Received: 21 April 2017 | Revised: 24 April 2018 | Accepted: 27 April 2018

DOI: 10.1002/sim.7827

RESEARCH ARTICLE

WILEY Statistics in Medicine

Models for the propensity score that contemplate the positivity assumption and their application to missing data and causality

J. Molina[1] | M. Sued[1,2] | M. Valdora[3]

[1]Universidad de Buenos Aires, Ciclo Básico Común, Buenos Aires, Argentina

[2]Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina

[3]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Instituto de Cálculo, Buenos Aires, Argentina

Correspondence
M. Valdora, Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Intendente G'iraldes 2160, Ciudad Universitaria, Buenos Aires 1428, Argentina.
Email: mvaldora@dm.uba.ar

Funding information
Universidad de Buenos Aires, Grant/Award Numbers: 20020130100279BA and 20020150200110BA

Generalized linear models are often assumed to fit propensity scores, which are used to compute inverse probability weighted (IPW) estimators. To derive the asymptotic properties of IPW estimators, the propensity score is supposed to be bounded away from zero. This condition is known in the literature as strict positivity (or positivity assumption), and, in practice, when it does not hold, IPW estimators are very unstable and have a large variability. Although strict positivity is often assumed, it is not upheld when some of the covariates are unbounded. In real data sets, a data-generating process that violates the positivity assumption may lead to wrong inference because of the inaccuracy in the estimations. In this work, we attempt to conciliate between the strict positivity condition and the theory of generalized linear models by incorporating an extra parameter, which results in an explicit lower bound for the propensity score. An additional parameter is added to fulfil the overlap assumption in the causal framework.

KEYWORDS
average treatment effect, inverse probability weighting, missing data, observational studies, positivity

1 | INTRODUCTION

In the last 20 years, inverse probability weighted (IPW) estimators have attracted considerable attention in the statistical community. Among other things, they are used for estimating a population mean $E(Y)$ from an incomplete data set under the missing-at-random (MAR) assumption. Missing at random establishes that the variable of interest Y and the response indicator A are conditionally independent given an always observed vector $\mathbf{X}$ of covariates. See Robins et al,[1,2] Little and An,[3] and Kang and Schafer.[4] In the causal framework, IPW estimators are used for estimating the average effect of a binary treatment on a scalar outcome under the assumption of no unmeasured confounders in observational studies. See Rosenbaum,[5] Hirano et al,[6] Lunceford and Davidian,[7] and Crump et al.[8]

In the missing data setting, the propensity score, also known as the selection probability, is defined as the response probability given the vector of covariates $\mathbf{X}$, whereas, in the causal context, the propensity score is the conditional probability of treatment assignment given a set of measured baseline covariates. Inverse probability weighted estimators are essentially weighted means of observed responses in which the weights are determined as the inverse of the estimated propensity scores. The aim of these weights is to compensate for the missing responses.

One of the conditions required to derive the asymptotic properties of IPW estimators is the strict positivity condition (also known as the positivity assumption). It states, in the missing data setting, that the propensity score is bounded away from 0 and, in the causal context, that the propensity score is bounded away from 1 and 0 (see Robins et al,[2]

Kang and Schafer,[4] Lunceford and Davidian,[7] and Crump et al[8]). Besides theoretical issues, the violation of the strict positivity condition causes the estimates to be very unstable and to have a large variability. See also Little and Rubin[9] and Cole and Hernan.[10]

Most users of IPW procedures assume models for the propensity score that are usually incompatible with the positivity condition. For instance, this is the case of the generalized linear models (GLMs), when some of the covariates are unbounded. Despite this incompatibility, GLMs are the most popular models considered in the field.

In this work, we present a slight modification to the GLM originally postulated for the propensity score that is compatible with the strict positivity condition. When the covariate is 1-dimensional and a logistic regression is considered, this model agrees with the so-called 4-parameter logistic regression model in the dose-response literature; see, for instance, Ritz et al.[11] The proposed model can also be used in designed experiments, in which the treatment is assigned randomly based on $\mathbf{X}$. In this way, the validity of the overlap assumption is guaranteed.

To explore plausible applications of the modified model, we revise a real example in the causal context. We consider the data collected by Tager et al,[12,13] who investigate the effects of cigarette smoking on children's pulmonary function. They study the effect of parental smoking on the pulmonary function of their children as well as the effect of direct smoking by the children themselves. Besides, in his book, Rosner[14] performs another analysis of the mentioned data. The author considers the forced expiratory volume (FEV) of a group of children as a response variable as well as their height, age, sex, and a binary variable indicating whether or not they smoke. Kahn[15] studies these data using linear regression. In this work, we study the effect of smoking in the FEV of children by estimating the average treatment effect (ATE). This is an example of an observational study in which, under the assumption of no unmeasured confounders, the ATE can be estimated using IPW procedures. The validity of no unmeasured confounders is beyond the scope of this work, and thus, the estimates here presented for the ATE are included only to illustrate the use of the new model for the propensity score. Since the proportion of smokers among the younger children is very small, the corresponding propensity scores take on values near 0, and therefore, the classical IPW estimators may have a large variance. In situations such as this, we expect that the models we propose in this work will provide more stable estimators than the classical IPW procedures.

To achieve more stability in IPW estimators, Lunceford and Davidian[7] proposed adding an extra set of weights. These weights help to moderate the extreme values of the propensity score. They showed both theoretically and by simulations that this set of weights decreases the variance of IPW estimators. In Section 4, we give a precise description of this method, which we shall note *LD* for brevity. The problem of near violation of the positivity assumption has been attacked by many authors by means of trimming extreme values of the propensity score. In particular, Crump, Hotz, Imbens, and Mitnik[8] showed that, by eliminating from the sample observations with propensity score near 0 or 1, the variance of the final estimator decreases. They found the optimal cut-off value for this trimming, which depends on the sample. Even though this method is shown to decrease the variance of the ATE estimator, there is no control on the bias introduced by the trimming process. In fact, the estimators they proposed are not consistent for the ATE. Instead, they estimate a different parameter, namely, the ATE conditional to the propensity score belonging to a subinterval of (0,1). This estimator shall be noted *CHIM* in the rest of this work

In recent years, several authors have applied machine learning methods, such as classification and regression trees (CART), random forests, bagging, and boosted CART, to estimate the propensity score needed for IPW estimators; see, for example, Lee et al.[16] These methods are extremely useful when the number of covariates is large. Besides, most of them have the great advantage that no model needs to be assumed in order to apply them. However, the presence of covariates with estimated missing probability very near 0 or with probability of treatment assignment near 0 or 1 can still cause instability and high variability in the final estimators. On the other hand, IPW methods are still widely applied to datasets with a small or moderate number of covariates in situations in which the access to fast computers and advanced technology is still limited. In these cases, the use of computer intensive methods may be unjustified or even unfeasible. A combination of the proposal of machine learning methods and trimming has also been considered in Kang et al.[17]

The paper is structured as follows: in Section 2, we review methods based on propensity scores for estimating a population mean from incomplete data, and we present our model for the propensity score in that missing data context. In Section 3, we review methods for estimating ATEs that use weighting by the inverse of the probability of treatment and present our model in the causality context. In Section 4, we study the performance of our model through its impact in IPW estimators by means of a Monte Carlo study and compare it with other existing proposals. In Section 5, we apply the proposed method to estimate the mean cost of hospital stay in a group of patients of a particular hospital, with artificially omitted responses. As an illustration, in Section 6, we apply these methods to estimate the effect of smoking

in children. The proofs corresponding to the results presented in Section 2 are relegated to Appendix A, while those related to Section 3 can be similarly deduced and are, therefore, omitted.

2 | MISSING DATA

Assume that we are interested in estimating the mean value of a scalar response Y, based on a sample composed by an always observed vector $\mathbf{X}\in\mathbb{R}^p$ of covariates, while the response of interest is missing by happenstance for some subjects. We will assume that data are MAR.[18] This means that the missing mechanism is not related to the response of interest and it is only related to $\mathbf{X}$, the observed vector of covariates. Let A be a binary variable indicating whether Y is observed or not, namely, $A=1$ if Y is observed and $A=0$ if Y is missing. Missing at random establishes that

$$P(A=1|\boldsymbol{X},Y)=P(A=1|\boldsymbol{X})=\pi(\mathbf{X}). \tag{1}$$

$\pi(\boldsymbol{X})$ is known in the literature as the propensity score or selection probability.[19] This condition will be assumed along the remainder of this section. Up to integrability conditions, under the MAR assumption, $\mu_0=E(Y)$ can be represented in terms of the distribution of the observed data as $E(Y)=E\{AY/\pi(\mathbf{X})\}$. This representation invites us to estimate μ_0 by

$$\hat{\mu}(\hat{\pi}_n)=P_n\left\{\frac{AY}{\hat{\pi}_n(\mathbf{X})}\right\}, \tag{2}$$

where $\hat{\pi}(\mathbf{X})$ is a consistent estimator of $\pi(\mathbf{X})$ and P_n is the empirical mean operator, ie, $P_nV=n^{-1}\sum_{i=1}^n V_i$. These estimators are in consonance with those proposed by Horvitz and Thompson.[20] Observed responses corresponding to low values of the estimated propensity score are highly weighted since they should compensate for the high missing rate associated to such a level of covariates. For more details, see Robins et al[1] and Robins and Rotnitzky.[21] Different ways of estimating $\pi(\mathbf{X})$ lead to different estimators for μ_0; parametric and nonparametric estimators of the propensity score have been considered. Most of the asymptotic results established for $\hat{\mu}(\hat{\pi}_n)$, defined in (2), require the strict positivity condition, which establishes that $P\{\pi(\mathbf{X})\geq\varepsilon\}=1$, for some $\varepsilon>0$ (see Robins et al[2] and Kang and Schafer[4]). For the sake of completeness, we include the following result, which establishes the consistency of $\hat{\mu}(\hat{\pi}_n)$, defined in (2), under the following assumptions:

