Biometrika (2018), **105**, 2, *pp*. 1–7
doi: doi: 10.1093/biomet/asy008
Advance Access publication on 12 March 2018

Printed in Great Britain

Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores

BY S. YANG

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
syang24@ncsu.edu

P. DING

Department of Statistics, University of California, Berkeley, California 94720, U.S.A.
pengdingpku@berkeley.edu

SUMMARY

Causal inference with observational studies often relies on the assumptions of unconfoundedness and overlap of covariate distributions in different treatment groups. The overlap assumption is violated when some units have propensity scores close to 0 or 1, so both practical and theoretical researchers suggest dropping units with extreme estimated propensity scores. However, existing trimming methods often do not incorporate the uncertainty in this design stage and restrict inference only to the trimmed sample, due to the non-smoothness of the trimming. We propose a smooth weighting, which approximates sample trimming and has better asymptotic properties. An advantage of our estimator is its asymptotic linearity, which ensures that the bootstrap can be used to make inference for the target population, incorporating uncertainty arising from both the design and analysis stages. We extend the theory to the average treatment effect on the treated, suggesting trimming samples with estimated propensity scores close to 1.

Some key words: Bootstrap; Limited overlap; Non-smooth estimator; Potential outcome; Unconfoundedness.

1. INTRODUCTION

In the potential outcomes framework, there is an extensive literature on estimating causal effects based on the assumptions of unconfoundedness and overlap of the covariate distributions (Rosenbaum & Rubin, 1983; Angrist & Pischke, 2008; Imbens & Rubin, 2015). Unfortunately, it is common to have limited overlap in covariates between the treatment and control groups, which affects the credibility of all methods attempting to estimate causal effects for the population (King & Zeng, 2005; Imbens, 2015). A consequence is that extreme estimated propensity scores induce large weights, which can result in a large variance and poor finite-sample properties (Kang & Schafer, 2007; Khan & Tamer, 2010), so it may seem desirable to modify the estimand to averaging only over that part of the covariate space with treatment probabilities bounded away from 0 and 1. For example, in a medical study of a particular chemotherapy for breast cancer, because patients with stage I breast cancer have never been treated with chemotherapy, clinicians then redefine the study population to be patients with stage II to stage IV breast cancer, omitting patients with stage I breast cancer for whom the propensity scores are zero. This effectively

alters the estimand by changing the reference population to a different target population. Li et al. (2017) proposed a general representation for the target population.

Trimming observational studies based on estimated propensity scores was first used in medical applications (e.g., Vincent et al., 2002; Grzybowski et al., 2003; Kurth et al., 2005) and then formalized by Crump et al. (2009), who suggested dropping units from the analysis with estimated propensity score outside an interval $[\alpha_1, \alpha_2]$, so that the average treatment effect for the target population can be estimated with the smallest asymptotic variance. Other methods, e.g., Traskin & Small (2011) and Fogarty et al. (2016), construct the study population based on covariates themselves. But with moderate or high-dimensional covariates, these rules for discarding units become complicated. In these cases, dimension reduction, for example, seeking a scalar summary of the covariates, seems important. This was the original motivation of the propensity score (Rosenbaum & Rubin, 1983), which is auguably the most interpretable scalar function of the covariates.

Existing methods rarely incorporate the uncertainty in this design stage and restrict inference to the trimmed sample. We incorporate uncertainty in both the design and analysis stages. The non-smooth nature of trimming renders the target causal estimand not root-n estimable (Crump et al., 2009), so, instead of making a binary decision to include or exclude units from analysis, we propose to use a smooth weight function to approximate the existing sample trimming. This allows us to derive the asymptotic properties of the corresponding causal effect estimators using conventional linearization methods for two-step statistics. We show that the new weighting estimators are asymptotically linear, so the bootstrap can be used to construct confidence intervals.

2. POTENTIAL OUTCOMES, CAUSAL EFFECTS AND ASSUMPTIONS

For each unit i, the treatment is $A_i \in \{0, 1\}$, where 0 and 1 are labels for control and treatment. There are two potential outcomes, one for treatment and the other for control, denoted by $Y_i(1)$ and $Y_i(0)$, respectively. The observed outcome is $Y_i = Y_i(A_i)$. Let X_i be the observed pre-treatment covariates. We assume that $\{A_i, X_i, Y_i(1), Y_i(0)\}_{i=1}^N$ are independent draws from the distribution of $\{A, X, Y(1), Y(0)\}$. Given the observed covariates, the conditional average causal effect is $\tau(X) = E\{Y(1) - Y(0) \mid X\}$. The average treatment effect is $\tau = E\{Y(1) - Y(0)\} = E\{\tau(X)\}$. The common assumptions to identify τ are as follows (Rosenbaum & Rubin, 1983):

Assumption 1 (*Unconfoundedness*). $Y(a) \perp\!\!\!\perp A \mid X$ for $a = 0, 1$;

Assumption 2 (*Overlap*). there exist constants c_1 and c_2 such that with probability $1, 0 < c_1 \leq e(X) \leq c_2 < 1$, where $e(X) = \text{pr}(A = 1 \mid X)$ is the propensity score.

In observational studies, the propensity score is not known and therefore must be estimated from data. Following Rosenbaum & Rubin (1983) and most of the empirical literature, we assume that the propensity score is correctly specified by a generalized linear model $e(X) = e(X'\theta^*)$. We focus on $\hat{\theta}$, the maximum likelihood estimator of the true parameter θ^*, although our method is also applicable to other asymptotically linear estimators of θ^*. Then, a simple weighting estimator of τ is $N^{-1}\sum_{i=1}^N \hat{\tau}(X_i)$, where

$$\hat{\tau}(X_i) = \frac{A_i Y_i}{e(X_i'\hat{\theta})} - \frac{(1 - A_i)Y_i}{1 - e(X_i'\hat{\theta})}.$$

If we further estimate $\mu(a, X) = E(Y \mid A = a, X)$ by $\hat{\mu}(a, X)$ and obtain the residual $\hat{R}_i = Y_i - \hat{\mu}(A_i, X_i)$, then the augmented weighting estimator is $N^{-1}\sum_{i=1}^N \hat{\tau}^{\text{aug}}(X_i)$ (Lunceford &

Davidian, 2004; Bang & Robins, 2005), where

$$\hat{\tau}^{\text{aug}}(X_i) = \left\{ \frac{A_i \hat{R}_i}{e(X_i'\hat{\theta})} + \hat{\mu}(1, X_i) \right\} - \left\{ \frac{(1-A_i)\hat{R}_i}{1-e(X_i'\hat{\theta})} + \hat{\mu}(0, X_i) \right\}.$$

The augmented weighting estimator features a double robustness property in the sense that under Assumptions 1 and 2, it is consistent for τ if either $e(X)$ or $\mu(a, X)$ is correctly specified.

The weighting estimators may be variable when Assumption 2 is violated or nearly violated. When there is limited overlap, define the set with adequate overlap to be $\mathcal{O} = \{X : \alpha_1 \leq e(X) \leq \alpha_2\}$, where α_1 and α_2 are fixed cut-off values; e.g., $\alpha_1 = 0.1$ and $\alpha_2 = 0.9$ (Crump et al., 2009). The target population is then represented by $\mathcal{O}$, and the estimand of interest becomes $\tau(\mathcal{O}) = E\{\tau(X) \mid X \in \mathcal{O}\}$. The trimmed sample based on the estimated propensity score is $\hat{\mathcal{O}} = \{X : \alpha_1 \leq e(X'\hat{\theta}) \leq \alpha_2\}$. Correspondingly, the inclusion weight is

$$\omega(X_i'\hat{\theta}) = 1\{\alpha_1 \leq e(X_i'\hat{\theta}) \leq \alpha_2\}, \tag{1}$$

where $1(\cdot)$ is the indicator function, and the weighting estimators of $\tau(\mathcal{O})$ become

$$\hat{\tau} = \hat{\tau}(\hat{\theta}) = \left\{ \sum_{i=1}^{N} \omega(X_i'\hat{\theta}) \right\}^{-1} \sum_{i=1}^{N} \omega(X_i'\hat{\theta}) \hat{\tau}(X_i), \tag{2}$$

$$\hat{\tau}^{\text{aug}} = \hat{\tau}^{\text{aug}}(\hat{\theta}) = \left\{ \sum_{i=1}^{N} \omega(X_i'\hat{\theta}) \right\}^{-1} \sum_{i=1}^{N} \omega(X_i'\hat{\theta}) \hat{\tau}^{\text{aug}}(X_i). \tag{3}$$

The main question we address is how the estimated support affects the inference. To make inference for $\tau(\mathcal{O})$, we need to take into account first the sampling variability in $\hat{\theta}$, which induces variability of the estimated set $\hat{\mathcal{O}}$ and second the sampling variability in $\hat{\tau}$ and $\hat{\tau}^{\text{aug}}$. We cannot directly apply conventional asymptotic linearization methods because the weight function (1) is non-smooth, so we consider a smooth weight function

$$\omega_\epsilon(X_i'\hat{\theta}) = \Phi_\epsilon \left\{ e(X_i'\hat{\theta}) - \alpha_1 \right\} \Phi_\epsilon \left\{ \alpha_2 - e(X_i'\hat{\theta}) \right\}, \tag{4}$$

where $\Phi_\epsilon(z)$ is a normal cumulative distribution with mean zero and variance ϵ^2. The normal distribution is can be changed to any differentiable distribution whose variance increases with ϵ. As $\epsilon \to 0$, (4) converges to the indicator weight function (1). Both functions include units with non-extreme propensity scores with probability 1. In contrast, another smooth weight function, the overlap weight function $\omega\{e(X)\} = e(X)\{1 - e(X)\}$ recently proposed by Li et al. (2017), overweighs units with propensity scores close to 0.5 and thus does not target $\tau(\mathcal{O})$.

3. Main Results for the average causal effect

We derive the asymptotic results for the smooth weighting estimators. Based on data $\{(A_i, X_i)\}_{i=1}^N$, let the score function and the Fisher information matrix of θ be

$$S(\theta) = \frac{1}{N} \sum_{i=1}^{N} X_i \frac{A_i - e(X_i'\theta)}{e(X_i'\theta)\{1 - e(X_i'\theta)\}} f(X_i'\theta), \quad \mathcal{I}(\theta) = E\left[\frac{f(X'\theta)^2}{e(X'\theta)\{1 - e(X'\theta)\}} XX' \right],$$

where $f(t) = \mathrm{d}e(t)/\mathrm{d}t$. Let $\sigma^2(a, X) = \text{var}(Y \mid A = a, X)$ for $a = 0, 1$. Let $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\text{aug}}$ denote the weighting estimators (2) and (3) with the smooth weight function (4), respectively. Let $\tau_\epsilon = E\{\omega_\epsilon(X'\theta^*)\tau(X)\}$ and $\omega_\epsilon(\theta) = E\{\omega_\epsilon(X'\theta)\}$. We show that $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\text{aug}}$ are consistent for τ_ϵ.

