were imprecise. Variance estimates were more stable under varying model specifications for the empirical sandwich variance estimator than the influence function based variance estimator.

Keywords causal inference; double robustness; empirical sandwich variance; augmented inverse probability weighting; M-estimation.

1 Introduction

Two common approaches for estimating the causal effect of a point exposure or treatment with observational studies are methods based on inverse probability of treatment weighting (IPTW) and the g-formula. IPTW estimators (Robins, 1998; Robins et al., 2000; Hernán et al., 2000) rely on estimation of the probability of exposure given a set of covariates, i.e., the propensity score. Alternatively, the g-formula (Robins, 1986) is a direct standardization approach where the outcome is modeled as a function of the exposure and covariates. When estimated nonparametrically, IPTW and g-formula estimators are equivalent (Hernán and Robins, 2020). However, in high dimensional settings (e.g., settings with continuous covariates), analysts typically resort to parametric models for the exposure and outcome, and the IPTW and g-formula estimators are no longer equivalent due to differing parametric constraints. Both approaches provide consistent estimators of the average causal effect (ACE) if the corresponding model is correctly specified. Because the functional forms of these “nuisance” models are typically unknown, correct specification remains an obstacle to valid effect estimation.

Augmented inverse probability weighted (AIPW) estimators and targeted maximum likelihood estimators (TMLE) combine estimates from the exposure and outcome models in such a way that resulting estimators are “doubly robust.” That is, they are consistent when either the exposure or outcome model is correctly specified, but not necessarily both (Robins et al., 2007; Bang and Robins, 2005; Funk et al., 2011; Kang and Schafer, 2007; Hernán and Robins, 2020; van der Laan and Rubin, 2006). In other words, doubly robust methods give the analyst two opportunities for correct model specification. Two common AIPW estimators are an estimator that incorporates estimated propensity scores

from the exposure model and pseudo-outcomes from the outcome model (i.e., the classic AIPW) (Hernán and Robins, 2020; Lunceford and Davidian, 2004) and a g-formula type estimator where the outcome model is weighted by the IPTW (i.e., the weighted regression AIPW) (Kang and Schafer, 2007; Robins et al., 2007; Vansteelandt and Keiding, 2011; Gabriel et al., 2023). TMLE is a variation of the weighted regression AIPW estimator that instead combines IPTW and the outcome model through a “targeting” model fit using the outcome predictions and weights (van der Laan and Rubin, 2006). Both the weighted regression AIPW estimator and TMLE respect the parameter space of the estimand, while the classic AIPW can produce estimates outside of the range of the parameter space (Gabriel et al., 2023; Gruber and Van Der Laan, 2012).

In addition to the double robustness property, doubly robust estimators also offer precision advantages over IPTW estimators. When both nuisance models are correctly specified, the three doubly robust estimators previously discussed are semiparametric efficient, having the smallest asymptotic variance of all IPTW estimators (Robins et al., 1994; van der Laan and Rubin, 2006; Gabriel et al., 2023). This means that even in settings where the analyst is confident in the specification of one of the models (e.g., the exposure mechanism is well known), doubly robust estimators can offer improved efficiency relative to IPTW estimators, provided that the outcome model is not grossly misspecified. Given these potential precision advantages and that the true specification of both models is typically unknown, doubly robust approaches have gained popularity (Sullivan et al., 2022; Josey et al., 2023; Lee et al., 2019; Zhong et al., 2018; Gerds et al., 2022; Chua et al., 2022; Hoffman et al., 2022; Yazdi et al., 2021).

Despite adoption of doubly robust estimators, there is a major limitation with the corresponding influence function based variance estimator commonly used in applications and provided by software. Namely, this variance estimator is only consistent if both the propensity and outcome models are correctly specified (Daniel, 2014). This limitation has been highlighted previously (Funk et al., 2011; Daniel, 2014), yet the influence function based variance estimator is still commonly used. A systematic review published in October 2023 found that 67% of papers published using TMLE did not indicate how variances were estimated. Of the 33% that did report the method, the majority (73%) relied on the influence function based variance estimator (Smith et al., 2023). Widespread application of the influence function based variance estimator is aided by available software, as it is used for confidence interval (CI) construction in the R AIPW (Zhong et al., 2021) and `tmle` (Gruber and Van Der Laan, 2012) packages as well as the `zEpid` (Zivich et al., 2022a) Python package.

The influence function based variance estimator’s reliance on correctly specifying both exposure and outcome models is particularly concerning in instances where researchers understand either the exposure or outcome mechanism well, but not both. If the outcome process is well known but the exposure process is not well understood, leading to correct specification of the outcome model but misspecification of the propensity model, the influence function based variance estimator is inconsistent (Laan and Robins, 2003; Muñoz and Van Der Laan, 2012). In this setting, consistent point estimators may be paired with CIs that do not achieve the nominal level of coverage. Alternatively, if the exposure process is well understood but the outcome process is not, leading to correct specification of the propensity model but misspecification of the outcome model, the influence function based variance estimator is conservative (Muñoz and Van Der Laan, 2012). Precision advantages of doubly robust estimators may be lost with application of a conservative variance estimator. For example, we aim to estimate the effect of maternal anemia on birth weight among women with HIV. While most causes of anemia among women with HIV are well understood, doubly robust estimators may be preferred over IPTW estimators due to efficiency advantages. Modeling the relationship between anemia and birth weight among women with HIV is challenging due to complex dependencies between anemia and other drivers of birth weight. Furthermore, low birth weight may result from two separate pathways - preterm birth or fetal growth restrictions - each of which is a complex syndrome with several causes. For these reasons, the outcome model may be susceptible to misspecification, thus application of the influence function based variance estimator may lead to conservative variance estimates in this study.

Alternative, but less commonly used, variance estimation approaches are themselves doubly robust, providing consistent estimates of the variance when only one of the two nuisance models is correctly specified. Two such approaches are the nonparametric bootstrap and the empirical sandwich variance estimator based on M-estimation theory. The nonparametric bootstrap, while doubly robust (Funk et al., 2011), is computationally burdensome. For large datasets, this computational burden may preclude its use. M-estimation (Stefanski and Boos, 2002) has been proposed for use with doubly robust estimators (Lunceford and Davidian, 2004; Daniel, 2014), but we are not aware of prior work highlighting the doubly robust property of the empirical sandwich variance estimator. Furthermore, the recent systematic review (Smith et al., 2023) did not mention application of the empirical sandwich variance estimator in any of the papers evaluated. This suggests that the doubly robust property of the empirical sandwich variance estimator is not well known.

This paper demonstrates how M-estimation can be applied for AIPW estimators and TMLE to obtain doubly robust point *and* variance estimators. In Section 2, three common doubly robust estimators are described along with two

approaches for inference: the influence function based variance estimator and M-estimation paired with the empirical sandwich variance estimator. Estimators are applied to data from the Improving Pregnancy Outcomes with Progesterone (IPOP) trial to estimate the average causal effect of anemia on birth weight in Section 3. The doubly robust property of the empirical sandwich variance estimator is demonstrated in a simulation study and is contrasted with the performance of the influence function based variance estimator in Section 4. The benefits and limitations of each variance estimation approach are discussed in Section 5. R code for implementing each of the doubly robust estimators is provided on GitHub.

2 Methods

2.1 Preliminaries

Consider an observational study that seeks to estimate the effect of a binary exposure (or treatment) X on an outcome Y . Assume n independent and identically distributed copies of $O_i = (X_i, Y_i, Z_i)$ are observed, where Z are a set of baseline covariates. Let the potential outcomes for individual i under exposure and no exposure be denoted by Y_i^1 and Y_i^0 such that by causal consistency $Y_i = X_i Y_i^1 + (1 - X_i) Y_i^0$. Assume the set of covariates Z satisfy $Y^x \perp\!\!\!\perp X \mid Z$ for $x \in \{0, 1\}$, i.e., conditional exchangeability. Also assume positivity holds such that $\Pr(X = x \mid Z = z) > 0$ for all z such that $dF_Z(z) > 0$ and $x \in \{0, 1\}$, where F_Z is the cumulative distribution function of Z . The estimand is the average causal effect (ACE), $\mu^1 - \mu^0$ where $\mu^x = E(Y^x)$ for $x \in \{0, 1\}$. Unless otherwise specified, vectors are assumed to be column vectors.

2.2 Estimators

Three doubly robust approaches for estimating the ACE are considered: the classic AIPW estimator, the weighted regression AIPW estimator, and TMLE. All three methods combine estimates from a model for the probability of exposure conditional on covariates (propensity model) and a model for the outcome conditional on exposure and covariates (outcome model). The approach for modeling the propensity score is common across the three methods, while there are slight differences in fitting the outcome model. For all three approaches, the propensity score for each participant, $e_i = \Pr(X_i = 1 \mid Z_i)$, is estimated. In practice, often the model $\text{logit}(e_i) = G_i^T \alpha$ is fit, where $G_i = g(Z_i)$ is a vector of predictors for participant i for some user-specified function g of Z_i and α is the vector of regression coefficients from the propensity model. The predicted propensity score for each participant is calculated as $\hat{e}_i = e(Z_i, \hat{\alpha}) = \text{logit}^{-1}(G_i^T \hat{\alpha})$ where $\hat{\alpha}$ is the maximum likelihood estimate (MLE) of α . The IPTW is then estimated for each participant as $\hat{W}_i = X_i \hat{e}_i^{-1} + (1 - X_i)(1 - \hat{e}_i)^{-1}$.

2.2.1 Classic AIPW estimator

For the classic AIPW estimator, the $Y \mid Z, X$ model is fit and used to predict pseudo-outcomes for each participant under both exposures. That is, $a_x(Z_i, \hat{\gamma}) = \hat{E}(Y_i \mid Z_i, X_i = x)$ for $x \in \{0, 1\}$, where $\hat{\gamma}$ are the MLEs of the parameters from the assumed outcome model. Then, the classic AIPW estimator for the ACE is

$$\widehat{DR}_C = \hat{\mu}_C^1 - \hat{\mu}_C^0 \quad (1)$$

where $\hat{\mu}_C^1 = n^{-1} \sum_{i=1}^n \hat{e}_i^{-1} \{X_i Y_i - (X_i - \hat{e}_i) a_1(Z_i, \hat{\gamma})\}$ and $\hat{\mu}_C^0 = n^{-1} \sum_{i=1}^n (1 - \hat{e}_i)^{-1} \{(1 - X_i) Y_i + (X_i - \hat{e}_i) a_0(Z_i, \hat{\gamma})\}$. The estimator (1) was originally proposed by Robins et al. (1994) and was further examined in simulation studies, such as Lunceford and Davidian (2004), Kang and Schafer (2007), and Funk et al. (2011).