C.1 $\sup_{\mathbf{X}\in K}|\hat{\pi}_n(\mathbf{X})-\pi(\mathbf{X})|\to 0$ almost surely (a.s.), for all compact set $K\subset\mathbb{R}^p$.
C.2 There exists $\eta_0>0$ such that

$$P\left\{\inf_{\mathbf{X}\in\mathbb{R}^p}\pi(\mathbf{X})>\eta_0\right\}=1 \quad \text{and} \quad P\left\{\liminf_{n\to\infty}\inf_{\mathbf{X}\in\mathbb{R}^p}\hat{\pi}_n(\mathbf{X})>\eta_0\right\}=1. \tag{3}$$

Assumption **C.1** states that the estimated propensity score converges to the true propensity score uniformly over all compact sets. The first part of **C.2** is the strict positivity condition, while the second part states that the same bound is eventually satisfied by the estimated propensity score.

> **Lemma 1.** *Consider* $(\boldsymbol{X}_i,\mathrm{A}_i,\mathrm{Y}_i)_{i\geq 1}$ *i.i.d., distributed as* $(\boldsymbol{X},\mathrm{A},\mathrm{Y})$, *with* $\pi(\boldsymbol{X})=\mathrm{P}(\mathrm{A}=1|\mathbf{X})$. *Let* $\hat{\pi}_n(\mathbf{X})$ *be a sequence of estimators of* $\pi(\boldsymbol{X})$. *Assume that conditions* **C.1** *and* **C.2** *hold. Then,* $\hat{\mu}(\hat{\pi}_n)$, *defined in Equation 2, converges to* $\mathrm{E}(\mathrm{Y})$ *a.s.*

The class of estimators $\hat{\mu}(\hat{\pi}_n)$, defined in (2), has been considered by many authors. Robins, Rotnitzky, and Zhao[1] studied an enlarged family that include this class as a particular case and established the asymptotic normality of such estimators when π follows a parametric model. Little and An[3] proposed to estimate the propensity by fitting a spline to the logistic regression of the missing-data indicator A on $\boldsymbol{X}$. Kang and Shafer[4] presented $\hat{\mu}(\hat{\pi}_n)$, where π follows a linear logistic regression model. More generally, parametric models are often assumed for the propensity score, and, in particular, GLMs play a prominent role among them. These models postulate that

$$\pi(\mathbf{X})=\phi(\boldsymbol{\beta}_0^T\mathbf{X}), \tag{4}$$

where ϕ is a strictly increasing cumulative distribution function and $\boldsymbol{\beta}_0\in\mathbb{R}^p$. Note that the intercept, if present, is included in $\boldsymbol{\beta}_0$, assuming that the first entry of $\mathbf{X}$ equals 1. Taking $\phi(u)=1/\{1+exp(-u)\}$ results in the linear logistic

model, which is one of the most popular choices in the literature. Unfortunately, GLMs prevent the validity of the strict positivity condition, except when $\boldsymbol{\beta}_0^T\mathbf{X}$ is bounded from below. Thus, strict positivity is typically violated as far as unbounded covariates are included in the vector $\boldsymbol{X}$. In practice, values of $\hat{\pi}_n(\boldsymbol{X}_i)$ close to 0 may arise causing $\hat{\mu}(\hat{\pi}_n)$ to be a very unstable estimator, having a large variability (see also Little and Rubin[9]).

In this work, we attempt to conciliate between the strict positivity condition and the most popular parametric models used for the propensity score. To do so, we slightly perturb the original parametric model postulated by the practitioner by incorporating an explicit lower bound for the propensity score. Namely, model (4) is replaced by

$$\pi(\mathbf{X}) = (1-\varepsilon_0)\phi(\boldsymbol{\beta}_0^T\mathbf{X}) + \varepsilon_0,$$

where $\varepsilon_0 \in [0,1]$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. This model contemplates the validity of the strict positivity condition as far as $\varepsilon_0 > 0$, and thus, we call it strictly positive propensity score (SPPS) model for the missing mechanism. Let $\theta = (\varepsilon, \boldsymbol{\beta}^T)^T \in \Theta \subset \mathbb{R}^{p+1}$, with $\varepsilon \in [0,1]$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. Define

$$\pi(\mathbf{X}, \theta) = (1-\varepsilon)\phi(\boldsymbol{\beta}^T\mathbf{X}) + \varepsilon, \tag{5}$$

and assume that, for some $\theta_0 \in \boldsymbol{\Theta}$, $\pi(\mathbf{X}) = \pi(\mathbf{X}, \theta_0)$. To identify the parameter θ_0 from the distribution of $(\mathbf{X}, A)$, we will make the following assumptions.

A.1 The function $\phi: \mathbb{R} \to (0,1)$ is a strictly increasing smooth cumulative function ($\phi \in C^2(\mathbb{R})$).
A.2 The distribution of the observed covariates is not concentrated at a hyperplane.
A.3 For all $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $A|\mathbf{X} \sim A|\boldsymbol{\beta}^T\mathbf{X}$, the support of $\boldsymbol{\beta}^T\mathbf{X}$ is unbounded from below:

$$P(\boldsymbol{\beta}^T\mathbf{X} \leq u) > 0 \text{ , for all } u \in \mathbb{R}.$$

Conditions **A.1** and **A.2** are typically required for the identification of $\boldsymbol{\beta}_0$ in a GLM (4) (see McCullagh and Nelder[22]), while condition **A.3** is used for the identification of ε_0. Roughly speaking, it means that any linear combination $\boldsymbol{\beta}^T\mathbf{X}$ that carries all the information on $\mathbf{X}$ relevant for A is necessarily unbounded from below. If this condition does not hold, the positivity assumption is automatically satisfied, and there is no need for the enlarged model. Estimating using the extra parameter would induce an overfitting phenomenon. The identifiability of the parameter θ_0, indexing the proposed model, is established in the following lemma.

Lemma 2. *Let $(\boldsymbol{X}, \mathrm{A})$ be a random vector, such that $\mathrm{P}(\mathrm{A}=1|\boldsymbol{X}) = \pi(\boldsymbol{X}, \theta_0)$, with $\pi(\boldsymbol{X}, \theta)$ defined in (5). Under **A.1** to **A.3**, we get that*

$$P\{\pi(\mathbf{X}, \theta) = \pi(\mathbf{X}, \theta_0)\} < 1 \text{ , } \quad \forall \theta \neq \theta_0. \tag{6}$$

Substituting $\pi(\mathbf{X}, \hat{\theta}_n)$ for $\hat{\pi}_n(\mathbf{X})$ in (2), we propose to estimate the mean of Y by

$$\hat{\mu}_n = P_n\left\{\frac{AY}{\pi(\mathbf{X}, \hat{\theta}_n)}\right\}, \tag{7}$$

where $\hat{\theta}_n$ is any consistent estimator of θ_0 under model (5). In principle, the parametric nature of the proposed model invites to estimate θ_0 following a maximum likelihood principle. On the other hand, we are postulating a parametric model for the regression function of a binary response, and, therefore, nonlinear least squares (LS) procedures can also be invoked to estimate θ_0. These 2 estimators are consistent and asymptotically normal under regularity conditions. Such conditions typically involve finite moment assumptions and, thus, are satisfied when the support of $\mathbf{X}$ is compact. This compactness condition is not necessary but helps to avoid technical complications. Similar arguments as those used in Valdora and Yohai[23] can be used to prove the consistency and asymptotic normality of these estimators under minimal assumptions. A rigorous study of this issue is beyond the scope of this work.

We now establish the asymptotic properties of $\hat{\mu}_n$, defined in (7). To do so, let $\dot{\phi}(u)$ and $\ddot{\phi}(u)$ denote the first and second derivatives of $\phi(u)$, respectively, and let $\dot{\pi}(\mathbf{X}, \theta)$ denote the column vector of partial derivatives of $\pi(\mathbf{X}, \theta)$ with respect to each entry of the vector θ.

