Robust doubly protected estimators for quantiles with missing data

Mariela Sued · Marina Valdora · Víctor Yohai

Received: 27 July 2018 / Accepted: 10 November 2019 / Published online: 21 November 2019
© Sociedad de Estadística e Investigación Operativa 2019

Abstract
Doubly protected methods are widely used for estimating the population mean of an outcome $Y$ from a sample where the response is missing in some individuals. To compensate for the missing responses, a vector $X$ of covariates is observed at each individual, and the missing mechanism is assumed to be independent of the response, conditioned on $X$ (missing at random). In recent years, many authors have turned from the estimation of the mean to that of the median, and more generally, doubly protected estimators of the quantiles have been proposed. In this work, we present doubly protected estimators for the quantiles in semiparametric models that are also robust, in the sense that they are resistant to the presence of outliers in the sample.

Keywords Robust estimators · Missing data · Median · Quantiles · Propensity score · Doubly protected estimators

This research was partially supported by Grant PICT 2014-0351 from ANPCYT and Grants 20020150200110BA and 20020130100279BA from the Universidad de Buenos Aires at Buenos Aires, Argentina.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11749-019-00689-9) contains supplementary material, which is available to authorized users.

Marina Valdora
mvaldora@gmail.com; mvaldora@dm.uba.ar

Mariela Sued
marielasued@gmail.com

Víctor Yohai
vyohai@dm.uba.ar

1 Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Intendente Guiraldes 2160, Ciudad Universitaria, Pabellón II - 2do. piso, C1428EGA Buenos Aires, Argentina

2 Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Intendente Guiraldes 2160, Ciudad Universitaria, Pabellón II - 2do. piso, C1428EGA Buenos Aires, Argentina
1 Introduction

Missing values occur in many situations when dealing with data sets from different fields and have attracted the attention of the statistical community during the last decades. In particular, the estimation of the quantiles of an outcome $Y$ from an incomplete data set under the missing at random (MAR) assumption has recently been considered by many authors; see, for instance, Bianco et al. (2018), Díaz (2017) and Zhang et al. (2012). MAR establishes that the variable of interest $Y$ and the response indicator $A$ are conditionally independent given an always observed vector $X \in \mathbb{R}^p$ of covariates; see Rubin (1976).

Causal inference is an area where missing data inevitably occur because counterfactual variables can never be observed simultaneously. The average treatment effect and the effect of a treatment on the quantiles are defined in terms of the means and quantiles of counterfactual variables, and thus, the estimation of the mean and of the quantiles of an outcome under the MAR assumption (no unmeasured confounders in the causal framework) is of great interest in the causal literature.

Most of the existing proposals to estimate the mean are based on three different approaches: inverse probability weighted (IPW), outcome regression (OR) and doubly protected (DP) methods. IPW methods, born in the seminal work of Horvitz and Thompson (1952), are based on estimating the propensity score $\pi(X) = P(A = 1|X)$, while OR procedures require a consistent estimator of $g(X) = E(Y|X)$, the regression of $Y$ on $X$. Doubly protected estimators combine IPW and OR methods providing consistent estimators when either the propensity score or the regression function is consistently estimated.

The propensity score can be estimated using nonparametric techniques, as in Hirano et al. (2003), or postulating a parametric working model; for example, a logistic regression model has been considered in Kang and Schafer (2007). Methods to estimate $g(X)$ include kernel smoothing (Cheng 1994) and semiparametric estimation (Wang et al. 2004; Robins et al. 1995), among others. A comprehensive overview of parametric regression estimators is given in Little and Rubin (2014). Rotnitzky et al. (2017), in a much more general setting, propose doubly protected methods that use nonparametric machine learning estimators of nuisance functionals in lieu of parametric models.

When the interest lies in the estimation of the median or other quantiles, rather than the mean, IPW, OR and DP methods can be used to obtain an estimator $\hat{F}_n$ of the probability distribution function $F_0$ of $Y$. Then, the quantiles of $\hat{F}_n$ are used to estimate those of $F_0$, as in Bianco et al. (2018), Díaz (2017) and Zhang et al. (2012).