2.2.2 Weighted regression AIPW estimator

An alternative AIPW estimator incorporates the estimated propensity scores when fitting the outcome regression model. The parameters of the $Y \mid Z, X$ model are estimated with IPTW-weighted maximum likelihood estimation. Pseudo-outcomes are obtained for each participant under both exposures, i.e., $b_x(Z_i, \hat{\beta}) = \hat{E}(Y_i \mid Z_i, X_i = x)$ for $x \in \{0, 1\}$, where $\hat{\beta}$ are the MLEs from the weighted outcome model. The weighted regression AIPW estimator is

$$\widehat{DR}_{WR} = \hat{\mu}_{WR}^1 - \hat{\mu}_{WR}^0 \quad (2)$$

where $\hat{\mu}_{WR}^x = n^{-1} \sum_{i=1}^n b_x(Z_i, \hat{\beta})$ for $x \in \{0, 1\}$. The weighted regression AIPW estimator (2) has been evaluated previously (Kang and Schafer, 2007; Robins et al., 2007; Vansteelandt and Keiding, 2011; Gabriel et al., 2023) and is expected to be more stable than (1) when IPTWs are extreme due to bounding within the parameter space (Kang and Schafer, 2007; Robins et al., 2007; Vansteelandt et al., 2012).

2.2.3 TMLE

TMLE is an alternative doubly robust estimator that consists of a two-stage process. In the first stage, $Y \mid Z, X$ is modeled as in Section 2.2.1, and $a_x(Z_i, \hat{\gamma})$ are calculated for $x \in \{0, 1\}$. In the second stage, referred to as the targeting stage (van der Laan and Rubin, 2006), corrections to outcome regression predictions from stage one are made that incorporate the estimated propensity scores. Different methods have been proposed for fitting the targeting model; here, the weighted regression approach is considered, as it is thought to handle random positivity violations better (van der Laan et al., 2011). Specifically, the models $\text{logit}(Y_i) = \eta_x + \text{logit}\{a_x(Z_i, \hat{\gamma})\}$ are fit for $x \in \{0, 1\}$. Parameters η_x are estimated using weighted maximum-likelihood, with weights $(1 - X)\hat{W}_i$ and $X\hat{W}_i$ for the $x = 0$ and $x = 1$ targeting model, respectively. In the targeting models, logit-transformed pseudo-outcomes ($\text{logit}\{a_x(Z_i, \hat{\gamma})\}$) for each participant are included as offset terms, and MLEs $\hat{\eta}_0$ and $\hat{\eta}_1$ can be thought of as corrections to these pseudo-outcomes (Zivich et al., 2022b). The stage two pseudo-outcomes are computed as $c_x(O_i, \hat{\gamma}, \hat{\eta}_x) = \text{expit}[\text{logit}\{a_x(Z_i, \hat{\gamma})\} + \hat{\eta}_x]$. The TMLE estimator of ACE is

$$\widehat{DR}_{TMLE} = \hat{\mu}_{TMLE}^1 - \hat{\mu}_{TMLE}^0 \quad (3)$$

where $\hat{\mu}_{TMLE}^x = n^{-1} \sum_{i=1}^n c_x(O_i, \hat{\gamma}, \hat{\eta}_x)$ for $x \in \{0, 1\}$. Note that for continuous outcomes, such as birth weight considered in Section 3, the outcome is scaled prior to implementing TMLE, i.e., $Y_i^* = (Y_i - a)/(b - a)$ is defined for each participant, where (a, b) are the bounds of Y . Then, Y_i is replaced with Y_i^* in the steps above, and $c_x(O_i, \hat{\gamma}, \hat{\eta}_x) = \text{expit}[\text{logit}\{a_x(Z_i, \hat{\gamma})\} + \hat{\eta}_x](b - a) + a$ re-scales the outcome back to the original bounds of Y .

TMLE methods were originally proposed by van der Laan and Rubin (2006) and have been extended and applied in a number of settings (Stitelman et al., 2012; Balzer et al., 2023; van der Laan et al., 2018; van der Laan and Gruber, 2012). When parametric models are used in both stages of estimation, (3) is expected to be similar to (2) (Tran et al., 2019).

2.3 Variance Estimation

Variance estimation for doubly robust estimators (1), (2), and (3) is not straightforward, because the variance of each estimator depends on estimating the parameters of the nuisance models. Here, we consider two methods for variance estimation: the commonly used influence function based variance estimator and the empirical sandwich variance estimator. Both variance estimators can be used to construct Wald-typed CIs for the ACE .

2.3.1 Influence function based variance estimator

Estimators (1), (2), and (3) are suggested by the efficient influence function (Luque-Fernandez et al., 2018; Gabriel et al., 2023). The commonly used influence function based variance estimator has the form (Glynn and Quinn, 2010; Gruber and Van Der Laan, 2012)

$$\hat{V}(\widehat{DR}) = n^{-2} \sum_{i=1}^n \hat{I}_i^2 \quad (4)$$

where $\hat{I}_i = \hat{e}_i^{-1} X_i Y_i - (1 - \hat{e}_i)^{-1} (1 - X_i) Y_i - \{\hat{e}_i^{-1} (1 - \hat{e}_i)^{-1} (X_i - \hat{e}_i)\} \{ (1 - \hat{e}_i) \hat{Y}_i^1 + \hat{e}_i \hat{Y}_i^0 \} - \widehat{DR}$. Here, pseudo-outcomes \hat{Y}_i^x equal $a_x(Z_i, \hat{\gamma})$, $b_x(Z_i, \hat{\beta})$, and $c_x(O_i, \hat{\gamma}, \hat{\eta}_x)$ for the classic AIPW, weighted regression AIPW, and TMLE methods, respectively, and \widehat{DR} equals (1), (2), and (3) for the three corresponding estimators.

The estimators (4) are consistent for the variance of (1), (2), and (3) when both propensity score and outcome models are correctly specified (Daniel, 2014; Gruber and Van Der Laan, 2012). When one of the two nuisance models is misspecified, the influence function based variance estimator is not consistent. As noted in the introduction, the estimator (4) is conservative when the weight model is correctly specified but the outcome model is misspecified (Muñoz and Van Der Laan, 2012), and may be conservative or anti-conservative when the outcome model is correctly specified but the weight model is misspecified (Laan and Robins, 2003; Muñoz and Van Der Laan, 2012). In empirical studies, the variance estimator (4) for the classic AIPW estimator (1) was found to be unstable for small sample sizes ($n < 1000$), even under correct model specification (Funk et al., 2011). Similar instability has been reported for (4) with TMLE ($n = 100$) (van der Laan and Gruber, 2011).

2.3.2 Empirical sandwich variance estimator

An alternative to the influence function based variance estimator (4) is the empirical sandwich variance estimator. The estimators (1), (2), and (3) can each be expressed as the solution to a set of unbiased estimating equations. The M-estimator for each set of estimating equations, $\hat{\theta}$, is the solution (for θ) to $\sum_{i=1}^n \psi_q(O_i, \theta) = 0$, where $\psi_q(O_i, \theta)$ are the estimating functions corresponding to each of the three estimators $q \in \{1, 2, 3\}$ with parameter vector θ . Parameter vectors and estimating functions for (1), (2), and (3) can be specified as follows.

Classic AIPW estimator

For the classic AIPW estimator, $\theta = [\alpha, \gamma, \mu_C^1, \mu_C^0, DR_C]$. The estimating function ψ_1 is specified as

$$\psi_1(O_i; \theta) = \begin{bmatrix} \psi_\alpha \\ \psi_\gamma \\ \psi_{\mu_{C1}} \\ \psi_{\mu_{C0}} \\ \psi_C \end{bmatrix} = \begin{bmatrix} \{X_i - \text{expit}(G_i^T \alpha)\} G_i \\ \{Y_i - \phi^{-1}(H_i^T \gamma)\} H_i \\ \{X_i Y_i - (X_i - e_i) a_1(Z_i, \gamma)\} e_i^{-1} - \mu_C^1 \\ \{(1 - X_i) Y_i + (X_i - e_i) a_0(Z_i, \gamma)\} (1 - e_i)^{-1} - \mu_C^0 \\ \mu_C^1 - \mu_C^0 - DR_C \end{bmatrix}$$

where ψ_α is the vector of score functions from the propensity score model, ψ_γ is the vector of score functions for the outcome model (with link function ϕ), and $H_i = h(X_i, Z_i)$ is a vector of predictors for the outcome model for participant i for some user-specified function h of X_i and Z_i . The estimating functions $\psi_{\mu_{C1}}$ and $\psi_{\mu_{C0}}$ are for the causal means μ_C^1 and μ_C^0 , respectively, and ψ_C is a delta method transformation of the causal means for estimation of the ACE using the classic AIPW, i.e., \widehat{DR}_C . Shook-Sa et al. (2023) demonstrate that this set of estimating equations is unbiased when at least one of the models is correctly specified.

Weighted regression AIPW estimator

For the weighted regression AIPW estimator, $\theta = [\alpha, \beta, \mu_{WR}^1, \mu_{WR}^0, DR_{WR}]$. The estimating function ψ_2 is specified as

$$\psi_2(O_i; \theta) = \begin{bmatrix} \psi_\alpha \\ \psi_\beta \\ \psi_{\mu_{WR1}} \\ \psi_{\mu_{WR0}} \\ \psi_{WR} \end{bmatrix} = \begin{bmatrix} \{X_i - \text{expit}(G_i^T \alpha)\} G_i \\ W_i \{Y_i - \phi^{-1}(H_i^T \beta)\} H_i \\ b_1(Z_i, \beta) - \mu_{WR}^1 \\ b_0(Z_i, \beta) - \mu_{WR}^0 \\ \mu_{WR}^1 - \mu_{WR}^0 - DR_{WR} \end{bmatrix}$$

where ψ_β is the vector of IPTW-weighted score functions for the outcome model. The estimating functions $\psi_{\mu_{WR1}}$ and $\psi_{\mu_{WR0}}$ are for the causal means μ_{WR}^1 and μ_{WR}^0 , respectively, and ψ_{WR} is a delta method transformation of the causal means for estimation of the ACE using the weighted AIPW estimator \widehat{DR}_{WR} . Gabriel et al. (2023) demonstrate unbiasedness of this set of estimating equations when at least one of the nuisance models is correctly specified.