Theorem 1. *Consider $(\boldsymbol{X}_\mathrm{i}, \mathrm{A}_\mathrm{i}, \mathrm{Y}_\mathrm{i})_{\mathrm{i} \geq 1}$ i.i.d., distributed as $(\boldsymbol{X}, \mathrm{A}, \mathrm{Y})$, with $\pi(\boldsymbol{X}) = \mathrm{P}(\mathrm{A}=1|\mathbf{X})$. Assume that $\pi(\boldsymbol{X}) = \pi(\boldsymbol{X}, \theta_0)$, for $\pi(\boldsymbol{X}, \theta)$ defined in (5), with $\varepsilon_0 > 0$.*

(i) Consistency: Assume that $\hat{\theta}_n$ converges to θ_0 a.s. (or in probability). If $\dot{\phi}(u)$ is a continuous function, then $\hat{\mu}_n$, defined in (7), converges to μ_0 a.s. (or in probability).
(ii) Asymptotic normality: Assume that there exists an influence function $IF_{\theta_0}(\mathbf{X}, A)$ with $E\{IF_{\theta_0}(\mathbf{X}, A)\} = 0$ and $E\{||IF_{\theta_0}(\mathbf{X}, A)||^2\} < \infty$, such that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF_{\theta_0}(\mathbf{X}_i, A_i) + o_P(1). \tag{8}$$

Then, if $\ddot{\phi}(u)$ is continuous, $\sqrt{n}(\hat{\mu}_n - \mu_0)$ is asymptotically normal, with mean zero and asymptotic variance:

$$asvar\{\sqrt{n}(\hat{\mu}_n - \mu_0)\} = E\left[\left\{AY/\pi(\mathbf{X}, \theta_0) - \mu_0 - \mathbf{C}^T \mathrm{IF}_{\theta_0}(\mathbf{X}, A)\right\}^2\right], \tag{9}$$

$$where \quad \mathbf{C} = E\left\{\frac{m_1(\mathbf{X})\dot{\pi}(\mathbf{X}, \theta_0)}{\pi(\mathbf{X}, \theta_0)}\right\}, \quad with \quad m_1(\mathbf{X}) = E(Y|A = 1, \mathbf{X}). \tag{10}$$

The theorem says that in order to get the asymptotic distribution of $\hat{\mu}_n$, we need to have a linear expansion for the estimator $\hat{\theta}_n$ of θ_0.

At this point, it is worth mentioning that both maximum likelihood and nonlinear LS estimators are indeed M-estimators, which, under regularity conditions, are consistent and satisfy the linear expansion presented in (8) with influence function $IF_{\theta_0}(\mathbf{X}, A)$ given by

$$\mathrm{IF}^{ml}_{\theta_0}(\mathbf{X}, A) = \left[E\left\{\pi(\mathbf{X}, \theta_0)^{-1}(1 - \pi(\mathbf{X}, \theta_0))^{-1}\dot{\pi}(\mathbf{X}, \theta_0)\dot{\pi}(\mathbf{X}, \theta_0)^T\right\}\right]^{-1} \frac{A - \pi(\mathbf{X}, \theta_0)}{\pi(\mathbf{X}, \theta_0)(1 - \pi(\mathbf{X}, \theta_0))}\dot{\pi}(\mathbf{X}, \theta_0)$$

and

$$\mathrm{IF}^{ls}_{\theta_0}(\mathbf{X}, A) = \left[E\{\dot{\pi}(\mathbf{X}, \theta_0)\dot{\pi}(\mathbf{X}, \theta_0)^T\}\right]^{-1}\{A - \pi(\mathbf{X}, \theta_0)\}\dot{\pi}(\mathbf{X}, \theta_0),$$

for the maximum likelihood and the LS procedure, respectively; (see Van der Vaart[24]).

3 | AVERAGE TREATMENT EFFECT

Causal inference is a second field where IPW procedures play a crucial role. Consider, for instance, a dichotomous treatment variable T, where $T = 1$ represents an active treatment and $T = 0$ means that a control is assigned. The potential outcomes framework, introduced by Neyman[25] and Rubin,[26] is used to quantify the effect of the treatment on some response of interest, whenever this difference is different from zero. To do so, 2 potential outcomes (or counterfactual variables) $Y^{(0)}$ and $Y^{(1)}$ are defined to represent the outcome variable of interest that would be seen if an individual were to receive the treatment and the control, respectively. We are interested in estimating the ATE, defined as the difference between the mean values of the potential outcomes: $\tau_0 = E(Y^{(1)}) - E(Y^{(0)})$. $E(Y^{(1)})$ (respectively $E(Y^{(0)})$) represents the hypothetical mean of the response for the population of individuals where all of them are assumed to receive treatment (respectively control). So the difference between these means may be considered a resultant of the treatment, meaning that it has a "causal effect" on the response of interest, whenever this difference is different from zero.

The potential outcomes $Y^{(1)}$ and $Y^{(0)}$ constitute an artificial contraption that allows us to conceptualize what we mean by "causality." Only one of these variables is observed in each individual, and it is related to the observed response through the consistency assumption, which establishes (see Cole and Frangakis[27]) that if an individual follows the treatment ($T = 1$), then the potential outcome $Y^{(1)}$ is precisely his observed outcome. It also establishes the same thing for the control case. Therefore, under the consistency assumption, the observed outcome Y is related to the counterfactual variables through the identity $Y = TY^{(1)} + (1 - T)Y^{(0)}$. Thus, estimating the ATE is a missing data problem, since one of the counterfactuals ($Y^{(1)}$ or $Y^{(0)}$) is missing for each individual. To identify τ_0, we assume that a vector $\boldsymbol{X}$ with all possible counfounders is observed at each subject. This means that potential outcomes $Y^{(1)}$ and $Y^{(0)}$ are conditionally independent of the treatment exposure T given $\boldsymbol{X}$:

$$(Y^{(0)}, Y^{(1)}) \coprod T|\boldsymbol{X}. \tag{11}$$

This condition is known in the literature as strongly ignorable treatment assignment or no unmeasured confounders (see Rosenbaum and Rubin[19]) and will be assumed in the remainder of this section. In particular, (11) implies that

$$E\left(Y^{(1)}\right) = E\left\{\frac{TY}{\pi(\boldsymbol{X})}\right\}, \quad E\left(Y^{(0)}\right) = E\left\{\frac{(1-T)Y}{1-\pi(\boldsymbol{X})}\right\}, \tag{12}$$

where now the propensity score $\pi(\boldsymbol{X})$ is defined by $\pi(\boldsymbol{X}) = P(T=1|\boldsymbol{X})$. These representations of $E\left(Y^{(1)}\right)$ and $E\left(Y^{(0)}\right)$ immediately suggest the estimator for the ATE proposed by Rosenbaum,[5] given by

$$\hat{\tau}(\hat{\pi}_n) = P_n\left\{\frac{TY}{\hat{\pi}_n(\mathbf{X})}\right\} - P_n\left\{\frac{(1-T)Y}{1-\hat{\pi}_n(\mathbf{X})}\right\}, \tag{13}$$

where $\hat{\pi}_n(\mathbf{X})$ is a consistent estimator of $\pi(\mathbf{X})$. Notice that (13) involves 2 estimators such as those treated in Section 2. In the first term, T plays the role of A, while in the second one, $1-T$ does. The conditions required of A in the missing data setting now have to be required of both T and $1-T$. This is the reason for conditions **C.3** and **A.3** below.

C.3 There exists $\eta_1 < 1$ such that

$$P\left\{\sup_{\mathbf{X}\in\mathbb{R}^p} \pi(\mathbf{X}) < \eta_1\right\} = 1 \quad \text{and} \quad P\left\{\limsup_{n\to\infty} \sup_{\mathbf{X}\in\mathbb{R}^p} \hat{\pi}_n(\mathbf{X}) < \eta_1\right\} = 1.$$

Lemma 3. *Consider* $(\boldsymbol{X}_i, \mathrm{T}_i, \mathrm{Y}_i)_{i\geq 1}$ *i.i.d., distributed as* $(\boldsymbol{X}, \mathrm{T}, \mathrm{Y})$, *with* $\pi(\boldsymbol{X}) = \mathrm{P}(\mathrm{T}=1|\mathbf{X})$. *Let* $\hat{\pi}_n(\mathbf{X})$ *be a sequence of estimators of* $\pi(\boldsymbol{X})$. *Assume that conditions* ***C.1*****-*****C.3*** *hold. Then,* $\hat{\tau}(\hat{\pi}_n)$, *defined in (13), converges to* $\mathrm{E}(\mathrm{Y}^{(1)}) - \mathrm{E}(\mathrm{Y}^{(0)})$ *a.s.*