Atypical observations, called outliers, are quite common in real datasets. Classical procedures do not contemplate their existence and therefore their application may lead to wrong conclusions. For instance, the sample mean or the least-squares fit of a regression model can be very adversely influenced by outliers, even by a single one. Robust methods arise to cope with these atypical observations, mitigating their impact in the estimation procedure. The median is, probably, the most popular example of a robust procedure to summarize a univariate dataset; other location parameters can also
be considered. For instance, in the missing data setting, Bianco et al. (2010) proposed IPW and also OR procedures to construct robust estimators of any location parameter of the marginal distribution of $Y$. Asymptotic properties of these estimators are presented in Bianco et al. (2011). Sued and Yohai (2013) deal with the estimation of the marginal distribution of $Y$ considering a semiparametric regression model. Predicted values are combined with observed residuals to emulate a complete data set. A robust fit of the regression model is used to take care of anomalous observations. Later, Statti et al. (2018) kept the observed responses and only emulate pseudo-observations to compensate for the missing ones. Recently, Bianco et al. (2018) considered both IPW and OR procedures based on the estimation of the propensity score and the regression function of $Y$ on $X$, respectively, under very general regression settings; they also contemplate the existence of missing covariates. All these proposals can be used to robustly and consistently estimate any parameter defined through a weakly continuous functional at the marginal distribution of the outcome.

As far as doubly protected estimation of the quantiles is concerned, Díaz (2017) considers efficient estimators that use machine learning techniques for estimating the propensity score and regression function, while Zhang et al. (2012) propose methods based on parametric models for both of these functionals. However, none of these methods is robust.

In this work we reformulate the robust OR procedure presented in Sued and Yohai (2013) to estimate quantiles of the marginal distribution of $Y$ to get an estimator that, additionally, is doubly protected. In this way, we get a robust doubly protected estimator of the quantiles. Robustness in this context means that, if a small proportion of observations do not follow the true regression model that generates the majority of the data, the estimations of the quantiles behave almost as if these atypical observations were not present, even if the propensity score is not estimated consistently. On the other hand, if there are no atypical observations, our proposed estimator gives results that are comparable to those given by existing doubly protected non-robust procedures: Díaz (2017) and Zhang et al. (2012). This means that the loss in efficiency that typically happens when using robust estimators is negligible in the present case. The robustness of our method to estimate the quantiles of the marginal distribution of $Y$ is accomplished by using robust procedures for estimating the regression function. Besides, our proposal does not impose any parametric model for the conditional distribution of $Y$ on $X$, as assumed in Zhang et al. (2012). For this reason, it can be considered a semiparametric method. The robustness and efficiency of our proposed method are shown by means of a Monte Carlo simulation study.

The paper is organized as follows. In Sect. 2 we review the existing procedures for estimating quantiles from samples with missing data presented in the literature and we introduce our estimators. In Sect. 3 we show their consistency. In order to assess the performance of our estimating procedure in finite samples, in Sect. 4 we present the results of a Monte Carlo simulation study and in Sect. 5 we apply it to estimate the median cost of hospital stay in a real data set. The proof of the results presented in Sect. 3 is relegated to Appendix. Additionally, a Supplementary Material is available, including a sketch of the proof of the $\sqrt{n}$ asymptotic normality under a semiparametric scenario, and additional results regarding the Monte Carlo simulation study presented in Sect. 4.
2 A robust doubly protected estimator for quantiles

In this section we introduce a new robust doubly protected estimator for any quantile of $Y$, based on a sample with missing data. Moreover, this estimator is robust; that is, it is not much influenced by a small fraction of atypical observations (outliers). Consider a sample $(X_i, A_i, Y_i), i = 1, \ldots, n$, distributed as $(X, A, Y)$, where $X \in \mathbb{R}^p$ is an always observed vector of covariates, $Y$ is the response of interest, which is missing at random (MAR), and $A$ denotes the indicator of whether or not $Y$ is missing, i.e., $A = 1$ if $Y$ is observed, and $A = 0$ when $Y$ is missing. MAR establishes that, conditioned on $X, Y$ and $A$ are independent. We use $\pi(X) = P(A = 1|X)$ for the so-called propensity score. In the sequel, we use $\hat{\pi}_n(X)$ to denote an estimator of the propensity score.