Moreover, the discrepancy between τ_ϵ and the target estimand $\tau(\mathcal{O})$ can be made arbitrarily small by choosing a small ϵ.

THEOREM 1. *Under Assumption 1, $\hat{\tau}_\epsilon$ is asymptotically linear. Moreover,*

$$N^{1/2}(\hat{\tau}_\epsilon - \tau_\epsilon) \to \mathcal{N}\left\{0, \sigma_\epsilon^2 + b_{1,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{1,\epsilon} - b_{2,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon}\right\},$$

in distribution, as $N \to \infty$, where

$$\begin{aligned}
b_{1,\epsilon} &= E\left[\frac{\partial}{\partial\theta}\left\{\omega_\epsilon(\theta^*)^{-1}\omega_\epsilon(X'\theta^*)\right\}\tau(X)\right],\\
b_{2,\epsilon} &= \omega_\epsilon(\theta^*)^{-1}E\left\{\omega_\epsilon(X'\theta^*)f(X'\theta^*)\left[\frac{E\{X\mu(1,X)\mid e(X)\}}{e(X)} + \frac{E\{X\mu(0,X)\mid e(X)\}}{1-e(X)}\right]\right\},\\
\sigma_\epsilon^2 &= \operatorname{var}\left\{\omega_\epsilon(\theta^*)^{-1}\omega_\epsilon(X^{\mathrm{T}}\theta^*)\tau(X)\right\}\\
&\quad +\omega_\epsilon(\theta^*)^{-2}E\left\{\omega_\epsilon(X'\theta^*)^2\left[\left\{\frac{1-e(X)}{e(X)}\right\}^{1/2}\mu(1,X) + \left\{\frac{e(X)}{1-e(X)}\right\}^{1/2}\mu(0,X)\right]^2\right\}\\
&\quad +\omega_\epsilon(\theta^*)^{-2}E\left[\omega_\epsilon(X'\theta^*)^2\left\{\frac{\sigma^2(1,X)}{e(X)} + \frac{\sigma^2(0,X)}{1-e(X)}\right\}\right].
\end{aligned}$$

Remark 1. We show in the Supplementary Material that $b_{1,\epsilon} \to 0$, as $\epsilon \to 0$. Therefore, the increased variability due to estimating the support, $b_{1,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{1,\epsilon}$, is close to 0 with a small ϵ.

Remark 2. The term $-b_{2,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon}$ implies that the estimated propensity score increases the precision of the simple weighting estimator of τ based on the true propensity score, a phenomenon that has previously appeared in the causal inference literature (e.g., Rubin & Thomas, 1992; Hahn, 1998; Abadie & Imbens, 2016).

THEOREM 2. *Under Assumption 1, $\hat{\tau}_\epsilon^{\mathrm{aug}}$ is asymptotically linear. Moreover,*

$$N^{1/2}(\hat{\tau}_\epsilon^{\mathrm{aug}} - \tau_\epsilon) \to \mathcal{N}\left\{0, \tilde{\sigma}_\epsilon^2 + b_{1\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{1\epsilon} + (C_0+C_1)'\mathcal{I}(\theta^*)^{-1}(C_0+C_1) + \tilde{B}'(C_0-C_1)\right\},$$

in distribution, as $N \to \infty$, where $b_{1,\epsilon}$ is defined in Theorem 1,

$$\begin{aligned}
\tilde{\sigma}_\epsilon^2 &= \operatorname{var}\left\{\omega_\epsilon(\theta^*)^{-1}\omega_\epsilon(X^{\mathrm{T}}\theta^*)\tau(X)\right\} + \omega_\epsilon(\theta^*)^{-2}E\left[\omega_\epsilon(X'\theta^*)^2\left\{\frac{\sigma^2(1,X)}{e(X)} + \frac{\sigma^2(0,X)}{1-e(X)}\right\}\right],\\
C_a &= E\left\{X\omega_\epsilon(X'\theta^*)f(X'\theta^*)\frac{\tilde{\mu}(a,X)-\mu(a,X)}{\mathrm{pr}(A=a\mid X)}\right\} \quad (a=0,1),
\end{aligned}$$

with $\hat{\mu}(a,X) \to \tilde{\mu}(a,X)$ in probability, for $a = 0, 1$, and $\tilde{B} = b_{1,\epsilon} - C_0 - C_1$.

Remark 3. If the outcome model is correctly specified, then $\tilde{\mu}(a,X) = \mu(a,X)$ and thus $C_0 = C_1 = 0$. Consequently, the asymptotic variance of $\hat{\tau}_\epsilon^{\mathrm{aug}}$ reduces to $\tilde{\sigma}_\epsilon^2 + b_{1\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{1\epsilon}$, which is smaller than the asymptotic variance of $\hat{\tau}_\epsilon$. Intuitively, by regressing Y on X and A, we use the residual as the new outcome, which in general has a smaller variance than Y.

Remark 4. Because $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\mathrm{aug}}$ are asymptotically linear, the bootstrap can be used to estimate the variances of $\hat{\tau}_\epsilon$ and $\hat{\tau}_\epsilon^{\mathrm{aug}}$ (Shao & Tu, 2012). We evaluate the finite sample properties of the bootstrap variance estimator by simulation in the Supplementary Material. Let $\mathcal{S} = \{X : e(X'\theta^*) = \alpha_1 \text{ or } \alpha_2\}$. We also show that if $\mathrm{pr}(X \in \mathcal{S}) = 0$, the bootstrap works for the weighting estimator with the indicator function, which is confirmed in the simulation study.

Remark 5. Although some robust nonparametric methods (Hirano et al., 2003; Lee et al., 2010) can be used for propensity score estimation, the majority of the literature uses parametric generalized linear models. When the propensity score model is misspecified, the weighting estimators are not consistent for the causal effect defined on the target population $\mathcal{O} = \{X : \alpha_1 \leq e(X) \leq \alpha_2\}$. However, our estimators can still be helpful to inform treatment effects for the population defined as $\mathcal{O}^* = \{X : \alpha_1 \leq e(X'\theta^*) \leq \alpha_2\}$, where $e(X'\theta^*)$ is the propensity score projected to the generalized linear model family. This new study population is defined as between two hyperplanes of the covariate space, which is slightly more complicated than the study population defined by the trees in Traskin & Small (2011) or by the intervals of covariates in Fogarty et al. (2016). Moreover, the smooth weighting estimators are still asymptotically linear, and again the bootstrap can be used for constructing confidence intervals. See the Supplementary Material for more details.

Remark 6. An important issue regarding the smooth weight function is the choice of ϵ, which involves a bias-variance trade-off. On the one hand, the discrepancy between τ_ϵ and the target parameter $\tau(\mathcal{O})$ is $E\{[\omega_\epsilon(X'\theta^*) - 1\{\alpha_1 \leq e(X'\theta^*) \leq \alpha_2\}]\tau(X)\}$. Assuming that $\tau(X)$ is integrable, by the dominated convergence theorem, τ_ϵ converges to $\tau(\mathcal{O})$ as $\epsilon \to 0$. This implies that based on $\hat{\tau}_\epsilon$ or $\hat{\tau}_\epsilon^{\text{aug}}$, we can draw inference for $\tau(\mathcal{O})$ by choosing a small ϵ. On the other hand, as $\epsilon \to 0$, the smooth weight function (4) becomes closer to the indicator weight function (1), which increases the variance of the weighting estimators. In practice, we recommend a sensitivity analysis varying ϵ over a grid, for example, $10^{-4}, 10^{-5}, \ldots$, illustrated in the Supplementary Material and the application.

4. The National Health and Nutrition Examination Survey Data

We examine a data set from the 2007–2008 U.S. National Health and Nutrition Examination Survey to estimate the causal effect of smoking on blood lead levels (Hsu & Small, 2013). The data set includes 3340 subjects consisting of 679 smokers, denoted as $A = 1$, and 2661 non-smokers, denoted as $A = 0$. The outcome variable Y is the measured lead level in blood, with the observed range from 0.18 ug/dl to 33.10 ug/dl. The covariates are age, income-to-poverty level, gender, education and race.

The propensity score is estimated by a logistic regression model with linear predictors including all covariates. To help address lack of overlap, for the average smoking effect, because there is little overlap for the propensity score less than 0.05 and greater than 0.6, we restrict our estimand to the target population $\mathcal{O} = \{X : 0.05 \leq e(X) \leq 0.6\}$. The truncation of the propensity score at 0.6 is because there are only few subjects with propensity score above 0.6. This removes 794 subjects, with 111 smokers and 683 non-smokers. Thus, the analysis sample includes 2546 subjects, with 568 smokers and 1978 non-smokers. In the Supplementary Material, we display the summary statistics of the covariates, and give more detailed interpretation of the target population.

We consider the weighting estimators using both the indicator and smooth weight functions with $\epsilon = 10^{-4}$ and $\epsilon = 10^{-5}$. For the augmented weighting estimator, we use a linear outcome model adjusting for all covariates, separately for $A = 0, 1$. Table 1 shows the results. The weighting estimators with the smooth weight function are close to the counterparts with the indicator weight function, but have slightly smaller standard errors. The smooth weighting estimators are insensitive to the choice of ϵ. From the results, on average, smoking increases the lead level in blood at least by 0.65 ug/dl over the target population with $0.05 \leq e(X) \leq 0.6$.

Table 1. *Estimate, standard error based on* 100 *bootstrap replicates, and* 95% *confidence interval*

	ϵ	estimate	s.e.	95% c.i.		estimate	s.e.	95% c.i.
$\hat{\tau}(\hat{\theta})$	–	0.646	0.135	$(0.376, 0.916)$	$\hat{\tau}^{\text{aug}}(\hat{\theta})$	0.765	0.107	$(0.552, 0.978)$
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-4}	0.661	0.124	$(0.412, 0.909)$	$\hat{\tau}^{\text{aug}}_\epsilon(\hat{\theta})$	0.763	0.105	$(0.554, 0.973)$
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-5}	0.632	0.133	$(0.366, 0.899)$	$\hat{\tau}^{\text{aug}}_\epsilon(\hat{\theta})$	0.754	0.105	$(0.543, 0.964)$

5. EXTENSION TO THE AVERAGE TREATMENT EFFECT ON THE TREATED

Another estimand of interest is the average treatment effect for the treated $\tau_{\text{ATT}} = E\{Y(1) - Y(0) \mid A = 1\} = E\{\tau(X) \mid A = 1\}$. Similar to Crump et al. (2009), if $\sigma^2(1, X) = \sigma^2(0, X)$, we can show that the optimal overlap for estimating τ_{ATT} is of the form $\mathcal{O} = \{X : 1 - e(X) \geq \alpha\}$ for some α, for which the estimators have the smallest asymptotic variance. Intuitively, for the treated units with $e(X)$ close to 1, there are few similar units in the control group that can provide information to infer their $Y(0)$'s. Therefore, it is reasonable to drop these units with $e(X)$ close to 1 when inferring τ_{ATT}. We give a formal discussion in the Supplementary Material.