Lunceford and Davidian,[7] following the general framework of Robins et al[1] and the theory of M-estimation (see Stefanski and Boos[28]), presented large-sample theoretical properties of $\hat{\tau}(\hat{\pi}_n)$, defined in (13), when $\hat{\pi}_n$ is a parametric estimator of π, which is assumed to follow a linear logistic model. Yao, Sun, and Wang[29] postulated a GLM for the propensity score, as in (4), and estimated $\pi(\boldsymbol{X})$ with $\phi\left(\hat{\boldsymbol{\beta}}_n^T\boldsymbol{X}\right)$, where $\hat{\boldsymbol{\beta}}_n$ denotes the MLE under model (4). Nonparametric versions of (13) were presented and analyzed by Hirano, Imbens , and Ridder,[6] who proposed splines in order to estimate $\pi(\mathbf{X})$. Their estimator achieves the semiparametric efficiency bound established by Hahn.[30] Both $\hat{\pi}_n(\boldsymbol{X})^{-1}$ and $\{1-\hat{\pi}_n(\boldsymbol{X})\}^{-1}$ are now involved in the estimator presented in (13). Thus, in this scenario, the strict positivity assumption is restated in terms of a lower and an upper bound for the propensity score, which is now assumed to be bounded away from 0 and 1, in the sense that $P\{\varepsilon \leq \pi(\boldsymbol{X}) \leq 1-\delta\} = 1$, for some ε and δ greater than 0. This assumption, also known as (existence of) "overlap" in the covariate distribution, is usually assumed to derive the asymptotic properties of $\hat{\tau}(\hat{\pi}_n)$, for the different estimators $\hat{\pi}_n$ of π considered in the literature; see Robins et al,[2] Hirano et al,[6] Lunceford and Davidian,[7] and Crump et al.[8] Besides, the lack of overlap leads to an erratic behavior of $\hat{\tau}(\hat{\pi}_n)$, making the precise estimation of τ_0 difficult. To deal with this issue, some authors proposed trimmed versions of (13); see Crump et al[8] and Cole and Hernan.[10]

As already mentioned in Section 2, GLMs for the propensity score are typically incompatible with the strict positivity assumption. Thus, using the same ideas developed in the missing data setting, for $\theta = (\varepsilon, \delta, \boldsymbol{\beta}^T)^T \in \Theta \subseteq \mathbb{R}^{p+2}$, where $\varepsilon, \delta \in [0,1]$ with $\varepsilon + \delta < 1$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, we propose the following model

$$\pi(\mathbf{X}, \theta) = (1-\delta-\varepsilon)\phi(\boldsymbol{\beta}^T\mathbf{X}) + \varepsilon \tag{14}$$

and assume that $\pi(\mathbf{X}) = \pi(\mathbf{X}, \theta_0)$, for some $\theta_0 \in \boldsymbol{\Theta}$. This model can be used in designed experiments where the treatment is assigned randomly based on $\mathbf{X}$. In this way, the overlap condition will be automatically satisfied. When $\mathbf{X}$ is comprised of an intercept and a scalar variable and ϕ is the logistic function, we get the "4-parameter logistic regression model," which is used for fitting dose-response curves; see Ritz et al.[11] In addition to **A.1**, **A.2**, and **A.3**, we need the following assumption to identify the parameter θ_0:

A.4 For all $\boldsymbol{\beta}\in\mathbb{R}^p$ such that $T|\mathbf{X}\sim A|\boldsymbol{\beta}^T\mathbf{X}$, the support of $\boldsymbol{\beta}^T\mathbf{X}$ is unbounded from above:

$$P(\boldsymbol{\beta}^T\mathbf{X}\leq u)<1 \text{ , for all } u\in\mathbb{R}.$$

Lemma 4. *Under **A.1** to **A.4**, we get that* $\mathrm{P}\{\pi(\boldsymbol{X},\Theta)=\pi(\boldsymbol{X},\Theta_0)\}<1$ *for all* $\Theta\neq\Theta_0$.

Model (14) intends to preserve as much as possible from the original family postulated for the propensity, $\phi(\boldsymbol{\beta}^T\mathbf{X})$. The proposed modification contemplates the strict positivity condition by the inclusion of ε and δ. Model (14) will be called SPPS model for treatment assignment, or simply SPPS model if the context is clear. So the estimator for the ATE that we propose is given by

$$\hat{\tau}_n=\mathbb{P}_n\left\{\frac{TY}{\pi(\mathbf{X},\hat{\theta}_n)}\right\}-\mathbb{P}_n\left\{\frac{(1-T)Y}{1-\pi(\mathbf{X},\hat{\theta}_n)}\right\}, \tag{15}$$

where $\hat{\theta}_n$ is any consistent estimator of θ_0 under model (14). The following result establishes the consistency and asymptotic normality of $\hat{\tau}_n$. The asymptotic variance presented below is similar to that given in Yao et al[29] and Lunceford and Davidian,[7] where a GLM is assumed for the propensity score.

Theorem 2. *Consider* $(\boldsymbol{X}_i,\mathrm{T}_i,\mathrm{Y}_i)_{i\geq 1}$ *i.i.d., distributed as* $(\boldsymbol{X},\mathrm{T},\mathrm{Y})$, *with* $\pi(\boldsymbol{X})=\mathrm{P}(\mathrm{T}=1|\boldsymbol{X})$. *Assume that* $\pi(\boldsymbol{X})=\pi(\boldsymbol{X},\Theta_0)$, *for* $\pi(\boldsymbol{X},\Theta)$ *defined in (14), with* $\varepsilon_0>0$ *and* $\delta_0>0$. *Assume that* $\hat{\theta}_n$ *converges to* Θ_0 *a.s. (or in probability). Suppose* $\dot{\phi}(u)$ *is a continuous function, then* $\hat{\tau}_n$, *defined in (15), converges to* τ_0 *a.s. (or in probability). Moreover, if* $\hat{\theta}_n$ *satisfies a linear expansion, as in (8) but replacing* T_i *for* A_i *with influence function* $IF_{\theta_0}(\mathbf{X},T)$ *with* $E\{IF_{\theta_0}(\mathbf{X},T)\}=0$ *and* $E\{||IF_{\theta_0}(\mathbf{X},T)||^2\}<\infty$, *and* $\ddot{\phi}(u)$ *is continuous, then* $\sqrt{n}(\hat{\tau}_n-\tau_0)$ *is asymptotically normal, with mean zero and asymptotic variance*

$$asvar\{\sqrt{n}(\hat{\tau}_n-\tau_0)\}=E\left[\left\{\frac{TY}{\pi(\mathbf{X},\theta_0)}-\frac{(1-T)Y}{1-\pi(\mathbf{X},\theta_0)}-\tau_0-\mathbf{D}^T\,IF_{\theta_0}(\mathbf{X},T)\right\}^2\right], \tag{16}$$

$$where\quad \mathbf{D}=E\left\{\frac{m_1(\mathbf{X})\dot{\pi}(\mathbf{X},\theta_0)}{\pi(\mathbf{X},\theta_0)}+\frac{m_0(\mathbf{X})\dot{\pi}(\mathbf{X},\theta_0)}{1-\pi(\mathbf{X},\theta_0)}\right\},\quad with\quad m_t(\mathbf{X})=E(Y|T=t,\mathbf{X})\text{ , }t=0,1. \tag{17}$$

4 | MONTE CARLO STUDY

In this section, we report the results of a Monte Carlo study we performed in order to assess the advantages of considering the proposed model over the traditional approach. In practice, several covariates will be available for modeling the propensity score. To investigate performance in a realistic setting, Lunceford and Dadivian[7] carried out simulations involving continuous and discrete variables, some of them associated with both treatment exposure and potential response and others associated with potential responses but not treatment exposure. We generated variables as they did in one of their proposed scenarios, except for the variances involved in Σ, defined below. We considered covariates $\mathbf{X}=(1, X_1, X_2, X_3, V_1, V_2, V_3)$, a binary variable T and an outcome Y such that the variable T follows a Bernoulli distribution with

$$\pi(\mathbf{X})=\pi(\mathbf{X},\theta_0)=\varepsilon_0+(1-\delta_0-\varepsilon_0)\{1+exp(\boldsymbol{\beta}_0^T\mathbf{X})\},^{-1}$$

where $\theta_0=(\varepsilon_0,\ \delta_0,\ \boldsymbol{\beta}_0)$ and $\boldsymbol{\beta}_0$=(0,0.6, − 0.6,0.6,0,0,0).

We remark that, when $\varepsilon_0=\delta_0=0$, this SPPS model reduces to the usual linear logistic model. Different settings of ε_0 and δ_0 were chosen. The outcome was generated as

$$Y=\nu_0+\nu_1X_1+\nu_2X_2+\nu_3X_3+\nu_4T+\xi_1V_1+\xi_2V_2+\xi_3V_3+Z,$$

where Z is a normal standard variable independent of $\mathbf{X}$ and T, $(\nu_0,\nu_1,\nu_2,\nu_3,\nu_4)=(0,-1,1,-1,2)^T$ and $\xi=(-1,1,1)$. The joint distribution of $\mathbf{X}$ was specified by taking X_3 with Bernoulli(0.2) distribution and then generating V_3 as

Bernoulli with $P(V_3=1|X_3)=0.75X_3+0.25(1-X_3)$. Conditional on X_3, the vector (X_1,V_1,X_2,V_2) was generated as multivariate normal $\mathcal{N}(\rho_{X_3},\Sigma)$, where $\rho_0=(-1,-1,1,1)$; $\rho_1=(1,1,-1,-1)$; and

$$\Sigma=\begin{pmatrix} 3 & 0.5 & -0.5 & -0.5 \\ 0.5 & 3 & -0.5 & -0.5 \\ -0.5 & -0.5 & 3 & 0.5 \\ -0.5 & -0.5 & 0.5 & 3 \end{pmatrix}.$$

In Lunceford and Davidian,[7] the elements in the diagonal of Σ equal 1, instead of 3. We have increased the variance of the covariates to make the effect of working in an unbounded scenario more noticeable.