TMLE

With TMLE, $\theta = [\alpha, \gamma, \eta_1, \eta_0, \mu_{TMLE}^1, \mu_{TMLE}^0, DR_{TMLE}]$. For continuous outcomes, the estimating function ψ_3 is specified as

$$\psi_3(O_i; \theta) = \begin{bmatrix} \psi_\alpha \\ \psi_\gamma \\ \psi_{t1} \\ \psi_{t0} \\ \psi_{\mu_{TMLE1}} \\ \psi_{\mu_{TMLE0}} \\ \psi_{TMLE} \end{bmatrix} = \begin{bmatrix} \{X_i - \text{expit}(G_i^T \alpha)\} G_i \\ \{Y_i^* - \phi^{-1}(H_i^T \gamma)\} H_i \\ X_i e_i^{-1} (Y_i^* - \text{expit}[\eta_1 + \text{logit}\{a_1(Z_i, \gamma)\}]) \\ (1 - X_i) (1 - e_i)^{-1} (Y_i^* - \text{expit}[\eta_0 + \text{logit}\{a_0(Z_i, \gamma)\}]) \\ c_1(O_i, \gamma, \eta_1) - \mu_{TMLE}^1 \\ c_0(O_i, \gamma, \eta_0) - \mu_{TMLE}^0 \\ \mu_{TMLE}^1 - \mu_{TMLE}^0 - DR_{TMLE} \end{bmatrix}$$

where ψ_γ is the vector of score functions for the scaled outcome model. The estimating functions ψ_{t1} and ψ_{t0} are for the two targeting models, while $\psi_{\mu_{TMLE1}}$ and $\psi_{\mu_{TMLE0}}$ are for the causal means μ_{TMLE}^1 and μ_{TMLE}^0 , respectively. Note $c_x(O_i, \gamma, \eta_x) = \text{expit}[\text{logit}\{a_x(Z_i, \gamma)\} + \eta_x](b - a) + a$ are the pseudo-outcomes following the targeting step, as outlined in Section 2.2.3. Finally, ψ_{TMLE} is a delta method transformation of the causal means for estimation of the ACE using \widehat{DR}_{TMLE} . Unbiasedness of $\psi_3(O_i; \theta)$ when at least one nuisance model is correctly specified is demonstrated in the Appendix.

Because (1), (2), and (3) are each solutions to unbiased estimating equation vectors, it follows under suitable regularity conditions (Stefanski and Boos, 2002) that for each estimator, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V(\theta))$. Here, $V(\theta) = A(\theta)^{-1} B(\theta) \{A(\theta)^{-1}\}^T$ where $A(\theta) = E\{-\partial \psi_q(O_i; \theta) / \partial \theta\}$ and $B(\theta) = E\{\psi_q(O_i; \theta) \psi_q(O_i; \theta)^T\}$. The quantities $A(\theta)$ and $B(\theta)$ can be consistently estimated by replacing each expected value with its empirical counterpart, resulting in the empirical sandwich variance estimator $\hat{V}(\hat{\theta})$. The variance of (1), (2), and (3) can be consistently estimated by the bottom right element of its corresponding $\hat{V}(\hat{\theta})$. Importantly, because the estimating equations are unbiased when either the propensity or the outcome model is correctly specified, the empirical sandwich variance estimator is consistent under such model misspecification. This is in contrast to the influence function based variance estimator (4) which requires correct specification of both nuisance models for consistent variance estimation (Daniel, 2014).

3 Example: Improving Pregnancy Outcomes with Progesterone

Prior research has shown links between maternal anemia and low birth weight (Azizah et al., 2022; Figueiredo et al., 2018). The relationship between anemia and birth weight is particularly important to understand for women living with HIV, as they experience anemia more frequently than women without HIV (Levine et al., 2001) and are more prone to adverse birth outcomes (Tukey et al., 2021). The effect of maternal anemia on birth weight among women living with HIV was estimated with data from the IPOP trial.

The IPOP trial was a randomized, double-blind, placebo-controlled trial conducted in Lusaka, Zambia. Participants were pregnant women aged 18 or older living with HIV on (or intending to initiate) antiretroviral therapy (ART) who had viable singleton pregnancies at less than 24 weeks of gestation at baseline. The 800 enrolled participants were randomized with equal probability to either a placebo or 17 alpha-hydroxyprogesterone caproate (17P) treatment to evaluate the efficacy of 17P for prevention of preterm birth. Additional details about the IPOP trial have been published previously (Price et al., 2021). While IPOP was designed to examine the effect of 17P on preterm birth or stillbirth, covariates collected during the trial allow for examination of other causal questions, assuming that the identification assumptions hold.

The three doubly robust estimators described in Section 2.2 were applied to IPOP data for estimation of the *ACE* of maternal anemia on birth weight. The exposure, maternal anemia, was defined as having a baseline hemoglobin level below 10.5 g/dL. The IPOP data were limited to the 782 participants (98%) with no prior preterm births and with measured birth weights and baseline hemoglobin values. Note 14 participants experienced stillbirth. Birth weights for these participants were included in the analysis; we return to this issue in the discussion. It was assumed that potential outcomes are independent of the exposure conditional on the following baseline covariates: trimester (1 or 2), maternal age group (18-24, 25-29, 30-34, 35-39, 40+), number of prior births (0, 1, 2, 3, 4+), alcohol use during pregnancy (yes or no), ART use (yes or no), and maternal height (cm) (Table 1). At baseline, no IPOP participants reported tobacco or drug use during pregnancy, so these covariates were not included. Primary analyses were conducted using R and were independently replicated in Python.

As discussed in the introduction, application of doubly robust estimators for estimating the *ACE* of maternal anemia on birth weight may offer precision advantages over IPTW estimators, but the correct model specifications are unknown. To demonstrate differences in the influence function based variance estimator and the empirical sandwich variance estimator presented in Section 2.3, each of the doubly robust estimators were applied with three sets of model specifications. First, the covariates in Table 1 were included in both propensity and outcome models. Then, a naive outcome model was fit that included only the exposure variable (anemia status), while the propensity model included the full set of covariates. Under this specification, the outcome model was reasonably thought to be misspecified due to its simplicity. Finally, a naive propensity model was fit that included only an intercept term, while the outcome model included the full set of covariates. Due to the simplicity of the propensity model, it was reasonably thought to be misspecified. For all estimators, maternal height was modeled using restricted cubic splines with four knots placed at the 5th, 35th, 65th, and 95th percentiles (Perperoglou et al., 2019). The empirical sandwich variance estimator was computed using `geex` in R (Saul and Hudgens, 2020) and `delicatessen` in Python (Zivich et al., 2022c). Confidence intervals for the *ACE* were constructed based on the influence function based variance estimator and the empirical sandwich variance estimator. Estimates for each of the three methods and model specifications along with corresponding 95% CIs are presented in Table 2, with 95% CI half-widths compared in Figure 1.

IPOP participants with and without anemia at baseline had similar self-reported alcohol use during pregnancy and similar age and height distributions. Participants without anemia were more likely to have one or more prior births, be in their first trimester, and be on ART. The three doubly robust methods provided similar estimates of the *ACE* for all model specifications examined, with average birth weights if all women had anemia were approximately 40g lower than birth weights if no women had anemia, though all 95% CIs included zero.

Precision estimates based on the empirical sandwich variance estimator and the influence function based variance estimator were similar when the full covariate set was included in both the outcome and propensity models. Under the naive outcome model, standard errors for the influence function based variance estimator were larger than those for the empirical sandwich variance estimator, resulting in wider CIs. When the propensity model was naive, standard errors for the influence function based variance estimator were smaller than those for the empirical sandwich variance estimator, resulting in narrower CIs. Note there was more fluctuation in 95% CI half-widths across model specifications for the influence function based variance estimator than the empirical sandwich variance estimator (Figure 1).

Table 1: Baseline characteristics of IPOP participants by anemia status

		No anemia <i>n</i> = 664	Anemia <i>n</i> = 118
Age category	18-24	156 (23%)	28 (24%)
	25-29	190 (29%)	33 (28%)
	30-34	192 (29%)	31 (26%)
	35-39	100 (15%)	21 (18%)
	40+	26 (4%)	5 (4%)
Number of prior births	0	125 (19%)	30 (25%)
	1	160 (24%)	25 (21%)
	2	157 (24%)	28 (24%)
	3	125 (19%)	18 (15%)
	4+	97 (15%)	17 (14%)
First trimester		120 (18%)	12 (10%)
ART use		648 (98%)	109 (92%)
Alcohol use during pregnancy		64 (10%)	9 (8%)
Height	Median (Q1,Q3)	155 (150, 160)	154 (150, 159)
	Mean (SD)	155 (8)	153 (8)
	Min, Max	131, 185	132, 170

Note: Anemia defined as baseline hemoglobin below 10.5 g/dL

Table 2: Estimated effect of maternal anemia on birth weight by model specification and estimator

	ACE	ES-SE	ES-95% CI	IF-SE	IF-95% CI
Full Covariate Set					
Classic AIPW	-37	56	(-147, 73)	58	(-151, 76)
Weighted regression AIPW	-41	56	(-151, 69)	56	(-151, 68)
TMLE	-37	56	(-147, 73)	58	(-151, 76)
Naive Outcome Model					
Classic AIPW	-36	57	(-148, 76)	61	(-156, 84)
Weighted regression AIPW	-36	57	(-148, 77)	61	(-156, 84)
TMLE	-36	57	(-148, 77)	61	(-156, 84)
Naive Propensity Model					
Classic AIPW	-41	58	(-154, 73)	56	(-150, 69)
Weighted regression AIPW	-47	57	(-159, 65)	54	(-153, 59)
TMLE	-40	58	(-154, 73)	56	(-150, 69)

Note: SE=standard error, ES=empirical sandwich, IF=influence function, CI=confidence interval. The naive outcome model included only the exposure, and the naive propensity model included only an intercept.

4 Simulation Study

A simulation study was conducted to compare the empirical properties of the influence function based variance estimator and the empirical sandwich variance estimator in conjunction with each of the three doubly robust methods discussed in Section 2.

4.1 Simulation Setup

Simulations were designed based on the example in Section 3, which aimed to estimate the effect of maternal anemia on birth weight among women with HIV. The primary simulations were conducted with $n = 800$ participants, similar to the sample size in the example. Results for a sample size of $n = 2000$ are included in the Supporting Information. The simulation study was conducted in R with code provided on GitHub.