On the other hand, in the present formulation the construction of OR procedures assumes a regression model of the form

$$Y = g(X) + u, \quad (1)$$

with $(A, X)$ independent of $u$. We refer to $g(X)$ as the regression function. Since $u$ is independent of $A$, $g(X)$ can be estimated using the data for which the responses are available, namely using the $(X_i, Y_i)$ with $A_i = 1$. Let $\hat{g}_n(X)$ denote an estimator of the regression function and $\hat{u}_j = Y_j - \hat{g}(X_j)$, for $j$ with $A_j = 1$, denote the $j$th residual.

There exists a vast range of robust methods that can be used to estimate the regression function under different scenarios. For instance, robust nonparametric estimators of the regression function are discussed in Boente and Fraiman (1989). However, due to the curse of dimensionality, these procedures can only be applied when the dimension $p$ of $X$ is very low. A robust semiparametric partly linear model is considered in Bianco and Boente (2004), while a robust bandwidth selection procedure is discussed in Boente and Rodriguez (2008); a robust approach for a single index model is given in Boente and Rodriguez (2012) and Agostinelli et al. (2017). Robust and efficient estimators (MM) for the linear and nonlinear case are presented in Yohai (1987) and Bianco and Spano (2019) or Fasano et al. (2012), respectively. The use of these robust methods to estimate the regression function gives rise to a robust recipe to estimate the quantile of the marginal distribution of the outcome $Y$. The asymptotic properties of all these methods have been established assuming that the postulated model for the regression function is correct. However, under suitable conditions, the estimators converge to some limit function even when the assumed model is not true. This property will be used later, when analyzing the asymptotic behavior of the doubly protected estimator presented in this work.

The $p$-quantile of a distribution $F$ is defined as

$$T_p(F) = \inf\{x : F(x) \geq p\}. \quad (2)$$

When $p = 0.5$, $T_{0.5}(F)$ is the median of $F$. The representation of the $p$-quantile given in (2) suggests that it can be estimated by $T_p(\hat{F}_n)$, provided $\hat{F}_n$ approximates $F_0$. We now introduce a method to pointwise estimate $F_0$ that will induce a robust doubly protected estimator of its quantiles; that is to say an estimator that converges
to $T_p(F_0)$ if either the propensity score $\pi$ or the regression function $g$ is consistently estimated. Moreover, as far as the regression function $g(X)$ is estimated in a robust way, the estimator of the quantiles that we propose turns also to be robust.

In the MAR literature, it is known that

$$F_0(y) = \mathbb{E}\{1_{Y \leq y}\} = \mathbb{E}\left\{\frac{A1_{Y \leq y}}{p(X)}\right\} - \mathbb{E}\left\{\frac{A}{p(X)} \cdot r_y(X)\right\} + \mathbb{E}\{r_y(X)\}$$

supposing either $p(X) = \mathbb{P}(A = 1 \mid X)$ or $r_y(X) = \mathbb{P}(Y \leq y \mid X)$ holds. Under model (1), $r_y(X)$ can be estimated with $m^{-1} \sum_{j=1}^{n} A_j 1_{\hat{u}_j \leq y - \hat{\pi}_n(X)}$, where $m$ denotes the number of observed responses. Equation (3) motivated the construction of augmented IPW-type (AIPW) estimators of $F_0$. AIPW estimators were originally proposed by Robins et al. (1994) for the mean and by Zhang et al. (2012) for quantiles in a parametric setting and were also studied by Díaz (2017). The main novelty of our proposal is to estimate the regression function robustly. Let $\hat{\pi}_n(X)$ be a robust estimator of $\pi(X)$.