We generated $Nrep=1000$ samples of size $n=2000$ following the described model, under which the real value of the ATE is $\tau_0=2$. For each sample, we computed 4 estimators, namely, $\hat{\tau}_O$, $\hat{\tau}_P$, $\hat{\tau}_{LD}$, $\hat{\tau}_{LDP}$, and $\hat{\tau}_{CHIM}$. $\hat{\tau}_O$ is the original IPW estimator, defined by modeling the propensity score with the linear logistic model. That is to say, $\hat{\tau}_O=\hat{\tau}_n$ given in (15) assuming that $\pi(\mathbf{X})=\phi(\boldsymbol{\beta}_1^T\mathbf{X})$, with $\phi(u)=1/(1+exp(-u))$ for some $\boldsymbol{\beta}_1\in\mathbb{R}^7$. $\hat{\tau}_P$ is defined in (15), fitting an SPPS model (14) for the propensity score. On the other hand, $\hat{\tau}_{LD}$ denotes the estimator proposed by Lunceford and Davidian.[7] It is a modified IPW estimator of τ_0, in which the weights are redefined in such a way that they not do not take on extremely large values. The authors showed that this modification gives rise to a more stable procedure. We applied the same modification to the weights computed assuming the SPPS model and obtained the following estimator for the ATE, which combines the proposal of Lunceford and Davidian[7] and the SPPS model discussed in this work:

$$\begin{aligned}\hat{\tau}_{LDP}=&\left(P_n\left[\frac{T}{\pi(\mathbf{X},\hat{\theta}_n)}\left\{1-\frac{C_1}{\pi(\mathbf{X},\hat{\theta}_n)}\right\}\right]\right)^{-1}P_n\left[\frac{TY}{\pi(\mathbf{X},\hat{\theta}_n)}\left\{1-\frac{C_1}{\pi(\mathbf{X},\hat{\theta}_n)}\right\}\right]\\&-\left(P_n\left[\frac{1-T}{1-\pi(\mathbf{X},\hat{\theta}_n)}\left\{1-\frac{C_0}{1-\pi(\mathbf{X},\hat{\theta}_n)}\right\}\right]\right)^{-1}P_n\left[\frac{(1-T)Y}{1-\pi(\mathbf{X},\hat{\theta}_n)}\left\{1-\frac{C_0}{1-\pi(\mathbf{X},\hat{\theta}_n)}\right\}\right],\end{aligned}\tag{18}$$

where

$$C_1=\frac{P_n\left\{\frac{T-\pi(\mathbf{X},\hat{\theta}_n)}{\pi(\mathbf{X},\hat{\theta}_n)}\right\}}{P_n\left[\left\{\frac{T-\pi(\mathbf{X},\hat{\theta}_n)}{\pi(\mathbf{X},\hat{\theta}_n)}\right\}^2\right]},\quad C_0=-\frac{P_n\left\{\frac{T-\pi(\mathbf{X},\hat{\theta}_n)}{1-\pi(\mathbf{X},\hat{\theta}_n)}\right\}}{P_n\left[\left\{\frac{T-\pi(\mathbf{X},\hat{\theta}_n)}{1-\pi(\mathbf{X},\hat{\theta}_n)}\right\}^2\right]},$$

and $\hat{\theta}_n$ is the maximum likelihood (ML) estimator of θ_0 under model (14). Another possibility is to estimate θ_0 using the LS method. In all the simulations settings we considered, the ML method gave better results than the LS method; this is why we only report the results obtained with the former. Potential advantages of using LS instead ML for fitting the propensity score, such as robustness, may be the subject of further work. Finally, $\hat{\tau}_{CHIM}$ is the estimator proposed in Crump et al.[8] This estimator gave surprisingly good results in this setting, despite its lack of consistency. As in Crump et al,[8] we found in our simulations that the cut-off value obtained by the *CHIM* procedure is near $\alpha=0.1$, regardless of the values of ϵ and δ used in the SPPS model (14) to generate the data. Therefore, if the samples are generated with ϵ_0 and δ_0 greater than 0.1, then the propensity score is automatically bounded between 0.1 and 0.9 and thus trimming as in Crump et al[8] has almost no effect; the estimator is essentially the same as the IPW estimator. In these cases, the *CHIM* estimator is consistent and it benefits from the trimming procedure discarding observations with estimated propensity score which, because of random fluctuations, may be small.

Several experiments, unreported here, were made in order to study the performance of the machine learning techniques described in Lee et al[16] in this simulation setting. We found that the best results were achieved by the methods bagged CART and boosted CART. However, the large variability of the covariates in this simulation setting causes extreme estimated propensity scores to appear occasionally. For this reason, *LD*, *CHIM*, and the method we propose here give better results in this setting. Combining machine learning and trimming, as proposed in Kang et al,[17] improved these results, but they were still not as good as *LD*, *CHIM*, or the method proposed here.

To compute the maximum likelihood estimator for the SPPS model, we used an iterative procedure, combining a coordinate descent method with the iteratively reweighted LS method implemented in R. The initial estimator of $\boldsymbol{\beta}_0$

was taken to be the maximum likelihood estimator for the linear logistic model, while the initial estimators for ϵ_0 and $1-\delta_0$ were computed, respectively, as the minimum and maximum of the fitted values corresponding to the initial estimator of $\boldsymbol{\beta}_0$. If the estimates of ϵ_0 and $1-\delta_0$ are both near 0.5, then, in order to avoid convergence problems, we simply discard the modified model and estimate the propensity score using the linear logistic model. The functions necessary for computing these estimators are available in R upon request.

For each estimator $\hat{\tau}$, we computed an empirical mean squared error with the following formula:

$$MSE(\hat{\tau}) = \frac{1}{Nrep}\sum_{i=1}^{Nrep}\left(\hat{\tau}-\tau_0\right)^2. \tag{19}$$

The results of the Monte Carlo study are reported in Tables 1 and 2. In these simulations, we can see that in almost all the situations considered, our proposed estimators give better results than the corresponding classical ones. The *CHIM* method gives very good results in some situations. In others, such as when ϵ is small and δ is near 0.5, its performance is quite poor. We understand that this is due to the bias introduced by the trimming process. To summarize these results, we computed the maximum MSE attained by each estimator, among all the situations considered. The results are presented in Table 3. Since in observational studies, the true model for the propensity score is always unknown, the maximum MSE that can be attained by each estimator is an important measure of its performance. According to this measure, both our proposed methods (τ_P and τ_{LDP}) give much better results than the others.

The remaining situations, when samples are generated with large values of both ϵ_0 and δ_0, are only included in our report. In these cases, under the linear logistic model, even though the estimation of beta is not correct, the fitted values preserve approximately the bounds imposed by the generating data process and are very near to those obtained fitting the correct SPPS model. This is why there are not big differences in the final estimations of τ_0. Thus, in practice, our proposed estimator is useful when the fitted values of the propensity score are not all near 0.5 (larger than 0.3 with $T=0$ and smaller than 0.7 with $T=1$). When ϵ_0 and δ_0 are both 0.4 or greater, our estimator has convergence problems. If they are, our method may not converge, and the classical method has to be used.

The asymptotic normality of both $\hat{\mu}_n$ and $\hat{\tau}_n$ presented in Theorems 1 and 2 , respectively, can be used to derive asymptotic confidence intervals. In each case, the asymptotic variance can be estimated though a plug-in procedure. This yields the following estimator:

TABLE 1 Empirical mean squared errors of different estimators of ATE for simulated data sets generated following the SPPS model with different values of ϵ_0 and δ_0

$\delta \vert \epsilon$		**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
0.01	$\hat{\tau}$	0.904	0.473	0.078	0.083	0.179	0.266
	$\hat{\tau}_P$	0.05	0.045	0.047	0.054	0.069	0.089
	$\hat{\tau}_{LD}$	0.056	0.053	0.04	0.058	0.11	0.169
	$\hat{\tau}_{LDP}$	0.027	0.025	0.025	0.029	0.037	0.049
	$\hat{\tau}_{CHIM}$	0.026	0.051	0.075	0.039	0.008	0.006
0.02	$\hat{\tau}$	1.085	0.564	0.097	0.059	0.121	0.182
	$\hat{\tau}_P$	0.035	0.03	0.03	0.033	0.042	0.052
	$\hat{\tau}_{LD}$	0.061	0.063	0.04	0.04	0.073	0.114
	$\hat{\tau}_{LDP}$	0.021	0.019	0.018	0.021	0.027	0.035
	$\hat{\tau}_{CHIM}$	0.026	0.051	0.073	0.037	0.007	0.007
0.05	$\hat{\tau}$	1.214	0.69	0.144	0.037	0.044	0.065
	$\hat{\tau}_P$	0.022	0.017	0.013	0.015	0.018	0.021
	$\hat{\tau}_{LD}$	0.082	0.093	0.054	0.025	0.028	0.042
	$\hat{\tau}_{LDP}$	0.017	0.014	0.012	0.013	0.017	0.019
	$\hat{\tau}_{CHIM}$	0.024	0.047	0.067	0.032	0.006	0.008

Abbreviations: ATE, average treatment effect; SPPS, strictly positive propensity score.