Three covariates (Z_1, Z_2, Z_3), the exposure (X), and potential outcomes (Y^0 and Y^1) were simulated. The covariate Z_1 was distributed Normal with mean 155 and standard deviation 7.6. Two binary covariates Z_2 and Z_3 were simulated from Bernoulli distributions with means 0.25 and 0.75, respectively. The exposure X was simulated from a Bernoulli distribution with mean $\text{expit}(15 - 0.1Z_1 + 2.5Z_2 - 1Z_3 - 0.02Z_1Z_2 + 0.005Z_1Z_3)$. Potential outcomes Y^0 and Y^1 under exposure and no exposure, respectively, were simulated from a normal distribution with mean

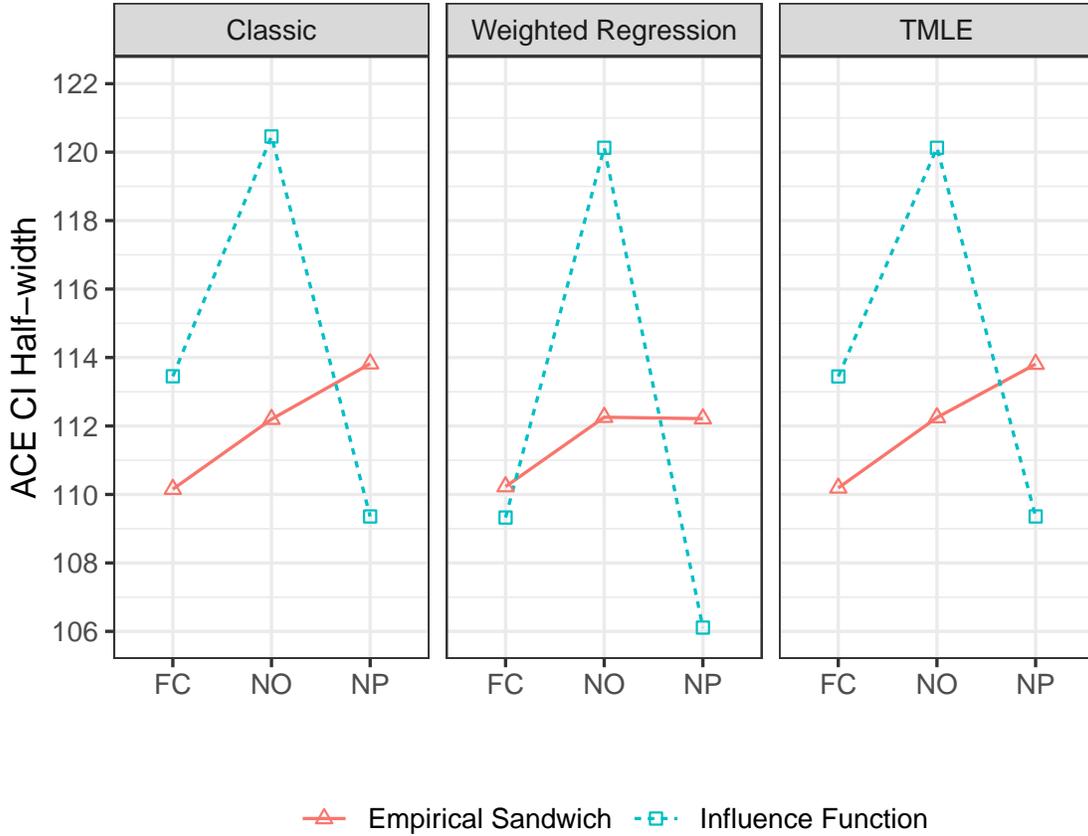


Figure 1: 95% confidence interval (CI) half-widths by model specification and estimator. FC=Full Covariate Set, NO=Naive Outcome Model, NP=Naive Propensity Model

$E(Y^x) = 1000 + 11.5Z_1 + 100Z_2 - 15Z_1Z_2 + 25x - 5.5xZ_1 - 30xZ_2 + 20xZ_1Z_2$ and standard deviation $\sigma = 400$ for $x \in \{0, 1\}$. Under this data generating mechanism, the ACE was approximately -60 . Additional simulations were conducted with $\sigma = 200$ with results in the Supporting Information.

The estimators \widehat{DR}_C , \widehat{DR}_{WR} , and \widehat{DR}_{TMLE} were applied to 5000 simulated samples to estimate the ACE . Estimators were computed four ways: (1) with both propensity and outcome models correctly specified, (2) with propensity models correctly specified but outcome models misspecified, (3) with outcome models correctly specified but propensity models misspecified, and (4) with both models misspecified. Misspecified propensity models included only an intercept and a linear term for $(Z_1 - 155)^2$, and misspecified outcome models included only an intercept and linear terms for X and $(Z_1 - 155)^2$.

Both the influence function based variance estimator and the empirical sandwich variance estimator were applied, as described in Section 2.3. For each of the three doubly robust estimators, 95% CIs for ACE were constructed based on each variance estimation approach. Simulation results were summarized by computing empirical bias, relative empirical bias ($|\text{bias}|/|ACE|$), average standard error (ASE), empirical standard error (ESE), standard error ratio ($SER=ASE/ESE$), and empirical 95% CI coverage. Coverage was defined as the proportion of simulated samples in which the CI included ACE . In addition, the ratio of the variance estimate for each simulation and the ESE, i.e., the variance ratio (VR), was computed and summarized. That is, $VR = se_s/ESE$, where $se_s = \sqrt{\widehat{V}(\widehat{DR})}$ for a given doubly robust estimator, model specification, and variance estimator for simulation s .

Table 3: Simulation summary results, $n = 800$, $\sigma = 400$, 5000 simulations. Bias, Relative Bias, ASE, ESE, SER, and 95% CI coverage calculated for the ACE.

Scenario	Estimator	Bias	Relative Bias (%)	ESE	ASE, ES	SER, ES	Cov, ES (%)	ASE, IF	SER, IF	Cov, IF (%)
CS	Classic	0.4	0.6	58.4	58.0	0.99	95	58.1	1.00	95
	WR	0.4	0.6	58.3	57.7	0.99	95	57.6	0.99	95
	TMLE	0.4	0.6	58.3	57.9	0.99	95	58.1	1.00	95
MO	Classic	-0.4	0.6	60.0	59.2	0.99	95	64.5	1.07	97
	WR	-1.7	2.8	59.2	58.4	0.99	95	62.4	1.05	96
	TMLE	-0.5	0.8	59.7	58.9	0.99	95	63.0	1.06	96
MW	Classic	0.3	0.5	57.8	57.6	1.00	95	55.8	0.97	94
	WR	0.2	0.4	57.7	57.6	1.00	95	55.8	0.97	94
	TMLE	0.3	0.5	57.8	57.6	1.00	95	55.8	0.97	94
MB	Classic	-23.8	39.8	57.0	56.8	1.00	92	56.8	1.00	92
	WR	-23.8	39.8	57.0	56.8	1.00	92	56.8	1.00	92
	TMLE	-23.8	39.8	57.0	56.8	1.00	92	56.8	1.00	92

ASE=average estimated standard error; ESE=empirical standard error; SER=standard error ratio (ASE/ESE); Cov = 95% confidence interval coverage; ES=empirical sandwich variance estimator; IF=influence function based variance estimator; CS=correct specification of both models; MO=misspecified outcome model, MW=misspecified propensity model; MB=misspecified both models; WR=weighted regression AIPW; ACE was approximately -60 ; Monte Carlo standard error for 95% CI coverage was 0.3% when coverage was 95%. Results exclude one simulation where models failed to converge.

4.2 Simulation Results

Both point and variance estimators performed as expected in simulations. The results of the simulation study for $n = 800$ and $\sigma = 400$ are presented in Table 3 and Figure 2. When at least one model was correctly specified, the classic AIPW, weighted regression AIPW, and TMLE displayed minimal bias, but all were substantially biased when both nuisance models were misspecified. Note bias was further reduced under correct specification of at least one model for $n = 2000$ (Table 4). Under correct specification of both nuisance models, the empirical sandwich and influence function based variance estimators tracked closely with the ESE, resulting in SERs close to one and CIs attaining the nominal level of coverage. When both models were misspecified, both variance estimators tracked with the ESE, but bias was substantial, resulting in CIs with below nominal coverage.

Differences between the variance estimators are apparent when examining the simulation study results for scenarios where one nuisance model was correctly specified and the other was misspecified. When either the outcome model or the propensity model was misspecified, CIs based on the empirical sandwich variance estimator attained the nominal level of coverage, and estimated standard errors tracked closely with the ESE. In Figure 2, note variance ratios for the empirical sandwich estimator are closely clustered around one for these scenarios. In contrast, the influence function based variance estimator was empirically biased when either nuisance model was misspecified. As expected, it tended to overestimate the variance when the outcome model was misspecified, leading to estimated standard errors that tracked above the ESE, SERs above one, and conservative CIs that exceeded the nominal level of coverage. When the propensity model was misspecified but the outcome model was correctly specified, the influence function based variance estimator underestimated the variance, resulting in estimated standard errors below the ESE and SERs below one. The biases observed for the influence function based variance estimator were also present when $n = 2000$ (Table 4 and Figure 3).

In the $\sigma = 200$ scenarios, the empirical sandwich variance estimator performance was similar to the $\sigma = 400$ scenario, but the performance of the influence function based variance estimator varied (Tables 5-6 and Figures 4-5). Specifically, when the propensity model was misspecified but the outcome model was correctly specified, the influence function based variance estimator demonstrated minimal empirical bias, resulting in SERs close to one.

In summary, the simulation study demonstrated the theoretical properties explained in Section 2. That is, the empirical sandwich variance estimator was consistent when at least one of the two nuisance models was correctly specified, and performed well in simulations. The influence function based variance estimator is not consistent under misspecification of either nuisance model. In line with expectations, the influence function based variance estimator was conservative when the outcome model was misspecified. For this data generating mechanism, it was anti-conservative when the propensity model was misspecified for $\sigma = 400$ and exhibited minimal bias when $\sigma = 200$.