By restricting to the subpopulation $\mathcal{O} = \{X : 1 - e(X) \geq \alpha\}$, the estimand of interest becomes $\tau_{\text{ATT}}(\mathcal{O}) = E\{\tau(X) \mid A = 1, X \in \mathcal{O}\}$. We propose two estimators with smooth inclusion weights $\omega_{\text{ATT},\epsilon}(X'\hat{\theta}) = \Phi_\epsilon\{1 - \alpha - e(X_i'\hat{\theta})\}e(X_i'\hat{\theta})$:

$$\hat{\tau}_{\text{ATT},\epsilon} = \frac{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X'\hat{\theta})\hat{\tau}(X_i)}{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X'\hat{\theta})}, \quad \hat{\tau}^{\text{aug}}_{\text{ATT},\epsilon} = \frac{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X'\hat{\theta})\hat{\tau}^{\text{aug}}(X_i)}{\sum_{i=1}^N \omega_{\text{ATT},\epsilon}(X'\hat{\theta})},$$

which are (2) and (3) with $\omega_\epsilon(X'\hat{\theta})$ replaced by $\omega_{\text{ATT},\epsilon}(X'\hat{\theta})$. Even without sample trimming, the augmented weighting estimator is different from the existing estimators in the literature (e.g., Mercatanti & Li, 2014; Shinozaki & Matsuyama, 2015; Zhao & Percival, 2017). We provide the motivation in the Supplementary Material. The asymptotic properties for $\hat{\tau}_{\text{ATT},\epsilon}$ and $\hat{\tau}^{\text{aug}}_{\text{ATT},\epsilon}$ can be derived similarly as in Theorems 1 and 2. In particular, the asymptotic linearity of these two estimators enables the bootstrap for inference.

Define $\tilde{b}_{1,\epsilon}$ and $\tilde{b}_{2,\epsilon}$ as the analogs of $b_{1,\epsilon}$ and $b_{2,\epsilon}$ with weights $\omega_{\text{ATT},\epsilon}(X'\hat{\theta})$. In contrast to Remark 1, for τ_{ATT}, the term $\tilde{b}_{1,\epsilon}$ does not converge to 0 as $\epsilon \to 0$. The correction term in the asymptotic variance formula due to the estimated propensity score instead of the true propensity score, $\tilde{b}_{1,\epsilon}'\mathcal{I}(\theta^*)^{-1}\tilde{b}_{1,\epsilon} - \tilde{b}_{2,\epsilon}'\mathcal{I}(\theta^*)^{-1}\tilde{b}_{2,\epsilon}$, can be negative, zero, or positive. Ignoring the uncertainty in the estimated propensity score, the inference can be either conservative or anti-conservative for τ_{ATT}, which differs from the inference for τ. This fundamental difference also appeared for matching estimators (Abadie & Imbens, 2016), which highlights the importance of incorporating the uncertainty in the design stage especially for τ_{ATT}.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs, a simulation study, an extension, and more details for the application.

ACKNOWLEDGMENTS

We benefited from the insightful comments from the Associate Editor and two reviewers. Peng Ding is partially supported by the U.S. Institute of Education Sciences and National Science Foundation.

REFERENCES

ABADIE, A. & IMBENS, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84**, 781–807.

ANGRIST, J. D. & PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

BANG, H. & ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. & MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–199.

FOGARTY, C. B., MIKKELSEN, M. E., GAIESKI, D. F. & SMALL, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Am. Stat. Assoc.* **111**, 447–458.

GRZYBOWSKI, M., CLEMENTS, E. A., PARSONS, L., WELCH, R., TINTINALLI, A. T., ROSS, M. A. & ZALENSKI, R. J. (2003). Mortality benefit of immediate revascularization of acute ST-segment elevation myocardial infarction in patients with contraindications to thrombolytic therapy: a propensity analysis. *The Journal of the American Medical Association* **290**, 1891–1898.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

HIRANO, K., IMBENS, G. W. & RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

HSU, J. Y. & SMALL, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69**, 803–811.

IMBENS, G. W. (2015). Matching methods in practice: Three examples. *J. Hum. Resour.* **50**, 373–419.

IMBENS, G. W. & RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge UK: Cambridge University Press.

KANG, J. D. & SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22**, 523–539.

KHAN, S. & TAMER, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica* **78**, 2021–2042.

KING, G. & ZENG, L. (2005). The dangers of extreme counterfactuals. *Political Analysis* **14**, 131–159.

KURTH, T., WALKER, A. M., GLYNN, R. J., CHAN, K. A., GAZIANO, J. M., BERGER, K. & ROBINS, J. M. (2005). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am. J. Epidemiol.* **163**, 262–270.

LEE, B. K., LESSLER, J. & STUART, E. A. (2010). Improving propensity score weighting using machine learning. *Stat. Med.* **29**, 337–346.

LI, F., MORGAN, K. L. & ZASLAVSKY, A. M. (2017). Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* , DOI: 10.1080/01621459.2016.1260466.

LUNCEFORD, J. K. & DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* **23**, 2937–2960.

MERCATANTI, A. & LI, F. (2014). Do debit cards increase household spending? evidence from a semiparametric causal analysis of a survey. *The Annals of Applied Statistics* **8**, 2485–2508.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

RUBIN, D. B. & THOMAS, N. (1992). Affinely invariant matching methods with ellipsoidal distributions. *Ann. Statist.* **20**, 1079–1093.

SHAO, J. & TU, D. (2012). *The Jackknife and Bootstrap*. New York: Springer.

SHINOZAKI, T. & MATSUYAMA, Y. (2015). Doubly robust estimation of standardized risk difference and ratio in the exposed population. *Epidemiology* **26**, 873–877.

TRASKIN, M. & SMALL, D. S. (2011). Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences* **3**, 94–118.

VINCENT, J. L., BARON, J.-F., REINHART, K., GATTINONI, L., THIJS, L., WEBB, A., MEIER-HELLMANN, A., NOLLET, G. & PERES-BOTA, D. (2002). Anemia and blood transfusion in critically ill patients. *The Journal of the American Medical Association* **288**, 1499–1507.

ZHAO, Q. & PERCIVAL, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference* **5**, DOI: https://doi.org/10.1515/jci–2016–0010.

Biometrika (2018), **105**, 2, *pp*. 1–14
doi: doi: 10.1093/biomet/asy008
Advance Access publication on 12 March 2018

Supplementary material for "Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores"

BY S. YANG

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

syang24@ncsu.edu

P. DING

Department of Statistics, University of California, Berkeley, California 94720, U.S.A.

pengdingpku@berkeley.edu

§S1 gives all the proofs, §S2 presents a simulation study, §S3 extends the theory to the average treatment effect on the treated, and §S4 provides more detailed analysis of the National Health and Nutrition Examination Survey Data.

Below we use $C \cong D$ for $C = D + O_p(N^{-1/2})$. Because $\hat{\theta}$ is the solution to the score equation $S(\theta) = 0$, under certain regularity conditions, $\hat{\theta} - \theta^* = \mathcal{J}(\theta^*)^{-1} S(\theta^*) + o_p(N^{-1/2})$, where $\mathcal{J}(\theta^*) = E\{\partial S(\theta^*)/\partial \theta'\}$ (e.g., van der Vaart, 2000). When the propensity model is correctly specified, then $\mathcal{J}(\theta^*) = \mathcal{I}(\theta^*)$; when the propensity score model is misspecified, $\mathcal{J}(\theta^*)$ is not necessarily equal to $\mathcal{I}(\theta^*)$.

S1. PROOFS

S1·1. *Proof of Theorem 1*

We write

$$\begin{aligned}
\hat{\tau}_\epsilon &= \hat{\tau}_\epsilon(\hat{\theta}) \\
&\cong \hat{\tau}_\epsilon(\theta^*) + E\left\{\frac{\partial \hat{\tau}_\epsilon(\theta^*)}{\partial \theta'}\right\}(\hat{\theta} - \theta^*) \qquad \text{(S1)} \\
&\cong \frac{1}{N}\sum_{i=1}^{N} \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left\{\frac{A_i Y_i}{e(X_i'\theta^*)} - \frac{(1-A_i)Y_i}{1-e(X_i'\theta^*)}\right\} + E\left\{\frac{\partial \hat{\tau}_\epsilon(\theta^*)}{\partial \theta'}\right\}\mathcal{I}(\theta^*)^{-1} S(\theta^*) \qquad \text{(S2)} \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left\{\frac{A_i Y_i}{e(X_i'\theta^*)} - \frac{(1-A_i)Y_i}{1-e(X_i'\theta^*)}\right\} \\
&\quad + B' \frac{1}{N}\sum_{i=1}^{N} X_i \frac{A_i - e(X_i'\theta^*)}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}} f(X_i'\theta^*),
\end{aligned}$$

where (S1) follows from the Taylor expansion, (S2) follows from $\hat{\theta} - \theta^* \cong \mathcal{I}(\theta^*)^{-1} S(\theta^*)$ and

$$B' = E\left\{\frac{\partial \hat{\tau}_\epsilon(\theta^*)}{\partial \theta'}\right\}\mathcal{I}(\theta^*)^{-1}. \qquad \text{(S3)}$$