TABLE 2 Empirical mean squared errors of different estimators of ATE for simulated data sets generated following the SPPS model with different values of ϵ_0 and δ_0

$\epsilon \vert \delta$		0.05	0.1	0.2	0.3	0.4	0.5
0.01	$\hat{\tau}$	1.437	1.139	0.31	0.1	0.183	0.353
	$\hat{\tau}_P$	0.071	0.048	0.049	0.054	0.064	0.079
	$\hat{\tau}_W$	0.059	0.064	0.068	0.081	0.127	0.206
	$\hat{\tau}_{LDP}$	0.031	0.028	0.03	0.033	0.039	0.05
	$\hat{\tau}_{CHIM}$	0.015	0.014	0.016	0.027	0.073	0.151
0.02	$\hat{\tau}$	1.498	1.206	0.367	0.102	0.108	0.204
	$\hat{\tau}_P$	0.044	0.033	0.033	0.036	0.041	0.056
	$\hat{\tau}_W$	0.064	0.072	0.065	0.059	0.075	0.12
	$\hat{\tau}_{LDP}$	0.025	0.023	0.023	0.026	0.03	0.04
	$\hat{\tau}_{CHIM}$	0.016	0.015	0.015	0.025	0.068	0.139
0.05	$\hat{\tau}$	1.214	1.046	0.393	0.114	0.047	0.054
	$\hat{\tau}_P$	0.022	0.021	0.02	0.022	0.028	0.038
	$\hat{\tau}_W$	0.082	0.098	0.08	0.049	0.032	0.037
	$\hat{\tau}_{LDP}$	0.017	0.016	0.016	0.017	0.022	0.03
	$\hat{\tau}_{CHIM}$	0.024	0.022	0.016	0.017	0.043	0.084

Abbreviations: ATE, average treatment effect; SPPS, strictly positive propensity score.

TABLE 3 Maximum mean squared errors of the ATE estimator computed by different methods for simulated data sets generated following the SPPS model with different values of ϵ_0 and δ_0

Estimator	Max MSE
$\hat{\tau}$	1.498
$\hat{\tau}_P$	0.089
$\hat{\tau}_{LD}$	0.206
$\hat{\tau}_{LDP}$	0.05
$\hat{\tau}_{CHIM}$	0.151

Abbreviations: ATE, average treatment effect; SPPS, strictly positive propensity score.

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{TY_i}{\pi(\mathbf{X}_i,\hat{\theta}_n)} - \frac{(1-T_i)Y_i}{1-\pi(\mathbf{X}_i,\hat{\theta}_n)} - \hat{\tau}_n - \hat{\mathbf{D}}^T \,\mathrm{IF}_{\hat{\theta}_n}(\mathbf{X}_i, T_i)\right\}^2, \tag{20}$$

with

$$\hat{\mathbf{D}} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i \dot{\pi}(\mathbf{X}_i,\hat{\theta}_n)}{\pi(\mathbf{X}_i,\hat{\theta}_n)^2} + \frac{(1-T_i)Y_i\dot{\pi}(\mathbf{X}_i,\hat{\theta}_n)}{(1-\pi(\mathbf{X}_i,\hat{\theta}_n))^2}. \tag{21}$$

Also, normal bootstrap intervals can be computed, as in Wasserman.[31, section 8.3] To illustrate these facts, we performed a final simulation study to evaluate empirical coverage of the asymptotically normal intervals, with the asymptotic variance estimated by formulas (20) and (21) and by the bootstrap method. We considered the $\epsilon_0 = \delta_0 = 0.1$ case in the data-generating process described above. The empirical coverage for nominal level 0.95 obtained was 0.9218, for the intervals computed using $\hat{v}$ defined in (20), and 0.9524 for the intervals computed using the bootstrap method.

Extensive simulations, unreported here, show that when the variance of the covariates is small, the estimation of the nuisance vector of parameters θ_0 is difficult and the estimations obtained are highly variable. The reason for this is that, in these cases, different combinations of $\boldsymbol{\beta}$, ϵ, and δ can result in very similar values of $\pi(\mathbf{X},\theta)$. The parameters of

interest, μ_0 and τ_0, however, can be estimated precisely because their estimators depend on $\hat{\theta}_n$ only through $\pi(\mathbf{X}, \hat{\theta}_n)$. Unfortunately, the estimators of the asymptotic variance given in (20) and (21) depend on the estimator of θ_0 and its asymptotic variance. We have found in our simulations that the estimation of the asymptotic variance of $\hat{\tau}_n$ improves when the variance of $\mathbf{X}$ increases but can be unreliable when the variance of $\mathbf{X}$ is small. For this reason, we prefer to use bootstrap intervals in real data examples.

5 | MISSING DATA EXAMPLE: HOSPITAL DATA

We considered a sample of 100 patients hospitalized in a Swiss hospital during 1999 for medical back problems. We studied the relationship between the cost of stay (Cost, in thousands of Swiss francs) and some explanatory variables that are available in administrative records: length of stay (LOS, in days); admission type (0 = planned, 1 = emergency); insurance type (0 = regular, 1 = private); age (years); sex (0 = female, 1 = male); and discharge destination (1 = home, 0 = another health institution). This data set was used in Marazzi and Yohai[32] and has no missing values. To study the performance of our proposed estimators, we artificially deleted some of the responses and computed the estimators in the sample with missing values. We repeated this procedure 1000 times.

In each replication, we generated a sample of dichotomous variables $A_1 \ldots A_n$ according to the following mechanism:

$$P(A_i = 1) = (1-\epsilon_0)\phi\{-0.1 * LOS_i + 1.1\} + \epsilon_0,$$

for different values of ϵ_0 and ϕ the logistic function. The responses with corresponding $A_i = 0$ were deleted from the sample and considered missing.

For each sample, we computed $\hat{\mu}_O$, $\hat{\mu}_P$, $\hat{\mu}_{LD}$, $\hat{\mu}_{LDP}$, and $\hat{\mu}_{CHIM}$ defined analogously to the corresponding $\hat{\tau}$ in the previous section. Even though *CHIM* estimator was proposed in a causal context, we adapted it to the missing data setting, by trimming only observations with estimated propensity score smaller than 0.1. The empirical mean cost of stay, $\mu_0 = 11.12$, was computed as the mean of all the $n = 100$ responses in the complete sample. The MSE of each estimator was computed as in (19), replacing μ_0 for $\bar{\hat{\tau}}$ and the corresponding estimator of μ_0 based on the $i-th$ sample for $\hat{\tau}_i$. The results are summarized in Table 4.

An inspection of the data reveals that observation 31 is atypical; it corresponds to a patient with $LOS = 64$ days, but a relatively low cost of 25 733.45 francs. Fitting a linear regression model to predict the cost of hospital stay using the available covariates results in an expected cost of 45 153.1 francs for patient 31.

In data sets such as this, where there is an observation with small propensity score and an atypical response, our proposed estimator will in general be a great improvement over the existing methods.

6 | ATE EXAMPLE: CHILDREN'S FEV DATA

This data set contains measurements of the FEV of 654 children and teenagers aged 3 to 19 years, together with their height, age, sex, and a binary variable indicating whether or not they smoke. It is basically the data set considered in Rosner,[14] which has been included in the R package covreg[33] with slight modifications. To estimate the average smoking effect in the FEV in this population, we consider only children aged 9 or older since there are not any smokers among the younger children in this data set.

TABLE 4 Estimates of the mean squared error of different estimators of the mean cost of stay for hospital data[a]

ϵ	$\hat{\tau}_O$	$\hat{\tau}_P$	$\hat{\tau}_{LD}$	$\hat{\tau}_{LDP}$	$\hat{\tau}_{CHIM}$
0.05	3.2557	2.0862	1.0806	1.0386	2.3712
0.1	2.5065	1.3951	0.7985	0.7678	1.8034
0.2	1.7195	0.7321	0.4505	0.4369	1.2349
0.3	1.0738	0.3610	0.2885	0.2753	0.8718
0.4	0.6507	0.1993	0.1692	0.1699	0.6374

Missing values are artificially generated using an SPPS model (5) with different values of ϵ.

Abbreviation: SPPS, strictly positive propensity score.

We wish to emphasize that this example is included to show the use of the proposed model in a causal inference context. The reliability of results depend on the no unmeasured counfounders assumption, and the discussion on its validity given the observed covariates is beyond the scope of this work. This example is therefore considered only as an illustration of the application of the proposed methods.

All 4 estimates yielded a negative ATE but differed in the size of this effect. As a means to determine the significance of these differences, we computed 95% normal confidence intervals where we estimated the standard error based on 1000 bootstrap samples. To obtain the bootstrap samples, we first broke the data set in 2, smokers (65 subjects) and nonsmokers (374 subjects). Then, we resampled separately 65 observations from the smokers group and 374 observations from the nonsmokers group. Let $\hat{\tau}_i$ be the estimator of the ATE based on the ith sample and $\bar{\hat{\tau}}$ the mean of $\hat{\tau}_1, \cdots \hat{\tau}_{Nboot}$. The bootstrap estimator of the standard error of each estimator $\hat{\tau}$ was computed by

$$\hat{SE} = \sqrt{\frac{1}{Nboot-1}\sum_{i=1}^{Nboot}\left(\hat{\tau}_i - \bar{\hat{\tau}}\right)^2}, \tag{22}$$

and the 95% bootstrap confidence interval based on $\hat{\tau}$ was computed as

$$\left[\hat{\tau} - 1.96\,\hat{SE}, \hat{\tau} + 1.96\,\hat{SE}\right]. \tag{23}$$

In Table 5, we report the estimates, together with their corresponding standard errors and 95% bootstrap confidence intervals for the ATE. The estimated values for both estimators based on SPPS model are smaller than those based on the classical one. However, the bootstrap confidence intervals show that the ATE is not significant. The conclusion is the same for all the estimators considered, except for the estimator based on the *CHIM* method. However, it is important to emphasize that this method is not really estimating the ATE, since it is not consistent for it, as explained in the introduction. As a consequence, the obtained that interval is for a different parameter and it is not possible to compare it with those based on the other estimators. We remark that the standard error of our proposed estimator $\hat{\tau}_P$ is slightly smaller than the standard errors of the classical ones.