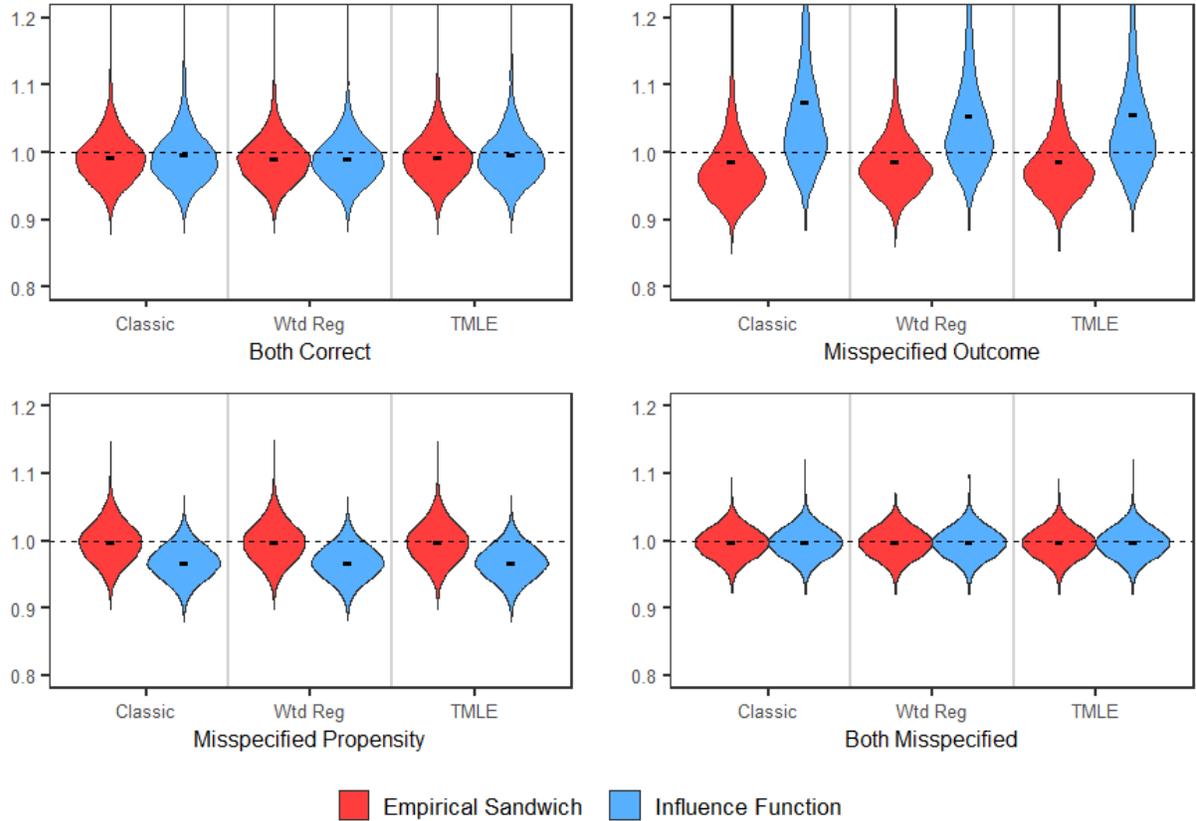


Figure 2: Ratio between each simulation’s estimated standard error and the empirical standard error by estimator and model specification, $n=800$, $\sigma = 400$, 5000 simulations. Black squares denote the mean variance ratio (=SER). Results exclude one simulation where models failed to converge. Note the 0.3% of correct model specification simulations and 4.5% of misspecified outcome model simulations where the ratio was above 1.2 are not displayed.

5 Discussion

Doubly robust estimators have increased in popularity in the field of causal inference due to their ability to provide consistent point estimates when either an outcome or propensity model is correctly specified (Smith et al., 2023). In this paper, the commonly-used influence function based variance estimator is compared with the empirical sandwich variance estimator in conjunction with three doubly robust estimators: the classic AIPW estimator, the weighted regression AIPW estimator, and TMLE. These three estimators can be specified as solutions to sets of unbiased estimating equations, and thus their variances can be consistently estimated by the empirical sandwich variance estimator. Because the estimating equations remain unbiased when either the outcome or propensity model is correctly specified, the empirical sandwich variance estimator is a doubly robust variance estimator. In contrast, the influence function based variance estimator is consistent only when both outcome and propensity models are correctly specified. This discrepancy may be of particular importance in settings where the functional form of one of the nuisance models is known but the other is not. For example, with randomized trials, randomization to treatment means that the exposure mechanism is known. Doubly robust estimators can offer improved efficiency in this setting, but some of that efficiency gain may not be realized with application of the influence function based variance estimator, which is conservative if the outcome model is misspecified.

While consideration in this paper was limited to finite-dimensional parametric modeling approaches, machine learning approaches can also be used with doubly robust estimators (Kennedy et al., 2016; van der Laan and Gruber, 2011, 2012; van der Laan et al., 2018; Balzer et al., 2023). Machine learning approaches allow for more complex functional forms for continuous covariates than fully parametric approaches and include higher order interaction terms that may not be included in investigator-specified parametric models, often leading to estimators that are more robust to model misspecification (Zivich et al., 2022). However, convergence of machine learning algorithms is typically slower and

consistent variance estimation is more challenging than with parametric modeling approaches. Machine learning methods are typically not compatible with the estimating equations approach discussed in this paper, though alternative methods have been developed for doubly robust variance estimation in this context for TMLE estimators (Benkeser et al., 2017).

In the example, the effect of anemia on birth weight among women with HIV was estimated by applying the three doubly robust methods to data from the IPOP randomized trial. Average birth weights if all women had anemia were estimated to be approximately 40g lower than birth weights if no women had anemia, though all 95% CIs included zero. The analysis included 14 participants who experienced stillbirth. Because stillbirth is a competing outcome for live birth weight, future work could reexamine this problem using alternative approaches for competing events.

Use of the empirical sandwich variance estimator in conjunction with doubly robust estimators is not new, though its doubly robust property has not been emphasized. Lunceford and Davidian (2004) discuss use of the empirical sandwich variance estimator for estimating the variance of the classic AIPW estimator and note that it tends to be more stable than the influence function based variance estimator. However, based on the systematic review discussed in the introduction (Smith et al., 2023), the influence function based variance estimator is frequently applied and the empirical sandwich variance estimator does not appear to be widely used with TMLE. While some existing software packages compute empirical sandwich variance estimators for doubly robust estimators (e.g., the “causaltrt” procedure in SAS and the “dr” procedure in Stata), other popular software packages compute the influence function based variance estimator but not the empirical sandwich variance estimator (e.g., the AIPW and `tmle` packages in R and the `zEpid` package in Python). We provide example R and Python code for estimating the variance of each of the three doubly robust estimators with the empirical sandwich variance estimator, which we hope will allow for easier implementation of doubly robust point and variance estimation.

Supplementary Material

R and Python code for computing the different estimators along with the corresponding standard error estimators is available at <https://github.com/bonnieshook/DoublyRobustVariance>.

Bibliography

References

- Azizah, F. K., Dewi, Y. L. R., and Murti, B. (2022). The effect of maternal anemia on low birth weight: a systematic review and meta analysis. *Journal of Maternal and Child Health*, 7(1):34–43.
- Balzer, L. B., van der Laan, M., Ayieko, J., Kanya, M., Chamie, G., Schwab, J., Havlir, D. V., and Petersen, M. L. (2023). Two-stage tmle to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics*, 24(2):502–517.
- Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.
- Benkeser, D., Carone, M., Laan, M. V. D., and Gilbert, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880.
- Chua, K.-P., Thorne, M. C., Ng, S., Donahue, M., and Brummett, C. M. (2022). Association between default number of opioid doses in electronic health record systems and opioid prescribing to adolescents and young adults undergoing tonsillectomy. *JAMA Network Open*, 5(6):e2219701–e2219701.
- Daniel, R. M. (2014). Double robustness. *Wiley StatsRef: Statistics Reference Online*, pages 1–14.
- Figueiredo, A. C., Gomes-Filho, I. S., Silva, R. B., Pereira, P. P., Mata, F. A. D., Lyrio, A. O., Souza, E. S., Cruz, S. S., and Pereira, M. G. (2018). Maternal anemia and low birth weight: a systematic review and meta-analysis. *Nutrients*, 10(5):601.
- Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.
- Gabriel, E. E., Sachs, M. C., Martinussen, T., Waernbaum, I., Goetghebeur, E., Vansteelandt, S., and Sjölander, A. (2023). Inverse probability of treatment weighting with generalized linear outcome models for doubly robust estimation. *Statistics in Medicine*, 10.1002/sim.9969:1–14.
- Gerds, C., Fuentes, L., Grossman, D., White, K., Keefe-Oates, B., Baum, S. E., Hopkins, K., Stolp, C. W., and Potter, J. E. (2022). Impact of clinic closures on women obtaining abortion services after implementation of a restrictive law in texas. *American Journal of Public Health*, 112(9):1297–1304.

- Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.
- Gruber, S. and Van Der Laan, M. (2012). tml: an r package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51:1–35.
- Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570.
- Hernán, M. Á. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Hoffman, K. L., Schenck, E. J., Satlin, M. J., Whalen, W., Pan, D., Williams, N., and Díaz, I. (2022). Comparison of a target trial emulation framework vs cox regression to estimate the association of corticosteroids with covid-19 mortality. *JAMA Network Open*, 5(10):e2234425–e2234425.
- Josey, K. P., Delaney, S. W., Wu, X., Nethery, R. C., DeSouza, P., Braun, D., and Dominici, F. (2023). Air pollution and mortality at the intersection of race and social class. *New England Journal of Medicine*, 388(15):1396–1404.
- Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.
- Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2016). Non-parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245.
- Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.
- Lee, M. M., Falbe, J., Schillinger, D., Basu, S., McCulloch, C. E., and Madsen, K. A. (2019). Sugar-sweetened beverage consumption 3 years after the berkeley, california, sugar-sweetened beverage tax. *American Journal of Public Health*, 109(4):637–639.
- Levine, A. M., Berhane, K., Masri-Lavine, L., Sanchez, M. L., Young, M., Augenbraun, M., Cohen, M., Anastos, K., Newman, M., Gange, S. J., and Watts, H. (2001). Prevalence and correlates of anemia in a large cohort of hiv-infected women: Women’s interagency hiv study. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 26(1):28–35.
- Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.
- Luque-Fernandez, M. A., Schomaker, M., Rachet, B., and Schnitzer, M. E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16):2530–2546.
- Muñoz, I. D. and Van Der Laan, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549.
- Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in r. *BMC Medical Research Methodology*, 19(1):1–16.
- Price, J. T., Vwalika, B., Freeman, B. L., Cole, S. R., Saha, P. T., Mbewe, F. M., Phiri, W. M., Peterson, M., Muyangwa, D., Sindano, N., et al. (2021). Weekly 17 alpha-hydroxyprogesterone caproate to prevent preterm birth among women living with hiv: a randomised, double-blind, placebo-controlled trial. *The Lancet HIV*, 8(10):e605–e613.
- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.
- Robins, J. (1998). Marginal structural models. In *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pages 1–10.
- Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science*, 22(4):544–559.
- Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in Epidemiology. *Epidemiology*, 11(5):550–560.
- Robins, J. M., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.
- Saul, B. C. and Hudgens, M. G. (2020). The calculus of M-estimation in R with geex. *Journal of Statistical Software*, 92(2):1–15.
- Shook-Sa, B. E., Hudgens, M. G., Knittel, A. K., Edmonds, A., Ramirez, C., Cole, S. R., Cohen, M., Adedimeji, A., Taylor, T., Michel, K. G., Kovacs, A., Cohen, J., Donohue, J., Foster, A., Fischl, M. A., Long, D., and Adimora, A. A. (2023). Exposure effects on count outcomes with observational data, with application to incarcerated women. *arXiv preprint: 2202.01650*.