Therefore, the asymptotic linearity of $\hat{\tau}_\epsilon$ follows. Moreover,

$$\begin{aligned}
N^{1/2}(\hat{\tau}_\epsilon - \tau_\epsilon) \cong\ & N^{-1/2}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\left[\frac{A_i\{Y_i-\mu(A_i,X_i)\}}{e(X_i'\theta^*)} - \frac{(1-A_i)\{Y_i-\mu(A_i,X_i)\}}{1-e(X_i'\theta^*)}\right] \\
& + N^{-1/2}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\left(\frac{\{A_i-e(X_i'\theta^*)\}[\mu(A_i,X_i)-\mu\{A_i,e(X_i'\theta^*)\}]}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}\right) \\
& + N^{-1/2}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\left[\frac{\{A_i-e(X_i'\theta^*)\}\mu\{A_i,e(X_i'\theta^*)\}}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}} - \tau\{e(X_i'\theta^*)\}\right] \\
& + N^{-1/2}\sum_{i=1}^{N}\left[\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\tau\{e(X_i'\theta^*)\} - \tau_\epsilon\right] \\
& + N^{-1/2}\sum_{i=1}^{N}B'X_i\frac{A_i-e(X_i'\theta^*)}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}f(X_i'\theta^*), \\
=\ & T_0+T_1+T_2+T_3,
\end{aligned}$$

where $\tau\{e(X'\theta^*)\} = E\{Y(1)-Y(0)\mid e(X'\theta^*)\}$, and by grouping different terms,

$$\begin{aligned}
T_0 =\ & N^{-1/2}\sum_{i=1}^{N}\left[\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\tau\{e(X_i'\theta^*)\} - \tau_\epsilon\right], \\
T_1 =\ & N^{-1/2}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\left[\frac{\{A_i-e(X_i'\theta^*)\}\mu\{A_i,e(X_i'\theta^*)\}}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}} - \tau\{e(X_i'\theta^*)\}\right] \\
& + N^{-1/2}\sum_{i=1}^{N}B'E\{X_i\mid e(X_i'\theta^*)\}\frac{A_i-e(X_i'\theta)}{e(X_i'\theta)\{1-e(X_i'\theta)\}}f(X_i'\theta^*), \\
T_2 =\ & N^{-1/2}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\left(\frac{\{A_i-e(X_i'\theta^*)\}[\mu(A_i,X_i)-\mu\{A_i,e(X_i'\theta^*)\}]}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}\right) \\
& + N^{-1/2}\sum_{i=1}^{N}B'[X_i-E\{X_i\mid e(X_i'\theta^*)\}]\frac{A_i-e(X_i'\theta^*)}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}f(X_i'\theta^*),
\end{aligned}$$

and

$$T_3 = N^{-1/2}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\left[\frac{A_i\{Y_i-\mu(A_i,X_i)\}}{e(X_i'\theta^*)} - \frac{(1-A_i)\{Y_i-\mu(A_i,X_i)\}}{1-e(X_i'\theta^*)}\right]. \quad \text{(S4)}$$

Define

$$\begin{aligned}
\mathcal{F}_0 = \left\{X_1'\theta^*,\ldots,X_N'\theta^*\right\},\ \mathcal{F}_1 &= \left\{A_1,\ldots,A_N,X_1'\theta^*,\ldots,X_N'\theta^*\right\}, \\
\mathcal{F}_2 &= \left\{A_1,\ldots,A_N,X_1'\theta^*,\ldots,X_N'\theta^*,X_1,\ldots,X_N\right\}.
\end{aligned}$$

By conditioning arguments, $E(T_0)=0$, for $k=1,\ldots,3$, $E(T_k)=E\{E(T_k\mid\mathcal{F}_{k-1})\}=0$, and for $k=1,\ldots,3$,

$$\begin{aligned}
\operatorname{cov}(T_0,T_k) &= \operatorname{cov}\{E(T_0\mid\mathcal{F}_0),E(T_k\mid\mathcal{F}_0)\} + E\{\operatorname{cov}(T_0,T_k\mid\mathcal{F}_0)\} \\
&= \operatorname{cov}\{E(T_0\mid\mathcal{F}_0),0\} + E\{0\} = 0,
\end{aligned}$$

for $k = 2, 3$,

$$\begin{aligned}\operatorname{cov}(T_1, T_k) &= \operatorname{cov}\{E(T_1 \mid \mathcal{F}_1), E(T_k \mid \mathcal{F}_1)\} + E\{\operatorname{cov}(T_1, T_k \mid \mathcal{F}_1)\} \\ &= \operatorname{cov}\{E(T_1 \mid \mathcal{F}_1), 0\} + E\{0\} = 0,\end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}(T_2, T_3) &= \operatorname{cov}\{E(T_2 \mid \mathcal{F}_2), E(T_3 \mid \mathcal{F}_2)\} + E\{\operatorname{cov}(T_2, T_3 \mid \mathcal{F}_2)\} \\ &= \operatorname{cov}\{E(T_2 \mid \mathcal{F}_2), 0\} + E\{0\} = 0.\end{aligned}$$

Also, we calculate the variances of T_i, for $i = 0, \ldots, 3$, as follows. For T_0,

$$\operatorname{var}(T_0) = E(T_0^2) = \operatorname{var}\left[\frac{\omega_\epsilon(X'\theta^*)\tau\{e(X'\theta^*)\}}{E\{\omega_\epsilon(X'\theta^*)\}}\right]$$

For T_1,

$$\begin{aligned}\operatorname{var}(T_1) &= E\{\operatorname{var}(T_1 \mid \mathcal{F}_0)\} = E\{E(T_1^2 \mid \mathcal{F}_0)\} \\ &= \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2} E\{\omega_\epsilon(X'\theta^*)^2 \left[\left\{\frac{1-e(X'\theta^*)}{e(X'\theta^*)}\right\}^{1/2} \mu\{1, e(X'\theta^*)\}\right. \\ &\quad \left. + \left\{\frac{e(X'\theta^*)}{1-e(X'\theta^*)}\right\}^{1/2} \mu\{0, e(X'\theta^*)\}\right]^2 \\ &\quad + 2\frac{1}{E\{\omega_\epsilon(X'\theta^*)\}} B' E\{\omega_\epsilon(X'\theta^*) E\{X \mid e(X'\theta^*)\} \\ &\quad \times \left[\frac{\mu\{1, e(X'\theta^*)\}}{e(X'\theta^*)} + \frac{\mu\{0, e(X'\theta^*)\}}{1-e(X'\theta^*)}\right] f(X'\theta^*)\} \\ &\quad + B' E\left[f(X'\theta^*)^2 \frac{E\{X \mid e(X'\theta^*)\} E\{X' \mid e(X'\theta^*)\}}{e(X'\theta^*)\{1-e(X'\theta^*)\}}\right] B.\end{aligned}$$

For T_2,

$$\begin{aligned}\operatorname{var}(T_2) &= E\{\operatorname{var}(T_2 \mid \mathcal{F}_1)\} = E\{E(T_2^2 \mid \mathcal{F}_1)\} \\ &= \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2} E\left\{\omega_\epsilon(X'\theta^*)^2 \left[\frac{\sigma^2\{1, e(X'\theta^*)\}}{e(X'\theta^*)} + \frac{\sigma^2\{0, e(X'\theta^*)\}}{1-e(X'\theta^*)}\right]\right\} \\ &\quad + 2\frac{1}{E\{\omega_\epsilon(X'\theta^*)\}} B' E\left\{\omega_\epsilon(X'\theta^*) f(X'\theta^*) \left[\frac{\operatorname{cov}\{X, \mu(1, X) \mid e(X'\theta^*)\}}{e(X'\theta^*)}\right.\right. \\ &\quad \left.\left. + \frac{\operatorname{cov}\{X, \mu(0, X) \mid e(X'\theta^*)\}}{1-e(X'\theta^*)}\right]\right\} \\ &\quad + B' E\left[f(X'\theta^*)^2 \frac{\operatorname{var}\{X \mid e(X'\theta^*)\}}{e(X'\theta^*)\{1-e(X'\theta^*)\}}\right] B.\end{aligned}$$

For T_3,

$$\begin{aligned}\operatorname{var}(T_3) &= E\{\operatorname{var}(T_3 \mid \mathcal{F}_2)\} = E\{E(T_3^2 \mid \mathcal{F}_2)\} \\ &\cong \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2} E\left[\omega_\epsilon(X'\theta^*)^2 \left\{\frac{\sigma_1^2(X)}{e(X'\theta^*)} + \frac{\sigma_0^2(X)}{1-e(X'\theta^*)}\right\}\right].\end{aligned}$$

Because

$$\frac{\partial\hat{\tau}_\epsilon(\theta^*)}{\partial\theta'} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}\left[\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\right]\left\{\frac{A_iY_i}{e(X_i'\theta^*)} - \frac{(1-A_i)Y_i}{1-e(X_i'\theta^*)}\right\}$$
$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\left[\frac{A_iY_i}{e(X_i'\theta^*)^2} + \frac{(1-A_i)Y_i}{\{1-e(X_i'\theta^*)\}^2}\right]f(X_i'\theta^*)X_i,$$

we have

$$E\left\{\frac{\partial\hat{\tau}_\epsilon(\theta^*)}{\partial\theta}\right\} = E\left(\frac{\partial}{\partial\theta}\left[\frac{\omega_\epsilon(X'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\right]\tau(X)\right) - \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}}E\left\{\omega_\epsilon(X'\theta^*)f(X'\theta^*)\right.$$
$$\left.\times\left[\frac{E\{X,\mu(1,X)\mid e(X'\theta^*)\}}{e(X'\theta^*)} + \frac{E\{X,\mu(0,X)\mid e(X'\theta^*)\}}{1-e(X'\theta^*)}\right]\right\}$$
$$= b_{1,\epsilon} - b_{2,\epsilon},$$

where $b_{1,\epsilon}$ and $b_{2,\epsilon}$ are defined in Theorem 1. Therefore, according to (S3), $B = (b_{1,\epsilon} - b_{2,\epsilon})'\mathcal{I}(\theta^*)^{-1}$. As a result,

$$\begin{aligned}
&\mathrm{var}(T_0) + \mathrm{var}(T_1) + \mathrm{var}(T_2) + \mathrm{var}(T_3)\\
&= \mathrm{var}\left[\frac{\omega_\epsilon(X'\theta^*)\tau\{e(X'\theta^*)\}}{E\{\omega_\epsilon(X'\theta^*)\}}\right] \qquad \text{(S5)}\\
&\quad + \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2}E\left\{\omega_\epsilon(X'\theta^*)^2\left[\left\{\frac{1-e(X'\theta^*)}{e(X'\theta^*)}\right\}^{1/2}\mu\{1,e(X'\theta^*)\}\right.\right.\\
&\quad \left.\left. + \left\{\frac{e(X'\theta^*)}{1-e(X'\theta^*)}\right\}^{1/2}\mu\{0,e(X'\theta^*)\}\right]^2\right\}\\
&\quad + \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2}E\left\{\omega_\epsilon(X'\theta^*)^2\left[\frac{\sigma^2\{1,e(X'\theta^*)\}}{e(X'\theta^*)} + \frac{\sigma^2\{0,e(X'\theta^*)\}}{1-e(X'\theta^*)}\right]\right\}\\
&\quad + \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2}E\left[\omega_\epsilon(X'\theta^*)^2\left\{\frac{\sigma^2(1,X)}{e(X'\theta^*)} + \frac{\sigma^2(0,X)}{1-e(X'\theta^*)}\right\}\right] \qquad \text{(S6)}\\
&\quad + 2\frac{1}{E\{\omega_\epsilon(X'\theta^*)\}}B'E\left\{\omega_\epsilon(X'\theta^*)f(X'\theta^*)\left[\frac{E\{X\mu(1,X)\mid e(X'\theta^*)\}}{e(X'\theta^*)}\right.\right.\\
&\quad \left.\left. + \frac{E\{X\mu(0,X)\mid e(X'\theta^*)\}}{1-e(X'\theta^*)}\right]\right\} + B'\mathcal{I}(\theta^*)B\\
&= \sigma_\epsilon^2 + b_{1,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{1,\epsilon} - b_{2,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon}, \qquad \text{(S7)}
\end{aligned}$$