The proposed SPPS model is an improvement over the classical model because it can tolerate unbounded covariates. In practice, finite samples can only consist of bounded covariates. However, covariates with a large variance, can bring about problems in estimations because they may behave as realizations of random vectors with unbounded support. To illustrate this phenomenon, we add 4 artificially generated data to the FEV data set, in order to increase the variance of the covariates. Two of the added observations correspond to very tall 18-year-old teenagers, and the other two correspond to very small 9-year-old children. Three of them are smokers, and 3 of them are males. The chosen FEV is a typical value for the given height and sex. The specific added points are given in Table 6. The estimates and 95% bootstrap confidence intervals are reported in Table 7.

7 | DISCUSSION

When estimating the mean of a sample with missing values or the ATE in a nonrandomized study, the propensity score is widely used in order to construct IPW estimators. These estimators are basically weighted means in which the weights are the inverse of the estimated values of the propensity score. For estimating the propensity score, several methods can be used being generalized liner models the most usual ones. However, GLM methods can produce estimated propensity

TABLE 5 Estimates for the ATE of smoking in children, their standard errors, and 95% bootstrap confidence intervals

	Estimator	Standard Error	Confidence Interval
$\hat{\tau}_O$	−0.2993	0.3601	(−1.0051 to 0.4065)
$\hat{\tau}_P$	−0.2655	0.1806	(−0.7685 to 0.2374)
$\hat{\tau}_{LD}$	−0.1919	0.2566	(−0.5458 to 0.1619)
$\hat{\tau}_{LDP}$	−0.2637	0.1927	(−0.6414 to 0.1140)
$\hat{\tau}_{CHIM}$	−0.2691	0.1364	(−0.5364 to −0.0018)

Abbreviations: ATE, average treatment effect.

TABLE 6 Artificially generated data for the ATE of smoking in children

Age	FEV	Height	Male	Smoke
18	5.00	76	1	0
18	5.12	76	1	1
9	1.50	52	0	1
9	1.62	54	1	1

Abbreviation: ATE, average treatment effect.

TABLE 7 Estimates for the ATE of smoking in children, their standard errors, and 95% bootstrap confidence intervals, for the modified data

	Estimator	Standard Error	Confidence Interval
$\hat{\tau}_O$	−0.2106	0.3064	(−0.8111 to 0.3899)
$\hat{\tau}_P$	−0.3026	0.1266	(−0.5791 to −0.0261)
$\hat{\tau}_{LD}$	−0.2477	0.1411	(−0.4959 to 0.0004)
$\hat{\tau}_{LDP}$	−0.2986	0.1365	(−0.5661 to −0.0311)
$\hat{\tau}_{CHIM}$	−0.4219	0.1510	(−0.7179 to −0.1260)

Abbreviation: ATE, average treatment effect.

scores that are very near 0 or 1, especially if the covariates have a large variance. When the estimated propensity score takes on extreme values (near 0 or 1), the resulting estimators become very unstable. Some authors, such as Lunceford and Davidian,[7] and Crump et al,[8] have proposed estimators that are more stable than the classical IPW estimators. We show that the method proposed by the former can be combined with the SPPS model to yield even more stable and still consistent estimators. The latter, in an attempt to decrease the variance of the estimations, change the object of the estimation, giving rise to a nonconsistent method.

In this work, we have proposed to estimate the propensity score using a modified GLM, which we call SPPS model, that includes 2 extra parameters, intended to bound the propensity score away from 0 and 1. The proposed method is attractive because of its simplicity, which allowed us to establish theoretical properties, such as consistency and asymptotic normality. Moreover, it can be combined with other methods that use GLMs for the propensity score, by replacing the GLM by a SPPS model, that is to say, by simply including 2 extra parameters. Finally, its excellent performance in finite samples has been shown by means of a Monte Carlo study and real data examples.

So far, we have analyzed data sets with a small number of covariates. When the number of covariates is very large, machine learning techniques can be used to estimate the propensity score; see Lee et al[16] and Kang et al.[17] Among these, the ridge, lasso, and elastic net methods for GLMs, as developed in Friedman et al,[34] are interesting possibilities. The SPPS model can be combined with these regularization techniques in order to get a bounded propensity score and, as a consequence, stable final estimators. The asymptotic properties and finite sample performance of these estimators are the subject of further work.

ACKNOWLEDGEMENTS

This research was partly supported by grants 20020150200110BA and 20020130100279BA from Universidad de Buenos Aires.

The authors would like to thank Dr Graciela Boente and Dr Víctor Yohai for helpful discussions and generous advice and the referees for their valuable comments and corrections.

Mariela Sued acknowledges the Abdus Salam International Center for Theoretical Physics (ICTP) for hosting her during a 2-month fruitful visit and providing her with all the resources that made possible the realization of the present work.

ORCID

M. Valdora http://orcid.org/0000-0001-7232-8775

REFERENCES

1. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866.
2. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc*. 1995;90(429):106-121.
3. Little R, An H. Robust likelihood-based analysis of multivariate data with missing values. *Stat Sin*. 2004;14:949-968.
4. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22:523-539.
5. Rosenbaum PR. Propensity score. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*, Vol. 5. New York: Wiley: New York; 1998:3551-3555.
6. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71(4):1161-1189.
7. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937-2960.
8. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199.
9. Little RJ, Rubin DB. Statistical analysis with missing data john wiley and sons; 1987.
10. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656-664.
11. Ritz C, Baty F, Streibig JC, Gerhard D. Dose-response analysis using R. *PLos one*. 2015 Dec 30;10(12). e0146021.
12. Tager IB, Weiss A, Rosner B, Speizer FE. Effect of parental cigarette smoking on the pulmonary function of children. *Am J Epidemiol*. 1979;110(1):15-26.
13. Tager IB, Weiss ST, Muñoz A, Rosner B, Speizer FE. Longitudinal study of the effects of maternal smoking on pulmonary function in children. *N Engl J Med*. 1983;309(12):699-703.
14. Rosner B. *Fundamentals of Biostatistics*. 5th ed. CA, Duxbury: Pacific Grove; 1999.
15. Kahn M. An exhalent problem for teaching statistics. *J Stat Educ*. 2005;13(2).
16. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346.
17. Kang J, Chan W, Kim MO, Steiner PM. Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores. *Commun Stat Appl Methods*. 2016;23(1):1-20.
18. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
19. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
20. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47(260):663-685.
21. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell N, Dietz K, Farewell VT, eds. *Aids Epidemiology*. Birkhäuser, Boston: Springer; 1992:297-331.
22. McCullagh P, Nelder JA. *Generalized Linear CRC Monograph on Statistics and Applied Probability*. New York: Springer Verlag; 1989.
23. Valdora M, Yohai VJ. Robust estimators for generalized linear models. *J Stat Plan Infer*. 2014;146:31-48.
24. Van der Vaart AW. *Asymptotic statistics*, Vol 3. Cambridge University Press; 1998.
25. Neyman J. Sur les applications de la theorie des probabilits˜ aux experiences agricoles. *Essai des principes Roczniki Nauk Rolniczych*. 1923;10:1-51.
26. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688.
27. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20(1):3-5.
28. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29-38.
29. Yao L, Sun Z, Wang Q. Estimation of average treatment effects based on parametric propensity score model. *J Stat Plan Inference*. 2010;140(3):806-816.
30. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*. 1998:315-331.
31. Wasserman L. *All of Statistics: a Concise Course in Statistical Inference*. New York: Springer Science and Business Media; 2013.
32. Marazzi A, Yohai VJ. Adaptively truncated maximum likelihood regression with asymmetric errors. *J Stat Plan Inference*. 2004;122(1):271-291.
33. Niu X, Hoff P. Covreg: a simultaneous regression model for the mean and covariance R package version; 2014.
34. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1.