- Smith, M. J., Phillips, R. V., Luque-Fernandez, M. A., and Maringe, C. (2023). Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review. *Annals of Epidemiology*, 86:34–48.
- Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.
- Stitelman, O. M., De Gruttola, V., and van der Laan, M. J. (2012). A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The International Journal of Biostatistics*, 8(1):1–37.
- Sullivan, D. J., Gebo, K. A., Shoham, S., Bloch, E. M., Lau, B., Shenoy, A. G., Mosnaim, G. S., Gniadek, T. J., Fukuta, Y., Patel, B., et al. (2022). Early outpatient treatment for covid-19 with convalescent plasma. *New England Journal of Medicine*, 386(18):1700–1711.
- Tran, L., Yiannoutsos, C., Wools-Kaloustian, K., Siika, A., van der Laan, M., and Petersen, M. (2019). Double robust efficient estimators of longitudinal treatment effects: Comparative performance in simulations and a case study. *The International Journal of Biostatistics*, 15(2):1–27.
- Tukei, V. J., Hoffman, H. J., Greenberg, L., Thabelo, R., Nchephe, M., Mots’oane, T., Masitha, M., Chabela, M., Mokone, M., Mofenson, L., Guay, L., and Tiam, A. (2021). Adverse pregnancy outcomes among hiv-positive women in the era of universal antiretroviral therapy remain elevated compared with hiv-negative women. *The Pediatric Infectious Disease Journal*, 40(9):821.
- van der Laan, M. J. and Gruber, S. (2011). Targeted minimum loss based estimation of an intervention specific mean outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*, pages 1–38.
- van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1):1–39.
- van der Laan, M. J., Rose, S., Sekhon, J. S., Gruber, S., Porter, K. E., and van der Laan, M. J. (2011). Propensity-score-based estimators and c-tmle. *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 343–364.
- van der Laan, M. J., Rose, S., van der Laan, M. J., Bibaut, A., and Luedtke, A. R. (2018). Cv-tmle for nonpathwise differentiable target parameters. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 455–481.
- van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):1–38.
- Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1):7–30.
- Vansteelandt, S. and Keiding, N. (2011). Invited commentary: G-computation—lost in translation? *American Journal of Epidemiology*, 173(7):739–742.
- Yazdi, M. D., Wang, Y., Di, Q., Requia, W. J., Wei, Y., Shi, L., Sabath, M. B., Dominici, F., Coull, B., Evans, J. S., et al. (2021). Long-term effect of exposure to lower concentrations of air pollution on mortality among us medicare participants and vulnerable subgroups: a doubly-robust approach. *The Lancet Planetary Health*, 5(10):e689–e697.
- Zhong, Y., Auchincloss, A. H., Lee, B. K., and Kanter, G. P. (2018). The short-term impacts of the philadelphia beverage tax on beverage consumption. *American Journal of Preventive Medicine*, 55(1):26–34.
- Zhong, Y., Kennedy, E. H., Bodnar, L. M., and Naimi, A. I. (2021). AIPW: An R Package for Augmented Inverse Probability–Weighted Estimation of Average Causal Effects. *American Journal of Epidemiology*, 190(12):2690–2699.
- Zivich, P. N., Breskin, A., and Kennedy, E. H. (2022). Machine learning and causal inference. *Wiley StatsRef: Statistics Reference Online*, pages 1–8.
- Zivich, P. N., Davidson-Pilon, C., Diong, J., Reger, D., and Badger, T. G. (2022a). *pzivich/zEpid: v0.9.1*. Zenodo.
- Zivich, P. N., Hudgens, M. G., Brookhart, M. A., Moody, J., Weber, D. J., and Aiello, A. E. (2022b). Targeted maximum likelihood estimation of causal effects with interference: A simulation study. *Statistics in Medicine*, 41(23):4554–4577.
- Zivich, P. N., Klose, M., Cole, S. R., Edwards, J. K., and Shook-Sa, B. E. (2022c). Delicatessen: M-estimation in python. *arXiv preprint: 2203.11300*.

Acknowledgements

This work was supported by US National Institutes of Health under award numbers R01 AI157758, R01 AI085073, P30 AI050410, and K01 AI177102. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Bradley Saul at the University of North Carolina at Chapel Hill for coding support.

Appendix

Consistency proof of TMLE estimator under correct specification of either the outcome or propensity model

Recall the TMLE estimating function ψ_3 is specified as

$$\psi_3(O_i; \theta) = \begin{bmatrix} \psi_\alpha \\ \psi_\gamma \\ \psi_{t1} \\ \psi_{t0} \\ \psi_{\mu_{TMLE1}} \\ \psi_{\mu_{TMLE0}} \\ \psi_{TMLE} \end{bmatrix} = \begin{bmatrix} \{X_i - \text{expit}(G_i^T \alpha)\} G_i \\ \{Y_i^* - \phi^{-1}(H_i^T \gamma)\} H_i \\ X_i e_i^{-1} (Y_i^* - \text{expit}[\eta_1 + \text{logit}\{a_1(Z_i, \gamma)\}]) \\ (1 - X_i)(1 - e_i)^{-1} (Y_i^* - \text{expit}[\eta_0 + \text{logit}\{a_0(Z_i, \gamma)\}]) \\ c_1(O_i, \gamma, \eta_1) - \mu_{TMLE}^1 \\ c_0(O_i, \gamma, \eta_0) - \mu_{TMLE}^0 \\ \mu_{TMLE}^1 - \mu_{TMLE}^0 - DR_{TMLE} \end{bmatrix}$$

where $\theta = [\alpha, \gamma, \eta_1, \eta_0, \mu_{TMLE}^1, \mu_{TMLE}^0, DR_{TMLE}]$ is the vector of parameters.

Let

$\tilde{\theta} = [\tilde{\alpha}, \tilde{\gamma}, \tilde{\eta}_1, \tilde{\eta}_0, \tilde{\mu}_{TMLE}^1, \tilde{\mu}_{TMLE}^0, \widehat{DR}_{TMLE}]$ be the solution of $E\{\psi_3(O_i; \theta)\} = 0$ and \widehat{DR}_{TMLE} be the proposed TMLE estimator of ACE, obtained by solving the estimating equation $n^{-1} \sum_i \psi_3(O_i; \theta) = 0$. Then, if either the outcome model or the propensity score model is correctly specified, $\sqrt{n}(\widehat{DR}_{TMLE} - ACE) \xrightarrow{d} N(0, \sigma^2(\tilde{\theta}))$, where $\sigma^2(\tilde{\theta})$ is the bottom right element of $V(\tilde{\theta}) = V(\theta)|_{\theta=\tilde{\theta}}$, and $V(\theta) = A(\theta)^{-1}B(\theta)A(\theta)^{-T}$, with $A(\theta) = E\{-\partial\psi_3(O_i; \theta)/\partial\theta\}$, $B(\theta) = E\{\psi_3(O_i; \theta)\psi_3(O_i; \theta)^\top\}$, which is a doubly robust variance estimator of \widehat{DR}_{TMLE} .

Proof.

Let $\hat{\theta}$ be the solution of the estimating equation $n^{-1} \sum_i \psi_3(O_i; \theta) = 0$. Under suitable regularity conditions (Stefanski and Boos, 2002), $\sqrt{n}(\hat{\theta} - \tilde{\theta}) \xrightarrow{d} N(0, V(\tilde{\theta}))$. Note the final element of $\hat{\theta}$ is \widehat{DR}_{TMLE} and therefore $\sqrt{n}(\widehat{DR}_{TMLE} - \widehat{DR}_{TMLE}) \xrightarrow{d} N(0, \sigma^2(\tilde{\theta}))$. It remains to show that if either the outcome model or the propensity score model is correctly specified, $\widehat{DR}_{TMLE} = ACE$.

First, let $\theta_0 = [\alpha_0, \gamma_0, \eta_{10}, \eta_{00}, \mu_{TMLE0}^1, \mu_{TMLE0}^0, DR_{TMLE0}]$ be the true parameter values from the data generating process. Then, $P(X_i = 1|Z_i) = e(Z_i, \alpha_0) = \text{expit}(G_i^T \alpha_0)$ and $E(Y_i^*|Z_i, X_i) = \phi^{-1}(H_i^T \gamma_0) = X_i a_1(Z_i, \gamma_0) + (1 - X_i) a_0(Z_i, \gamma_0)$.