where σ_ϵ^2 is defined as the sum of terms in (S5) to (S6), and (S7) follows by plugging the expression of B,

$$\begin{aligned}
2B'b_{2,\epsilon} + B'\mathcal{I}(\theta^*)B &= 2b_{1,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon} - 2b_{2,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon} + (b_{1,\epsilon}+b_{2,\epsilon})'\mathcal{I}(\theta^*)^{-1}(b_{1,\epsilon}+b_{2,\epsilon})\\
&= b_{1,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{1,\epsilon} - b_{2,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon}.
\end{aligned}$$

Moreover, σ_ϵ^2 can be further simplified as

$$\begin{aligned}\sigma_\epsilon^2 = \text{var}&\left[\frac{\omega_\epsilon(X'\theta^*)\tau(X)}{E\{\omega_\epsilon(X'\theta^*)\}}\right] \\ &+\frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2}E\left\{\omega_\epsilon(X'\theta^*)^2\left[\left\{\frac{1-e(X'\theta^*)}{e(X'\theta^*)}\right\}^{1/2}\mu(1,X)^2\right.\right. \\ &\left.\left.+\left\{\frac{e(X'\theta^*)}{1-e(X'\theta^*)}\right\}^{1/2}\mu(0,X)\right]^2\right\} \qquad \text{(S8)} \\ &+\frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2}E\left[\omega_\epsilon(X'\theta^*)^2\left\{\frac{\sigma^2(1,X)}{e(X'\theta^*)}+\frac{\sigma^2(0,X)}{1-e(X'\theta^*)}\right\}\right]. \qquad \text{(S9)}\end{aligned}$$

Finally, the Central Limit Theorem implies

$$N^{1/2}(\hat{\tau}_\epsilon - \tau_\epsilon) \to \mathcal{N}\left\{0, \sigma_\epsilon^2 + b_{1,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{1,\epsilon} - b_{2,\epsilon}'\mathcal{I}(\theta^*)^{-1}b_{2,\epsilon}\right\},$$

in distribution, as $N \to \infty$.

S1·2. *Proof of Theorem 2*

First, $\hat{\tau}^{\text{aug}}(X_i)$ can also be written as

$$\begin{aligned}\hat{\tau}^{\text{aug}}(X_i) = &\left[\frac{A_iY_i}{e(X_i'\hat{\theta})}+\left\{1-\frac{A_i}{e(X_i'\hat{\theta})}\right\}\hat{\mu}(1,X_i)\right] \\ &-\left[\frac{(1-A_i)Y_i}{1-e(X_i'\hat{\theta})}+\left\{1-\frac{1-A_i}{1-e(X_i'\hat{\theta})}\right\}\hat{\mu}(0,X_i)\right].\end{aligned}$$

Let $\hat{\mu}(A_i, X_i)$ converge to $\tilde{\mu}(A_i, X_i)$ as $N \to \infty$. If the model for $\mu(A_i, X_i)$ is correctly specified, $\tilde{\mu}(A_i, X_i) = \mu(A_i, X_i)$.

Write

$$\begin{aligned}\hat{\tau}_\epsilon^{\text{aug}} = \hat{\tau}_\epsilon^{\text{aug}}(\hat{\theta}) &\cong \hat{\tau}_\epsilon^{\text{aug}}(\theta^*) + E\left\{\frac{\partial\hat{\tau}_\epsilon^{\text{aug}}(\theta^*)}{\partial\theta'}\right\}(\hat{\theta}-\theta^*) \\ &\cong \frac{1}{N}\sum_{i=1}^N \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\hat{\tau}^{\text{aug}}(X_i) + E\left\{\frac{\partial\hat{\tau}_\epsilon^{\text{aug}}(\theta^*)}{\partial\theta'}\right\}\mathcal{I}(\theta^*)^{-1}S(\theta^*) \\ &\cong \frac{1}{N}\sum_{i=1}^N \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left[\frac{A_iY_i}{e(X_i)}+\left\{1-\frac{A_i}{e(X_i)}\right\}\tilde{\mu}(1,X_i)\right] \\ &\quad -\frac{1}{N}\sum_{i=1}^N \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left[\frac{(1-A_i)Y_i}{1-e(X_i)}+\left\{1-\frac{1-A_i}{1-e(X_i)}\right\}\tilde{\mu}(0,X_i)\right] \\ &\quad +\tilde{B}'\frac{1}{N}\sum_{i=1}^N X_i\frac{A_i-e(X_i'\theta^*)}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}f(X_i'\theta^*),\end{aligned}$$

where

$$\tilde{B}' = E\left\{\frac{\partial\hat{\tau}_\epsilon^{\text{aug}}(\theta^*)}{\partial\theta'}\right\}\mathcal{I}(\theta^*)^{-1}. \qquad \text{(S10)}$$

Therefore, the asymptotic linearity of $\hat{\tau}_\epsilon^{\text{aug}}$ follows. Moreover,

$$
\begin{aligned}
& N^{1/2}(\hat{\tau}_\epsilon^{\text{aug}} - \tau_\epsilon) \\
\cong\ & N^{-1/2}\sum_{i=1}^{N} \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}} \left[\frac{A_i\{Y_i - \mu(A_i, X_i)\}}{e(X_i'\theta^*)} - \frac{(1-A_i)\{Y_i - \mu(A_i, X_i)\}}{1 - e(X_i'\theta^*)}\right] \\
& + N^{-1/2}\sum_{i=1}^{N} \left[\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\tau(X_i) - \tau_\epsilon\right] \\
& + N^{-1/2}\sum_{i=1}^{N} \tilde{B}' X_i \frac{A_i - e(X_i'\theta^*)}{e(X_i'\theta^*)\{1 - e(X_i'\theta^*)\}} f(X_i'\theta^*), \\
& + N^{-1/2}\sum_{i=1}^{N} \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}} \left\{1 - \frac{A_i}{e(X_i)}\right\} \{\tilde{\mu}(1, X_i) - \mu(1, X_i)\} \\
& + N^{-1/2}\sum_{i=1}^{N} \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}} \left\{1 - \frac{1-A_i}{1 - e(X_i)}\right\} \{\tilde{\mu}(0, X_i) - \mu(0, X_i)\} \\
=\ & \tilde{T}_3 + \tilde{T}_0 + \tilde{T}_1 + \tilde{T}_2,
\end{aligned}
$$

where $\tilde{T}_3 = T_3$ is defined in (S4),

$$
\begin{aligned}
\tilde{T}_0 &= N^{-1/2}\sum_{i=1}^{N} \left[\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}}\tau(X_i) - \tau_\epsilon\right], \\
\tilde{T}_1 &= N^{-1/2}\sum_{i=1}^{N} \tilde{B}' X_i \frac{A_i - e(X_i'\theta)}{e(X_i'\theta)\{1 - e(X_i'\theta)\}} f(X_i'\theta^*), \\
\tilde{T}_2 &= N^{-1/2}\sum_{i=1}^{N} \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}} \left\{1 - \frac{A_i}{e(X_i)}\right\} \{\tilde{\mu}(1, X_i) - \mu(1, X_i)\} \\
&\quad + N^{-1/2}\sum_{i=1}^{N} \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}} \left\{1 - \frac{1-A_i}{1 - e(X_i)}\right\} \{\tilde{\mu}(0, X_i) - \mu(0, X_i)\}.
\end{aligned}
$$

By the same argument as in the proof of Theorem 1, $E(\tilde{T}_j) = 0$, for $j = 0, \ldots, 3$, and $\text{cov}(\tilde{T}_j, \tilde{T}_k) = 0$ for all $j \neq k$ expect $\text{cov}(\tilde{T}_1, \tilde{T}_2)$. Moreover,

$$
\begin{aligned}
& \text{var}(\tilde{T}_3) + \text{var}(\tilde{T}_0) + \text{var}(\tilde{T}_1) + \text{var}(\tilde{T}_2) + 2\text{cov}(\tilde{T}_1, \tilde{T}_2) \\
=\ & \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2} E\left[\omega_\epsilon(X'\theta^*)^2 \left\{\frac{\sigma^2(1, X)}{e(X'\theta^*)} + \frac{\sigma^2(0, X)}{1 - e(X'\theta^*)}\right\}\right] \\
& + \text{var}\left[\frac{\omega_\epsilon(X'\theta^*)\tau(X)}{E\{\omega_\epsilon(X'\theta^*)\}}\right] + \tilde{B}'\mathcal{I}(\theta^*)\tilde{B} \\
& + \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}^2} E\left\{\omega_\epsilon(X'\theta^*)^2 \left[\left\{\frac{1 - e(X'\theta^*)}{e(X'\theta^*)}\right\}^{1/2} \{\tilde{\mu}(1, X) - \mu(1, X)\}\right.\right.
\end{aligned}
$$

$$
\begin{aligned}
&- \left\{ \frac{e(X'\theta^*)}{1-e(X'\theta^*)} \right\}^{1/2} \{\tilde{\mu}(0,X) - \mu(0,X)\} \Bigg]^2 \Bigg\} \\
&+ \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}} \tilde{B}' E\left[\omega_\epsilon(X'\theta^*) X f(X'\theta^*) \left\{ -\frac{\tilde{\mu}(1,X_i) - \mu(1,X_i)}{e(X_i)} \right\} \right] \\
&+ \frac{1}{E\{\omega_\epsilon(X'\theta^*)\}} \tilde{B}' E\left[\omega_\epsilon(X'\theta^*) X f(X'\theta^*) \left\{ \frac{\tilde{\mu}(0,X) - \mu(0,X)}{1-e(X_i)} \right\} \right] \\
&= \tilde{\sigma}_\epsilon^2 + \tilde{B}' \mathcal{I}(\theta^*) \tilde{B} + \tilde{B}'(C_0 - C_1) \\
&= \tilde{\sigma}_\epsilon^2 + b_{1\epsilon}' \mathcal{I}(\theta^*)^{-1} b_{1\epsilon} + (C_0 + C_1)' \mathcal{I}(\theta^*)^{-1} (C_0 + C_1) + \tilde{B}'(C_0 - C_1)
\end{aligned}
$$