How to cite this article: Molina J, Sued M, Valdora M. Models for the Propensity Score that Contemplate the Positivity Assumption and their Application to Missing Data and Causality. *Statistics in Medicine*. 2018;1–16. https://doi.org/10.1002/sim.7827

APPENDIX A

Proof of Lemma 1. *Since* $E(Y) = E\{AY/\pi(\mathbf{X})\}$, *we have*

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\hat{\pi}_n(\mathbf{X}_i)} - E(Y)\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\hat{\pi}_n(\mathbf{X}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\pi(\mathbf{X}_i)}\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\pi(\mathbf{X}_i)} - E\left\{\frac{AY}{\pi(\mathbf{X})}\right\}\right|. \tag{A1}$$

By the law of large numbers, the second term in the right-hand side of display (A1) converges to zero. We first prove that the first term in the sum above converges to zero. To deal with the first term note that, by ***C.2****, there exists* n_0 *such that the following inequalities hold for* $\mathrm{n} \geq \mathrm{n}_0 = \mathrm{n}_o(\omega)$, *a.s.:*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\hat{\pi}_n(\mathbf{X}_i)} - \frac{A_iY_i}{\pi(\mathbf{X}_i)}\right\} < \frac{1}{n}\frac{1}{\eta_0^2}\sum_{i=1}^{n}|Y_i||\pi(\mathbf{X}_i) - \hat{\pi}_n(\mathbf{X}_i)| \\
&< \frac{1}{n}\frac{1}{\eta_0^2}\sum_{i=1}^{n}|Y_i||\pi(\mathbf{X}_i) - \hat{\pi}_n(\mathbf{X}_i)|I(|\mathbf{X}_i| \leq m) + \frac{1}{n}\frac{1}{\eta_0^2}\sum_{i=1}^{n}|Y_i||\pi(\mathbf{X}_i) - \hat{\pi}_n(\mathbf{X}_i)|I(|\mathbf{X}_i| > m) \\
&< \frac{1}{\eta_0^2}\sup_{|\mathbf{X}| \leq m}|\pi(\mathbf{X}) - \hat{\pi}_n(\mathbf{X})|\sum_{i=1}^{n}\frac{|Y_i|}{n} + \frac{2}{\eta_0^2}\sum_{i=1}^{n}\frac{|Y_i|}{n}I(|\mathbf{X}_i| > m).
\end{aligned}$$

The first term in the above sum converges to zero because of ***C.1****. By the law of large numbers, for each* $\mathrm{m} \in \mathrm{N}$, *there exists* $\Omega_\mathrm{m} \subseteq \Omega$ *such that* $\mathrm{P}(\Omega_\mathrm{m}) = 1$ *and*

$$\sum_{i=1}^{n}\frac{|Y_i|}{n}I(|\mathbf{X}_i| > m) \to E\{|Y|I(|\mathbf{X}| > m)\} \; in \Omega_m.$$

Therefore,

$$\lim_{n\to\infty}\sup\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\hat{\pi}_n(\mathbf{X}_i)} - \frac{A_iY_i}{\pi_\infty(\mathbf{X}_i)}\right\} \leq E\{|Y|I(|\mathbf{X}| > m)\} a.s. \; for \; all \; \; m \in N.$$

Taking limits when $\mathrm{m} \to \infty$, *we get that*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\hat{\pi}_n(\mathbf{X}_i)} - \frac{A_iY_i}{\pi(\mathbf{X}_i)}\right\} = 0 \; a.s.$$

From this, the result follows.

Proof of Lemma 2. *Suppose that* $\pi(\boldsymbol{X},\theta) = \pi(\boldsymbol{X},\theta_0)$ *with probability one, for some* $\theta = (\varepsilon,\boldsymbol{\beta})$ *with* $\theta \neq \theta_0$. *This implies that* $\mathrm{A}|\boldsymbol{X} \sim \mathrm{A}|\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}$ *and*

$$P\{(1-\varepsilon)\phi(\boldsymbol{\beta}^T\mathbf{X}) + \varepsilon = (1-\varepsilon_0)\phi(\boldsymbol{\beta}_0^T\mathbf{X}) + \varepsilon_0\} = 1. \tag{A2}$$

Now, since $\phi(\mathrm{u})$ *converges to zero as* u *goes to minus infinity, given* $\mathrm{h} > 0$, *there exists* $\mathrm{M_h} > 0$ *such that* $(1-\varepsilon)\phi(\mathrm{u}) + \varepsilon \leq \varepsilon + \mathrm{h}$, *for* $\mathrm{u} \leq -\mathrm{M_h}$. *Invoking* ***A3****, on a set* Ω_h *with positive probability, we get that* $\pi(\theta, \boldsymbol{X}) = \pi(\theta_0, \boldsymbol{X})$ *and* $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X} < -\mathrm{M_h}$.
Thus, on Ω_h, *we get that*

$$\varepsilon_0 \leq \pi(\boldsymbol{\beta}_0, \mathbf{X}) = \pi(\boldsymbol{\beta}, \mathbf{X}) \leq \varepsilon + h,$$

and therefore, $\varepsilon_0 \leq \varepsilon$. *Symmetrically, we get the opposite inequality and deduce that* $\varepsilon = \varepsilon_0$. *Appealing now to* ***A1*** *to* ***A2****, we deduce that* $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, *following standard arguments related to identifiability in GLMs. Thus, we have proved the validity of (6).*

Proof of Theorem 1. *(i) Consistency: We will show that* ***C.1*** *and* ***C.2*** *are satisfied with* $\hat{\pi}_n(\mathbf{X}) = \pi(\mathbf{X}.\hat{\theta}_n)$. *Then, the result follows from Lemma 1. Let* $C_\delta = \bigcup_{n_0 \in N}\bigcap_{n \geq n_0}\left\{\inf_{\mathbf{X} \in R^p}\pi(\mathbf{X}, \hat{\theta}_n) > \delta\right\}$.

Since $\hat{\varepsilon}_n$ *converges a.s. to* $\varepsilon_0 > 0 \hat{\epsilon}_n > \epsilon_0/2$ *a.s. if* n *is large enough. On the other hand,* $\pi(\mathbf{X}, \hat{\theta}_n) > \hat{\epsilon}_n$. *It follows that*

$$\left\{\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \to \boldsymbol{\theta}_0\right\} \subset \left\{\widehat{\varepsilon}_n \to \epsilon_0\right\} \subset \bigcup_{n_0 \in N} \bigcap_{n \geq n_0} \left\{\inf_{\mathbf{X} \in R^p} \pi(\mathbf{X}, \widehat{\theta}_n) > \varepsilon_0/2\right\}.$$

This implies ***C.1****, with* $\eta_0 = \varepsilon_0/2$. *To prove* ***C.2****, we write*

$$\left|\pi(\mathbf{X}, \hat{\theta}_n) - \pi(\mathbf{X}, \theta_0)\right| \leq \left|\dot{\pi}(\mathbf{X}, \boldsymbol{\xi}_n)^T \left(\hat{\theta}_n - \theta_0\right)\right| \leq \|\dot{\pi}(\mathbf{X}, \boldsymbol{\xi}_n)\| \left\|\hat{\theta}_n - \theta_0\right\|.$$

Therefore, since π *has a continuous derivative with respect to* Θ, *we have that, for every compact set* $K \subset R^p$, *there exists a constant* M *such that*

$$\sup_{\mathbf{X} \in K} \left|\pi(\mathbf{X}, \hat{\theta}_n) - \pi(\mathbf{X}, \theta_0)\right| \leq M \left\|\hat{\theta}_n - \theta_0\right\|,$$

which converges to zero a.s. because of the assumed consistency of $\hat{\theta}_n$. *(ii) Asymptotic normality: Consider the following expansion:*

$$\sqrt{n}(\hat{\mu}_n - \mu_0) = n^{-1/2} \sum_{i=1}^{n} A_i Y_i \left\{\pi(\mathbf{X}_i, \hat{\theta}_n)^{-1} - \pi(\mathbf{X}_i, \theta_0)^{-1}\right\} + n^{-1/2} \sum_{i=1}^{n} \left\{\frac{A_i Y_i}{\pi(\mathbf{X}_i, \theta_0)} - \mu_0\right\}. \tag{A3}$$

Performing a Taylor expansion and invoking the asymptotic linear representation for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *presented in (8), we get that*

$$n^{-1/2} \sum_{i=1}^{n} A_i Y_i \left\{\pi(\mathbf{X}_i, \hat{\theta}_n)^{-1} - \pi(\mathbf{X}_i, \theta_0)^{-1}\right\} = n^{-1/2} \sum_{i=1}^{n} -\frac{A_i Y_i}{\pi(\mathbf{X}_i, \hat{\boldsymbol{\xi}}_n)^2} \dot{\pi}(\mathbf{X}_i, \hat{\boldsymbol{\xi}}_n)(\hat{\theta}_n - \theta_0) = \tag{A4}$$

$$n^{-1/2} \sum_{i=1}^{n} -\mathbf{C}^T IF_{\theta_0}(\mathbf{X}_i, A_i) + o_P(1), \tag{A5}$$

$\mathbf{C} = E\left\{AY\dot{\pi}(\mathbf{X}, \theta_0)/\pi(\mathbf{X}, \theta_0)^2\right\}$, *which is equal to the vector defined in (10) by MAR. Thus,*

$$\sqrt{n}(\hat{\mu}_n - \mu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{A_i Y_i / \pi(\mathbf{X}_i, \theta_0) - \mu_0 - \mathbf{C}^T IF_{\theta_0}(\mathbf{X}_i, A_i)\right\} + o_P(1), \tag{A6}$$

and from this linear expansion, the asymptotic normality follows from the central limit theorem.