Now assume the outcome model is correctly specified, i.e., $\tilde{\gamma} = \gamma_0$ and $E(Y_i^*|Z_i, X_i) = X_i a_1(Z_i, \gamma_0) + (1 - X_i) a_0(Z_i, \gamma_0)$. From $E\{\psi_{t1}(O_i; \tilde{\theta})\} = 0$,

$$\begin{aligned} 0 &= E\{\psi_{t1}(O_i; \tilde{\theta})\} \\ &= E[E\{\psi_{t1}(O_i; \tilde{\theta})|X_i\}] \\ &= E[E\{X_i e(Z_i, \tilde{\alpha})^{-1} (Y_i^* - \text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \gamma_0)\}]) | X_i\}] \\ &= E\{e(Z_i, \tilde{\alpha})^{-1} (Y_i^* - \text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \gamma_0)\}]) | X_i = 1\} P(X_i = 1) \\ &= E[E\{e(Z_i, \tilde{\alpha})^{-1} (Y_i^* - \text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \gamma_0)\}]) | Z_i, X_i = 1\} | X_i = 1] P(X_i = 1) \\ &= E[e(Z_i, \tilde{\alpha})^{-1} \{E(Y_i^*|Z_i, X_i = 1) - \text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \gamma_0)\}]\} | X_i = 1] P(X_i = 1) \\ &= E[e(Z_i, \tilde{\alpha})^{-1} \{a_1(Z_i, \gamma_0) - \text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \gamma_0)\}]\} | X_i = 1] P(X_i = 1) \\ &= E[e(Z_i, \tilde{\alpha})^{-1} F(\tilde{\eta}_1) | X_i = 1] P(X_i = 1) \end{aligned}$$

where $F(t) = \text{expit}(d) - \text{expit}(t + d)$ and $d = \text{logit}\{a_1(Z_i, \gamma_0)\}$. Since the function $F(t)$ is strictly decreasing in t , it can be shown that the last line in the above is strictly decreasing in $\tilde{\eta}_1$ and attains 0 only at $\tilde{\eta}_1 = 0$. Therefore, $\tilde{\eta}_1 = 0$.

Then, from $E\{\psi_{\mu_{TMLE1}}(O_i; \tilde{\theta})\} = 0$,

$$\begin{aligned}
\tilde{\mu}_{TMLE}^1 &= E(Y_{i, TMLE}^1) \\
&= E\{\text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \gamma_0)\}](b-a) + a\} \\
&= E\{a_1(Z_i, \gamma_0)\}(b-a) + a \\
&= E[E\{Y_i^* | Z_i, X_i = 1\}](b-a) + a \\
&= E[E\{(Y_i - a)/(b-a) | Z_i, X_i = 1\}](b-a) + a \\
&= E\{E(Y_i | Z_i, X_i = 1)\} \\
&= \mu^1.
\end{aligned}$$

Similarly, $\tilde{\mu}_{TMLE}^0 = \mu^0$.

Next, assume the propensity score model is correctly specified, i.e., $\tilde{\alpha} = \alpha_0$ and $P(X_i = 1 | Z_i) = e(Z_i, \alpha_0)$. From $E\{\psi_{t1}(O_i; \tilde{\theta})\} = 0$,

$$\begin{aligned}
0 &= E\{\psi_{t1}(O_i; \tilde{\theta})\} \\
&= E[E\{\psi_{t1}(O_i; \tilde{\theta}) | Z_i\}] \\
&= E[E\{X_i e(Z_i, \alpha_0)^{-1} \{(Y_i^1 - a)/(b-a) - \text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \tilde{\gamma})\}]\} | Z_i\}] \\
&= E(E(X_i | Z_i) e(Z_i, \alpha_0)^{-1} [E\{(Y_i^1 - a)/(b-a) | Z_i\} - \text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \tilde{\gamma})\}]]) \\
&= E[E\{(Y_i^1 - a)/(b-a) | Z_i\}] - E(\text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \tilde{\gamma})\}]) \\
&= (\mu^1 - a)/(b-a) - E(\text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \tilde{\gamma})\}]).
\end{aligned}$$

Therefore, $E(\text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \tilde{\gamma})\}]) = (\mu^1 - a)/(b-a)$. Then, from $E\{\psi_{\mu_{TMLE1}}(O_i; \tilde{\theta})\} = 0$,

$$\begin{aligned}
\tilde{\mu}_{TMLE}^1 &= E\{c_x(O_i, \gamma, \eta_x)\} \\
&= E\{\text{expit}[\tilde{\eta}_1 + \text{logit}\{a_1(Z_i, \tilde{\gamma})\}](b-a) + a\} \\
&= \mu^1.
\end{aligned}$$

Similarly, $\tilde{\mu}_{TMLE}^0 = \mu^0$.

In conclusion, if either the outcome model or the propensity score model is correctly specified, $\widetilde{DR}_{TMLE} = \tilde{\mu}_{TMLE}^1 - \tilde{\mu}_{TMLE}^0 = \mu^1 - \mu^0 = ACE$ which proves the stated result.

Table 4: Simulation summary results, $n = 2000$, $\sigma = 400$, 5000 simulations. Bias, Relative Bias, ASE, ESE, SER, and 95% CI coverage calculated for the ACE.

Scenario	Estimator	Bias	Relative Bias (%)	ESE	ASE, ES	SER, ES	Cov, ES (%)	ASE, IF	SER, IF	Cov, IF (%)
CS	Classic	-0.1	0.2	36.6	36.7	1.00	95	36.8	1.00	95
	WR	-0.1	0.2	36.6	36.6	1.00	95	36.6	1.00	95
	TMLE	-0.1	0.2	36.6	36.7	1.00	95	36.8	1.00	95
MO	Classic	-0.3	0.6	37.2	37.4	1.01	95	40.6	1.09	97
	WR	-0.9	1.6	37.0	37.1	1.00	95	39.6	1.07	96
	TMLE	-0.4	0.6	37.1	37.3	1.01	95	39.8	1.07	96
MW	Classic	-0.2	0.3	36.2	36.4	1.01	95	35.3	0.98	94
	WR	-0.2	0.3	36.2	36.4	1.01	95	35.3	0.98	95
	TMLE	-0.2	0.3	36.2	36.4	1.01	95	35.3	0.98	94
MB	Classic	-23.8	39.8	35.9	35.9	1.00	90	35.9	1.00	90
	WR	-23.8	39.8	35.9	35.9	1.00	90	35.9	1.00	90
	TMLE	-23.8	39.8	35.9	35.9	1.00	90	35.9	1.00	90

ASE=average estimated standard error; ESE=empirical standard error; SER=standard error ratio (ASE/ESE); Cov = 95% confidence interval coverage; ES=empirical sandwich variance estimator; IF=influence function based variance estimator; CS=correct specification of both models; MO=misspecified outcome model, MW=misspecified propensity model; MB=misspecified both models; WR=weighted regression AIPW; ACE was approximately -60 ; Monte Carlo standard error for 95% CI coverage was 0.3% when coverage was 95%.

Table 5: Simulation summary results, $n = 800$, $\sigma = 200$, 5000 simulations. Bias, Relative Bias, ASE, ESE, SER, and 95% CI coverage calculated for the ACE.

Scenario	Estimator	Bias	Relative Bias (%)	ESE	ASE, ES	SER, ES	Cov, ES (%)	ASE, IF	SER, IF	Cov, IF (%)
CS	Classic	0.3	0.6	50.2	50.0	1.00	95	50.0	1.00	95
	WR	0.3	0.5	50.1	49.9	1.00	95	49.9	1.00	95
	TMLE	0.3	0.5	50.2	50.0	1.00	95	50.0	1.00	95
MO	Classic	-0.4	0.6	52.0	51.5	0.99	95	57.2	1.10	97
	WR	-1.6	2.7	51.3	50.7	0.99	95	55.1	1.07	96
	TMLE	-0.5	0.8	51.6	51.1	0.99	95	55.6	1.08	96
MW	Classic	0.3	0.5	50.0	49.9	1.00	95	49.4	0.99	95
	WR	0.3	0.4	50.0	49.9	1.00	95	49.4	0.99	95
	TMLE	0.3	0.5	50.0	49.9	1.00	95	49.4	0.99	95
MB	Classic	-23.6	39.4	50.7	50.4	0.99	92	50.4	0.99	92
	WR	-23.6	39.4	50.7	50.4	0.99	92	50.4	0.99	92
	TMLE	-23.5	39.3	50.7	50.4	0.99	92	50.4	0.99	92

ASE=average estimated standard error; ESE=empirical standard error; SER=standard error ratio (ASE/ESE); Cov = 95% confidence interval coverage; ES=empirical sandwich variance estimator; IF=influence function based variance estimator; CS=correct specification of both models; MO=misspecified outcome model, MW=misspecified propensity model; MB=misspecified both models; WR=weighted regression AIPW; ACE was approximately -60 ; Monte Carlo standard error for 95% CI coverage was 0.3% when coverage was 95%. Results exclude 14 simulations where models did not converge.

Table 6: Simulation summary results, $n = 2000$, $\sigma = 200$, 5000 simulations. Bias, Relative Bias, ASE, ESE, SER, and 95% CI coverage calculated for the ACE.

Scenario	Estimator	Bias	Relative Bias (%)	ESE	ASE, ES	SER, ES	Cov, ES (%)	ASE, IF	SER, IF	Cov, IF (%)
CS	Classic	0.0	0.0	31.6	31.6	1.00	95	31.7	1.00	95
	WR	0.0	0.1	31.5	31.6	1.00	95	31.6	1.00	95
	TMLE	0.0	0.0	31.6	31.6	1.00	95	31.7	1.00	95
MO	Classic	-0.2	0.4	32.3	32.4	1.00	95	36.0	1.11	97
	WR	-0.8	1.4	32.1	32.1	1.00	95	34.9	1.09	97
	TMLE	-0.3	0.4	32.1	32.3	1.00	95	35.1	1.09	97
MW	Classic	0.0	0.0	31.4	31.6	1.00	95	31.2	0.99	95
	WR	0.0	0.0	31.4	31.6	1.00	95	31.2	0.99	95
	TMLE	0.0	0.0	31.4	31.6	1.00	95	31.2	0.99	95
MB	Classic	-23.6	39.5	32.0	31.9	1.00	89	31.9	1.00	89
	WR	-23.6	39.5	32.0	31.9	1.00	89	31.9	1.00	89
	TMLE	-23.6	39.5	32.0	31.9	1.00	89	31.9	1.00	89

ASE=average estimated standard error; ESE=empirical standard error; SER=standard error ratio (ASE/ESE); Cov = 95% confidence interval coverage; ES=empirical sandwich variance estimator; IF=influence function based variance estimator; CS=correct specification of both models; MO=misspecified outcome model, MW=misspecified propensity model; MB=misspecified both models; WR=weighted regression AIPW; ACE was approximately -60 ; Monte Carlo standard error for 95% CI coverage was 0.3% when coverage was 95%.