where $\tilde{\sigma}_\epsilon^2$, C_0 and C_1 are defined in Theorem 2. Because

$$
\begin{aligned}
\frac{\partial \hat{\tau}_\epsilon^{\mathrm{aug}}(\theta^*)}{\partial \theta} = &\frac{1}{N} \sum_{i=1}^N \frac{\partial}{\partial \theta} \left[\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}} \right] \hat{\tau}^{\mathrm{aug}}(X_i) \\
&- \frac{1}{N} \sum_{i=1}^N \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}} X_i f(X_i'\theta^*) \frac{A_i\{Y_i - \tilde{\mu}(A_i,X_i)\}}{e(X_i'\theta^*)^2} \\
&- \frac{1}{N} \sum_{i=1}^N \frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X_i'\theta^*)\}} X_i f(X_i'\theta^*) \frac{(1-A_i)\{Y_i - \tilde{\mu}(A_i,X_i)\}}{\{1-e(X_i'\theta^*)\}^2},
\end{aligned}
$$

we have

$$
\begin{aligned}
E\left\{ \frac{\partial \hat{\tau}_\epsilon^{\mathrm{aug}}(\theta^*)}{\partial \theta} \right\} = &E\left(\frac{\partial}{\partial \theta} \left[\frac{\omega_\epsilon(X'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}} \right] \tau(X) \right) \\
&- \frac{1}{E\{\omega_\epsilon(X_i'\theta^*)\}} E\left\{ \omega_\epsilon(X_i'\theta^*) X f(X'\theta^*) \frac{\mu(1,X) - \tilde{\mu}(1,X)}{e(X'\theta^*)} \right\} \\
&- \frac{1}{E\{\omega_\epsilon(X_i'\theta^*)\}} E\left\{ \omega_\epsilon(X_i'\theta^*) X f(X'\theta^*) \frac{\mu(0,X) - \tilde{\mu}(0,X)}{1-e(X'\theta^*)} \right\} \\
&= b_{1,\epsilon} - C_0 - C_1.
\end{aligned}
$$

Therefore,

$$
N^{1/2}(\hat{\tau}_\epsilon^{\mathrm{aug}} - \tau_\epsilon) \to \mathcal{N}\left\{ 0, \tilde{\sigma}_\epsilon^2 + b_{1\epsilon}' \mathcal{I}(\theta^*)^{-1} b_{1\epsilon} + (C_0 + C_1)' \mathcal{I}(\theta^*)^{-1} (C_0 + C_1) + \tilde{B}'(C_0 - C_1) \right\},
$$

in distribution, as $N \to \infty$.

S1·3. *Proof of Remark 1*

We show that

$$
b_{1,\epsilon} = E\left[\frac{\partial}{\partial \theta} \left\{ \omega_\epsilon(\theta^*)^{-1} \omega_\epsilon(X'\theta^*) \right\} \tau(X) \right]
$$

goes to zero, as $\epsilon \to 0$.

We note

$$
\frac{\partial}{\partial \theta} \left\{ \omega_\epsilon(\theta^*)^{-1} \omega_\epsilon(X'\theta^*) \right\} = \omega_\epsilon(\theta^*)^{-2} \left[\frac{\partial \omega_\epsilon(X'\theta^*)}{\partial \theta} E\{\omega_\epsilon(X'\theta^*)\} - E\left\{ \frac{\partial \omega_\epsilon(X'\theta^*)}{\partial \theta} \right\} \omega_\epsilon(X'\theta^*) \right],
$$

where

$$\begin{aligned}\frac{\partial \omega_\epsilon(X'\theta^*)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left[\Phi_\epsilon\left\{e(X'\theta^*)-\alpha_1\right\}\Phi_\epsilon\left\{\alpha_2-e(X'\theta^*)\right\}\right] \\ &= \phi_\epsilon\left\{e(X'\theta^*)-\alpha_1\right\}\Phi_\epsilon\left\{\alpha_2-e(X'\theta^*)\right\}f(X'\theta^*)X \\ &\quad -\Phi_\epsilon\left\{e(X'\theta^*)-\alpha_1\right\}\phi_\epsilon\left\{\alpha_2-e(X'\theta^*)\right\}f(X'\theta^*)X,\end{aligned}$$

and $\phi_\epsilon(x) = \mathrm{d}\Phi_\epsilon(x)/\mathrm{d}x$. As $\epsilon \to 0$, $\phi_\epsilon(x) \to 0$ implies that $b_{1,\epsilon}$ goes to 0.

S1·4. *Proof of Remark 4*

We write

$$\begin{aligned}\hat{\tau} &= \hat{\tau}(\hat{\theta}) \\ &\cong \hat{\tau}(\theta^*) + E\left\{\frac{\partial \hat{\tau}(\theta^*)}{\partial \theta'}\right\}(\hat{\theta}-\theta^*) \\ &\cong \frac{1}{N}\sum_{i=1}^{N}\frac{1\{\alpha_1 \le e(X_i'\theta^*) \le \alpha_2\}}{\mathrm{pr}\{\alpha \le e(X'\theta^*) \le 1-\alpha\}}\left\{\frac{A_iY_i}{e(X_i'\theta^*)}-\frac{(1-A_i)Y_i}{1-e(X_i'\theta^*)}\right\} + E\left\{\frac{\partial \hat{\tau}(\theta^*)}{\partial \theta'}\right\}\mathcal{I}(\theta^*)^{-1}S(\theta^*) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{1\{\alpha_1 \le e(X_i'\theta^*) \le \alpha_2\}}{\mathrm{pr}\{\alpha_1 \le e(X'\theta^*) \le \alpha_2\}}\left\{\frac{A_iY_i}{e(X_i'\theta^*)}-\frac{(1-A_i)Y_i}{1-e(X_i'\theta^*)}\right\} \\ &\quad + E\left\{\frac{\partial \hat{\tau}(\theta^*)}{\partial \theta'}\right\}\frac{1}{N}\sum_{i=1}^{N}X_i\frac{A_i-e(X_i'\theta^*)}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}f(X_i'\theta^*).\end{aligned}$$

Let $\mathcal{S} = \{X : e(X'\theta^*) = \alpha_1 \text{ or } \alpha_2\}$. If $\mathrm{pr}(X \in \mathcal{S}) = 0$, then

$$\begin{aligned}E\left\{\frac{\partial \hat{\tau}(\theta^*)}{\partial \theta'}\right\} &= E\left(\frac{\partial}{\partial \theta'}\left[\frac{1\{\alpha_1 \le e(X'\theta^*) \le \alpha_2\}}{\mathrm{pr}\{\alpha_1 \le e(X'\theta^*) \le \alpha_2\}}\right]\left\{\frac{AY}{e(X'\theta^*)}-\frac{(1-A)Y}{1-e(X'\theta^*)}\right\}\right) \\ &\quad + E\left[\frac{1\{\alpha_1 \le e(X'\theta^*) \le \alpha_2\}}{\mathrm{pr}\{\alpha_1 \le e(X'\theta^*) \le \alpha_2\}}\frac{\partial}{\partial \theta'}\left\{\frac{AY}{e(X'\theta^*)}-\frac{(1-A)Y}{1-e(X'\theta^*)}\right\}\right]\end{aligned}$$

is finite and well-defined, because the only possible problem that prevents the use of the bootstrap is the derivative of the indicator function with respect to θ, which, however, has zero measure.

Therefore, $\hat{\tau}$ is asymptotically linear. According to Shao & Tu (2012), the bootstrap can be used to estimate $\mathrm{var}(\hat{\tau})$. A similar discussion applies to $\hat{\tau}^{\mathrm{aug}}$.

S1·5. *Proof of Remark 5*

We write

$$\begin{aligned}
\hat{\tau}_\epsilon &= \hat{\tau}_\epsilon(\hat{\theta}) \\
&\cong \hat{\tau}_\epsilon(\theta^*) + E\left\{\frac{\partial \hat{\tau}_\epsilon(\theta^*)}{\partial \theta'}\right\}(\hat{\theta}-\theta^*) \\
&\cong \frac{1}{N}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left\{\frac{A_iY_i}{e(X_i'\theta^*)}-\frac{(1-A_i)Y_i}{1-e(X_i'\theta^*)}\right\}+E\left\{\frac{\partial \hat{\tau}_\epsilon(\theta^*)}{\partial \theta'}\right\}\mathcal{J}(\theta^*)^{-1}S(\theta^*) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left\{\frac{A_iY_i}{e(X_i'\theta^*)}-\frac{(1-A_i)Y_i}{1-e(X_i'\theta^*)}\right\} \\
&\quad +\Gamma'\frac{1}{N}\sum_{i=1}^{N}X_i\frac{A_i-e(X_i'\theta^*)}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}f(X_i'\theta^*),
\end{aligned}$$

where

$$\Gamma' = E\left\{\frac{\partial \hat{\tau}_\epsilon(\theta^*)}{\partial \theta'}\right\}\mathcal{J}(\theta^*)^{-1}.$$

Therefore, the asymptotic linearity of $\hat{\tau}_\epsilon$ follows.

Write

$$\begin{aligned}
\hat{\tau}_\epsilon^{\mathrm{aug}} = \hat{\tau}_\epsilon^{\mathrm{aug}}(\hat{\theta}) &\cong \hat{\tau}_\epsilon^{\mathrm{aug}}(\theta^*) + E\left\{\frac{\partial \hat{\tau}_\epsilon^{\mathrm{aug}}(\theta^*)}{\partial \theta'}\right\}(\hat{\theta}-\theta^*) \\
&\cong \frac{1}{N}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\hat{\tau}^{\mathrm{aug}}(X_i)+E\left\{\frac{\partial \hat{\tau}_\epsilon^{\mathrm{aug}}(\theta^*)}{\partial \theta'}\right\}\mathcal{J}(\theta^*)^{-1}S(\theta^*) \\
&\cong \frac{1}{N}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left[\frac{A_iY_i}{e(X_i)}+\left\{1-\frac{A_i}{e(X_i)}\right\}\tilde{\mu}(1,X_i)\right] \\
&\quad -\frac{1}{N}\sum_{i=1}^{N}\frac{\omega_\epsilon(X_i'\theta^*)}{E\{\omega_\epsilon(X'\theta^*)\}}\left[\frac{(1-A_i)Y_i}{1-e(X_i)}+\left\{1-\frac{1-A_i}{1-e(X_i)}\right\}\tilde{\mu}(0,X_i)\right] \\
&\quad +\tilde{\Gamma}'\frac{1}{N}\sum_{i=1}^{N}X_i\frac{A_i-e(X_i'\theta^*)}{e(X_i'\theta^*)\{1-e(X_i'\theta^*)\}}f(X_i'\theta^*),
\end{aligned}$$

where

$$\tilde{\Gamma}' = E\left\{\frac{\partial \hat{\tau}_\epsilon^{\mathrm{aug}}(\theta^*)}{\partial \theta'}\right\}\mathcal{J}(\theta^*)^{-1}.$$

Therefore, the asymptotic linearity of $\hat{\tau}_\epsilon^{\mathrm{aug}}$ follows.

The asymptotic linearity of the weighting estimators allows for using the bootstrap to construct confidence intervals.

S2. SIMULATION

We assess the performance of the new weighting estimators of the average treatment effect over a target population. We consider $X = (X_1, X_2, X_3, X_4, X_5, X_6)'$, where X_1, X_2, and X_3

are multivariate normal with means $(0,0,0)$, variances $(2,1,1)$ and covariances $(1,-1,-0.5)$, $X_4 \sim \text{Uniform}[-3,3]$, $X_5 \sim \chi^2_1$, and $X_6 \sim \text{Bernoulli}(0.5)$. The treatment indicator A is generated from Bernoulli$\{e(X)\}$. We consider four propensity score models:

(P1) $e(X) = \text{logit}\{0.1(X_1 + X_2 + X_3 + X_4 + X_5 + X_6)\}$,
(P2) $e(X) = \text{logit}\{0.8(X_1 + X_2 + X_3 + X_4 + X_5 + X_6)\}$,
(P3) $e(X) = \text{logit}\{0.1(X_1 + X_2^2 + X_3^2 + X_4 + X_5 + X_6)\}$,
(P4) $e(X) = \text{logit}\{0.8(X_1 + X_2^2 + X_3^2 + X_4 + X_5 + X_6)\}$;

(P1) and (P3) represent weak separations, and (P2) and (P4) represent strong separations of propensity score distributions between the treatment and control groups. We consider both linear and nonlinear outcome models:

(O1) $Y(a) = a(X_1 + X_2 + X_3 - X_4 + X_5 + X_6) + \eta$, with $\eta \sim \mathcal{N}(0,1)$, for $a = 0, 1$,
(O2) $Y(a) = a(X_1 + X_2 + X_3)^2 + \eta$, with $\eta \sim \mathcal{N}(0,1)$, for $a = 0, 1$.

The target population is represented by $\mathcal{O} = \{X : 0.1 \le e(X) \le 0.9\}$, and the estimand of interest is the average treatment effect over the target population $\tau(\mathcal{O})$.

We consider the weighting estimators with the indicator and smooth weight functions, and $\tau(\mathcal{O}) = \{\sum_{i=1}^N 1(X_i \in \mathcal{O})\}^{-1} \sum_{i=1}^N 1(X_i \in \mathcal{O})\{Y_i(1) - Y_i(0)\}$ for benchmark comparison with $N = 500$. The propensity scores are estimated by a logistic regression model with linear predictors X. Therefore, the propensity score model is correctly specified under (P1) and (P2) but misspecified under (P3) and (P4). For the augmented weighting estimators, $\mu(a, X)$ is estimated by a simple linear regression of Y on X, separately for $A = 0, 1$. Therefore, the outcome regression model is correctly specified under (O1) but misspecified under (O2).

Table S1 shows the simulation results. Under Scenarios i, ii, v and vi when the propensity score model is correctly specified, the weighting estimators are nearly unbiased for $\tau(\mathcal{O})$, and the augmented weighting estimators are nearly unbiased and more efficient than the simple weighting estimators. However, under Scenarios iii, iv, vii and viii when the propensity score model is misspecified, all estimators are biased, even when the outcome regression model is correctly specified. The weighting estimators with the smooth weight function, $\hat\tau_\epsilon$ and $\hat\tau_\epsilon^{\text{aug}}$, show slightly smaller variances than the counterparts with the indicator weight function, $\hat\tau$ and $\hat\tau^{\text{aug}}$. Moreover, as ϵ becomes smaller, the performances of $\hat\tau_\epsilon$ and $\hat\tau_\epsilon^{\text{aug}}$ become closer to those of $\hat\tau$ and $\hat\tau^{\text{aug}}$. The bootstrap works well with the variance estimates close to the true variances for all estimators.

S3. AVERAGE TREATMENT EFFECT ON THE TREATED

S3·1. *Notation, Assumptions and Extension of Crump et al. (2009)*

Another estimand of interest is the average treatment effect for the treated $\tau_{\text{ATT}} = E\{Y(1) - Y(0) \mid A = 1\} = E\{\tau(X) \mid A = 1\}$. The outcome distribution for the treated is empirically identifiable, because $E\{Y(1) \mid A = 1\} = E(Y \mid A = 1)$. Therefore, Assumptions 1 and 2 can be weakened (Heckman et al., 1997).

Assumption S1. $Y(0) \perp\!\!\!\perp A \mid X$.

Assumption S2. There exists a constant c such that with probability 1, $e(X) \le c < 1$.

A simple weighting estimator (Hirano et al., 2003) is

$$\hat\tau_{\text{ATT}} = \frac{\sum_{i=1}^N A_i Y_i}{\sum_{i=1}^N e(X_i'\hat\theta)} - \frac{\sum_{i=1}^N (1-A_i) Y_i e(X_i'\hat\theta)/\{1 - e(X_i'\hat\theta)\}}{\sum_{i=1}^N e(X_i'\hat\theta)} = \frac{\sum_{i=1}^N e(X_i'\hat\theta)\hat\tau(X_i)}{\sum_{i=1}^N e(X_i'\hat\theta)}, \tag{S11}$$

Table S1. *Results: mean, variance*$\times 100$ *(var), and variance estimate*$\times 100$ *(ve) based on* 100 *bootstrap replicates under eight combinations of the outcome and propensity score models: for example, (O1)&(P1) means Outcome Model (O1) and Propensity Score Model (P1)*

Scenario		i (O1)&(P1)			ii (O1)&(P2)			iii (O1)&(P3)			iv (O1)&(P4)		
	ϵ	mean	var	ve	mean	var	ve	mean	var	ve	mean	var	ve
$\tau(\mathcal{O})$		1.46			1.33			1.44			1.37		
$\hat{\tau}(\hat{\theta})$	–	1.45	3.4	3.4	1.33	4.7	5.2	1.48	2.9	2.8	1.45	4.0	4.1
$\hat{\tau}^{\mathrm{aug}}(\hat{\theta})$	–	1.46	2.8	2.7	1.32	3.4	3.4	1.50	2.6	2.5	1.49	3.3	3.2
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-4}	1.45	3.3	3.3	1.33	4.5	4.7	1.48	2.8	2.8	1.45	3.9	3.8
$\hat{\tau}_\epsilon^{\mathrm{aug}}(\hat{\theta})$	10^{-4}	1.46	2.8	2.7	1.33	3.4	3.3	1.50	2.6	2.5	1.49	3.3	3.1
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-5}	1.45	3.4	3.3	1.33	4.6	5.0	1.48	2.9	2.8	1.45	3.9	4.0
$\hat{\tau}^{\mathrm{aug}}(\hat{\theta})$	10^{-5}	1.46	2.8	2.7	1.32	3.4	3.4	1.50	2.6	2.5	1.49	3.3	3.2
		v (O2)&(P1)			vi (O2)&(P2)			vii (O2)&(P3)			viii (O2)&(P4)		
$\tau(\mathcal{O})$		7.58			6.69			7.62			5.96		
$\hat{\tau}(\hat{\theta})$	–	7.58	94.0	89.1	6.69	89.8	98.1	8.75	92.0	91.2	8.93	142.0	138.1
$\hat{\tau}^{\mathrm{aug}}(\hat{\theta})$	–	7.59	85.4	76.5	6.67	79.2	84.2	8.82	84.9	79.3	9.06	122.6	109.6
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-4}	7.57	88.6	84.1	6.70	85.3	89.7	8.75	91.1	88.3	8.94	134.2	128.4
$\hat{\tau}_\epsilon^{\mathrm{aug}}(\hat{\theta})$	10^{-4}	7.58	82.7	74.7	6.68	76.6	79.4	8.82	84.4	78.4	9.07	119.0	105.5
$\hat{\tau}_\epsilon(\hat{\theta})$	10^{-5}	7.57	92.0	87.3	6.69	88.8	95.3	8.75	91.9	90.2	8.93	140.0	134.8
$\hat{\tau}_\epsilon^{\mathrm{aug}}(\hat{\theta})$	10^{-5}	7.59	84.1	75.9	6.68	78.7	82.5	8.82	84.7	79.0	9.06	121.7	108.2

which is a special case of the weighting estimator (4) by choosing $\omega(X_i'\hat{\theta}) = e(X_i'\hat{\theta})$. Analogously, we propose the augmented weighting estimator

$$\hat{\tau}_{\mathrm{ATT}}^{\mathrm{aug}} = \frac{\sum_{i=1}^N e(X_i'\hat{\theta})\hat{\tau}^{\mathrm{aug}}(X_i)}{\sum_{i=1}^N e(X_i'\hat{\theta})}. \tag{S12}$$

*Remark S*1. An existing augmented weighting estimator for τ_{ATT} is

$$\frac{\sum_{i=1}^N A_i Y_i}{\sum_{i=1}^N A_i} - \frac{1}{\sum_{i=1}^N A_i}\sum_{i=1}^N \frac{(1-A_i)e(X_i'\hat{\theta})Y_i + \hat{\mu}(0,X_i)\{A_i - e(X_i'\hat{\theta})\}}{1-e(X_i'\hat{\theta})}, \tag{S13}$$

which is doubly robust in the sense that the estimator is consistent for τ_{ATT} if either $\mu(0,X)$ or $e(X)$ is correctly specified (Mercatanti & Li, 2014). See also Shinozaki & Matsuyama (2015) and Zhao & Percival (2017) for other forms of doubly robust estimators for τ_{ATT}. The advantage of these estimators is that they do not require estimating $\mu(1,X)$ unlike (S12). However, (S12) is locally efficient in the sense that if the outcome and propensity score models are correctly specified, the asymptotic variance of (S12) achieves the efficiency bound. To show this, we recognize that (S12) is (3) with $\omega_\epsilon(X'\hat{\theta})$ replaced by $e(X'\hat{\theta})$. Let $p_1 = E\{e(X'\theta^*)\}$. Following a similar derivation as in Theorem 2, with correctly specified propensity score and outcome models, the asymptotic variance of (S12) is

$$p_1^{-2}\mathrm{var}\{e(X'\theta^*)\tau(X)\} + p_1^{-2}E\left[e(X'\theta^*)^2\left\{\frac{\sigma^2(1,X)}{e(X)} + \frac{\sigma^2(0,X)}{1-e(X)}\right\}\right],$$

which differs from the efficiency bound for τ_{ATT} (Hahn, 1998).