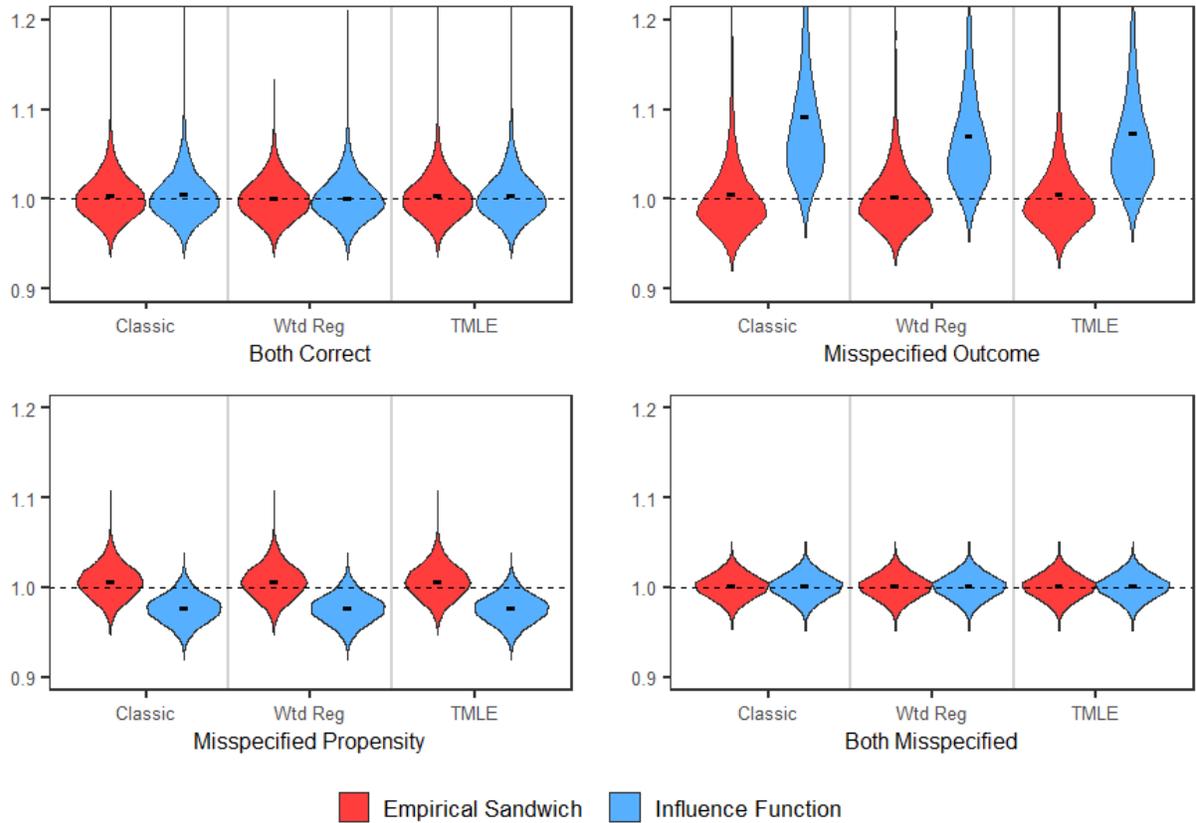


Figure 3: Ratio between each simulation's estimated standard error and the empirical standard error by estimator and model specification, $n=2000$, $\sigma = 400$, 5000 simulations. Black squares denote the mean variance ratio (=SER). Note the 0.08% of correct model specification simulations and 3.4% of misspecified outcome model simulations where the ratio was above 1.2 are not displayed.

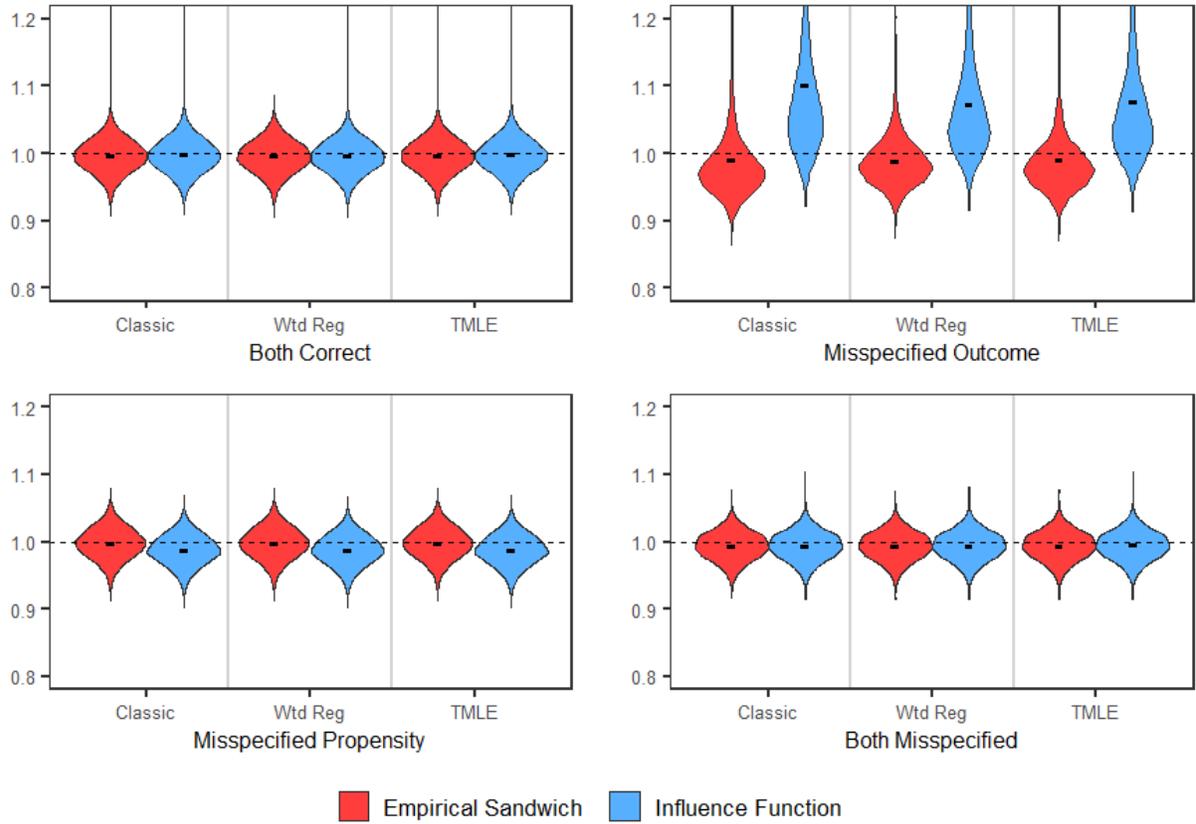


Figure 4: Ratio between each simulation's estimated standard error and the empirical standard error by estimator and model specification, $n=800$, $\sigma = 200$, 5000 simulations. Black squares denote the mean variance ratio (=SER). Results exclude 14 simulations where models did not converge. Note the 0.06% of correct model specification simulations, 5.2% of misspecified outcome model simulations, and 0.01% of simulations where both models were misspecified where the ratio was above 1.2 or below 0.8 are not displayed.

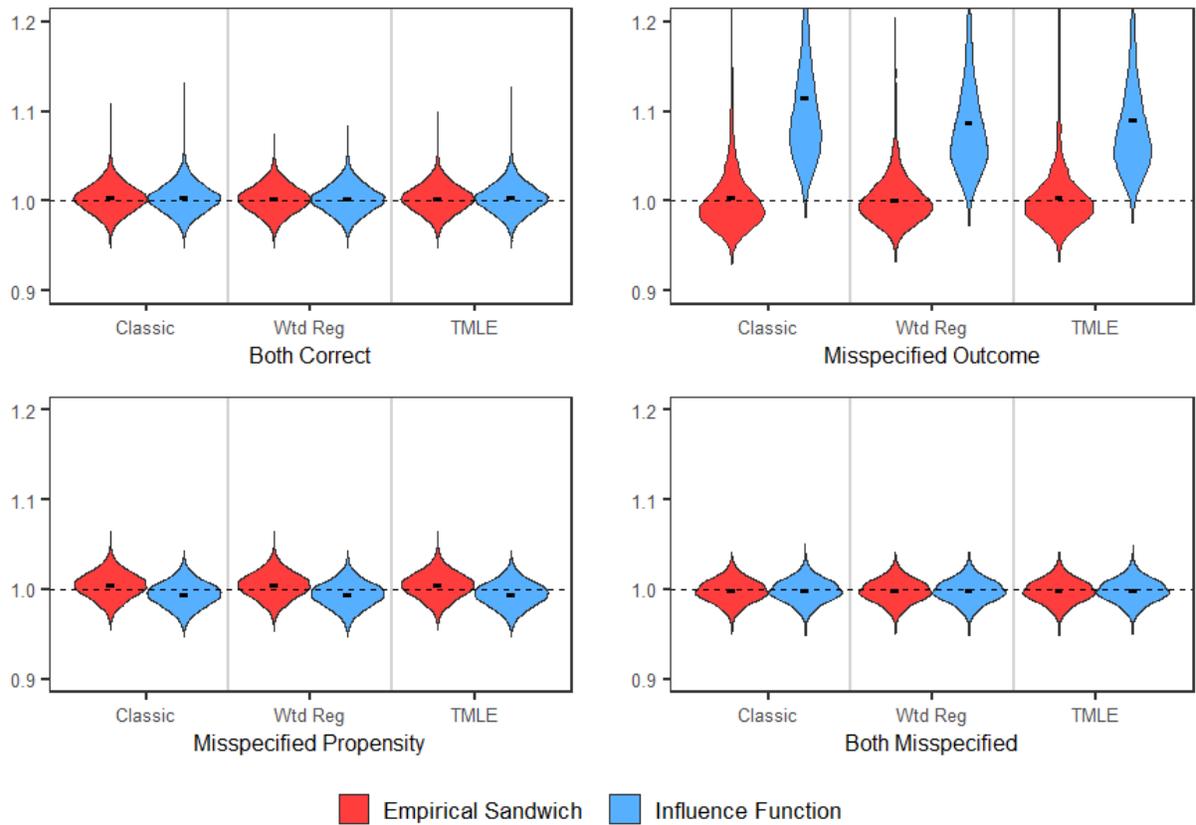


Figure 5: Ratio between each simulation's estimated standard error and the empirical standard error by estimator and model specification, $n=2000$, $\sigma = 200$, 5000 simulations. Black squares denote the mean variance ratio (=SER). Note the 3.9% of misspecified outcome model simulations where the ratio was above 1.2 are not displayed.