There is a limited literature dealing with lack of overlap for $\tau_{\rm ATT}$ when Assumption S2 may not hold. Dehejia & Wahba (1999) suggested dropping control units with estimated propensity scores lower than the smallest value of the estimated propensity score among the treated units. Heckman et al. (1997) and Smith & Todd (2005) proposed to discard units with covariate values at which the estimated density is below some threshold. However, few formal results have been established on properties of these procedures.

Similar to Crump et al. (2009), if $\sigma^2(1,X)=\sigma^2(0,X)$, we can show that the optimal overlap for estimating $\tau_{\rm ATT}$ is of the form $\mathcal{O}=\{X: 1-e(X)\geq\alpha\}$ for some α, for which the estimators have smallest asymptotic variance. Intuitively, for the treated units with $e(X)$ close to 1, there are no similar units in the control group that can provide information to infer $Y(0)$ for these treated units. Statistically, the control units with $e(X)$ close to 1 contribute large weights. Therefore, it is reasonable to drop these units with $e(X)$ close to 1. By restricting to the subpopulation, the estimand of interest becomes $\tau_{\rm ATT}(\mathcal{O})=E\{\tau(X)\mid A=1, X\in\mathcal{O}\}$. Below, we formalize this argument.

S3·2. *Theory of trimming for the average treatment effect on the treated*

Define a general weighting average treatment effect,

$$\tau_\omega(\mathcal{O})=\frac{\sum_{i:X_i\in\mathcal{O}}\omega(X_i)\tau(X_i)}{\sum_{i:X_i\in\mathcal{O}}\omega(X_i)}. \tag{S14}$$

According to the technique report in 2006 prior to Crump et al. (2009), the efficiency bound for $\tau_\omega(\mathcal{O})$ is

$$V_\omega(\mathcal{O})=\frac{1}{[E\{\omega(X)\mid X\in\mathcal{O}\}]^2}E\left[\omega(X)^2\left\{\frac{\sigma^2(1,X)}{e(X)}+\frac{\sigma^2(0,X)}{1-e(X)}\right\}\mid X\in\mathcal{O}\right]. \tag{S15}$$

Crump et al. (2009) showed that the optimal set with which $\hat\tau_\omega(\mathcal{O})$ achieves the smallest asymptotic variance over all choices of $\mathcal{O}$ is

$$\mathcal{O}=\left\{x:\omega(x)\left\{\frac{\sigma^2(1,x)}{e(x)}+\frac{\sigma^2(0,x)}{1-e(x)}\right\}\leq\gamma\right\}, \tag{S16}$$

where γ is defined through the following equation:

$$\gamma=2\times\frac{E\left[\omega^2(X)\left\{\frac{\sigma^2(1,X)}{e(X)}+\frac{\sigma^2(0,X)}{1-e(X)}\right\}\mid\omega(X)\left\{\frac{\sigma^2(1,X)}{e(X)}+\frac{\sigma^2(0,X)}{1-e(X)}\right\}\leq\gamma\right]}{E\left[\omega(X)\mid\omega(X)\left\{\frac{\sigma^2(1,X)}{e(X)}+\frac{\sigma^2(0,X)}{1-e(X)}\right\}\leq\gamma\right]}. \tag{S17}$$

The weighting estimator for the average treatment effect on the treated is (S14) with $\omega(X)=e(X)$. Assuming that $\sigma^2(1,X)=\sigma^2(0,X)=\sigma^2$, the optimal set (S16) reduces to $\mathcal{O}=\{X: 1-e(X)\geq\alpha\}$ with the cut-off value $\alpha=\sigma^2/\gamma$. In practice, α can be determined by the smallest value of α that satisfy the empirical estimate of (S17):

$$\frac{1}{\alpha}=2\times\frac{\sum_{i=1}^N e^2(X_i)\left\{\frac{1}{e(X_i)}+\frac{1}{1-e(X_i)}\right\}1\{1-e(X_i)\geq\alpha\}}{\sum_{i=1}^N e(X_i)1\{1-e(X_i)\geq\alpha\}}.$$

The choice of α in $\mathcal{O}=\{X: 1-e(X)\geq\alpha\}$ has two opposite effects on the asymptotic variance in (S15). On the one hand, as α increases, we reduce the denominator of the right hand side of (S15), $[E\{\omega(X)\mid X\in\mathcal{O}\}]^2=E[\{e(X)\mid X\in\mathcal{O}\}]^2$, and therefore increase the asymptotic variance. On the other hand, as α increases, we decrease the numerator of the right hand side of

Table S2. *Descriptive statistics for covariates X in the original population in the National Health and Nutrition Examination Survey Data*

Covariate	
Age, interquantile range	$[35, 63]$
Income-to-poverty level, interquantile range	$[1.18, 3.77]$
Missing, %	8.5
Male, %	41.7
Education, %	
Less than 9th grade (Edu.lt9)	13.3
$9-11$th grade (Edu.9to11)	16.5
High school graduate (Edu.hischl)	25.1
Some college (Edu.somecol)	25.4
College (Edu.college)	19.6
Unknown	0.1
Race, %	
White	45.8
Black	19.1
Mexican American (Mexicanam)	18.3
Other Hispanic (Otherhispan)	11.6
Other races	5.2

(S15),

$$E\left[\omega(X)^2\left\{\frac{\sigma^2(1,X)}{e(X)}+\frac{\sigma^2(0,X)}{1-e(X)}\right\} \mid X\in\mathcal{O}\right] \\ = E\left[e(X)\sigma^2(1,X)+\frac{e(X)^2\sigma^2(0,X)}{1-e(X)} \mid X\in\mathcal{O}\right],$$

and therefore decrease the asymptotic variance. The optimal value of α balances the two effects.

S4. The National Health and Nutrition Examination Survey Data

S4·1. *Interpretation of the trimmed population for the average smoking effect*

To interpret the target population $\mathcal{O}=\{X: 0.05\le e(X)\le 0.6\}$ for the average smoking effect, an effective strategy is to first present summary statistics for covariates X in the original population. See Table S2 for the description of the covariates. Then, from the fitted logistic regression for $e(X)$, the target population can be represented by $\{X: -2.944 \le -9 - 0.018\times\text{Age}+0.841\times\text{Male}+8.972\times\text{Edu.lt9}+9.331\times\text{Edu.9to11}+8.875\times\text{Edu.hischl}+8.546\times\text{Edu.somecol}+7.118\times\text{Edu.college}-0.254\times\text{Income}-0.145\times\text{Income.mis}+0.689\times\text{White}-0.067\times\text{Black}-1.639\times\text{Mexicanam}-1.304\times\text{Otherhispan}\le -0.405\}$.

Table S3. *Estimate of the average smoking effect on the smokers, estimated standard error based on* 100 *bootstrap replicates, and* 95% *confidence interval*

	ϵ	estimate	s.e.	95% c.i.		estimate	s.e.	95% c.i.
$\hat{\tau}_{\text{ATT}}(\hat{\theta})$	–	0.796	0.103	$(0.591, 1.001)$	$\hat{\tau}^{\text{aug}}_{\text{ATT}}(\hat{\theta})$	0.793	0.088	$(0.616, 0.970)$
$\hat{\tau}_{\text{ATT},\epsilon}(\hat{\theta})$	10^{-4}	0.796	0.102	$(0.593, 0.999)$	$\hat{\tau}^{\text{aug}}_{\text{ATT},\epsilon}(\hat{\theta})$	0.792	0.088	$(0.616, 0.968)$
$\hat{\tau}_{\text{ATT},\epsilon}(\hat{\theta})$	10^{-5}	0.796	0.109	$(0.579, 1.013)$	$\hat{\tau}^{\text{aug}}_{\text{ATT},\epsilon}(\hat{\theta})$	0.793	0.088	$(0.617, 0.968)$

S4·2. *Analysis for the average smoking effect on the smokers*

For the average smoking effect on the smokers, we drop subjects with estimated propensity scores greater than 0.7. This removes 36 subjects, with 29 smokers and 7 non-smokers. Thus, the analysis sample includes 3304 subjects, with 650 smokers and 2654 non-smokers. Following the main paper for the average treatment effect, we consider the weighting estimators using both the indicator and smooth weight functions with $\epsilon = 10^{-4}$ and $\epsilon = 10^{-5}$. For the augmented weighting estimator, we consider the outcome model to be a linear regression model adjusting for all covariates, separately for $A = 0, 1$.

Table S3 shows the results from the estimators for the average smoking effect on the smokers based on the trimmed samples. The weighting estimators with the smooth weight function are close to the counterparts with the indicator weight function, but have slightly smaller estimated standard errors. The smooth weighting estimators are insensitive to the choice of ϵ. From the results, on average, smoking increases the lead level in blood at least by 0.79 ug/dl for smokers with $e(X) \leq 0.7$.

REFERENCES

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. & MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–199.

DEHEJIA, R. H. & WAHBA, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Am. Stat. Assoc.* **94**, 1053–1062.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

HECKMAN, J. J., ICHIMURA, H. & TODD, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econ. Stud.* **64**, 605–654.

HIRANO, K., IMBENS, G. W. & RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

MERCATANTI, A. & LI, F. (2014). Do debit cards increase household spending? evidence from a semiparametric causal analysis of a survey. *The Annals of Applied Statistics* **8**, 2485–2508.

SHAO, J. & TU, D. (2012). *The Jackknife and Bootstrap*. New York: Springer.

SHINOZAKI, T. & MATSUYAMA, Y. (2015). Doubly robust estimation of standardized risk difference and ratio in the exposed population. *Epidemiology* **26**, 873–877.

SMITH, J. A. & TODD, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* **125**, 305–353.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge, MA: Cambridge University Press.

ZHAO, Q. & PERCIVAL, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference* **5**, DOI: https://doi.org/10.1515/jci–2016–0010.