The robust doubly protected AIPW-type estimator of $F_0$ that we propose is given by

$$\hat{F}_RDP = \frac{1}{C_n} \sum_{i=1}^{n} \frac{A_i \delta_{Y_i}}{\hat{\pi}_n(X_i)} - \frac{1}{C_m} \sum_{i,j=1}^{n} \frac{A_i A_j}{\hat{\pi}_n(X_i) + \hat{u}_j} + \frac{1}{nm} \sum_{i,j=1}^{n} A_j \delta_{\hat{\pi}_n(X_i) + \hat{u}_j},$$

where $\delta_s$ denotes the probability distribution function of the point mass probability measure concentrated at $s$.

The first term in (4) is an IPW estimator, while the third one coincides with the proposal presented in Sued and Yohai (2013). The second term is a correction to obtain consistency when either the propensity score or the regression function is consistently estimated. However, $\hat{F}_RDP$ may be not monotone because it is defined by a linear combination which is not convex. Therefore, it may fall out of the space of distribution functions. Nevertheless, Theorem 2 shows that $T_p(\hat{F}_RDP)$ is well defined and that it converges to $T_p(F_0)$, as far as the propensity score or the regression function is consistently estimated. For instance, any robust nonparametric or parametric method can be used.

Despite the predictive capability of modern learning methods, in many applied fields like biology, medical research, economy, etc., practitioners often prefer parametric models even if they not always predict as well as the former. This is understandable since parametric models are easier to understand and provide a natural framework to interpret the role of the covariates in the model. This last point is of major importance in applied science and justifies the use of parametric models. In such cases, when parametric models are postulated for the regression and the propensity score, doubly protected procedures became a fundamental tool, providing consistent procedures if at
least one of the two models is correct. As with all parametric methods, we recommend
the reader to be cautious when using these methods, verifying that the hypothesis holds
a convenient degree of approximation.

In the case of postulating parametric models \( \pi(X; \gamma) \) and \( g(X; \beta) \) for the propensity
score and the regression function, respectively, we estimate the propensity score with
\( \hat{\pi}_n(X) = \pi(X; \hat{\gamma}_n) \) and the regression function with \( \hat{g}_n(X) = g(X; \hat{\beta}_n) \), where \( \hat{\gamma}_n \)
is a consistent estimator under the postulated parametric model for the propensity
score and \( \hat{\beta}_n \) is an MM estimator. MM estimators for linear regression models were
introduced in Yohai (1987) and generalized for the nonlinear (but still parametric) case
in Fasano et al. (2012). MM estimators combine simultaneously high robustness and
high efficiency. A robust fit of the regression model provides good predictions for the
outcome even when a small percentage of atypical responses are present in the data.
This does not prevent the presence of atypical residuals, but their impact is mitigated
when estimating quantiles, especially for the median.

\section{3 Consistency}

The theoretic results developed in this section do not assume a parametric model for
the propensity score or for the regression function. In fact, in Theorems 1 and 2 stated
below, we will only require that \( \hat{\pi}_n(X) \) and \( \hat{g}_n(X) \) satisfy the following assumptions,
where \( \mathcal{X} \) denotes the support of \( X \):

A1 There exists \( \pi_\infty : \mathbb{R}^p \to (0, 1) \) such that \( \sup_{X \in \mathcal{X}} |\hat{\pi}_n(X) - \pi_\infty(X)| \to 0 \) a.s.

A2 \( \inf_{X \in \mathcal{X}} \pi_\infty(X) = i_\infty > 0 \).

A3 There exists \( g_\infty : \mathbb{R}^p \to \mathbb{R} \) such that, for every compact set \( \mathcal{K} \), \( \sup_{X \in \mathcal{K}} |\hat{g}_n(X) - g_\infty(X)| \to 0 \) a.s.

These are common assumptions in the present missing data framework. Their validity
when parametric models for both the propensity score and the regression function
are postulated is shown in the proof of Theorem 3, where we establish the double
robustness of \( T_p(\hat{F}_{\text{RDSP}}) \), for \( \hat{F}_{\text{RDSP}} \) defined below, in (10). Note that \( \hat{g}_n(X) \) is used both
to predict and to construct residuals \( \hat{\epsilon}_j = Y_j - \hat{g}_n(X_j) \), for \( j \) with \( A_j = 1 \). Let \( \hat{F}_1 \), \( \hat{F}_2 \)
and \( \hat{F}_3 \) denote each term on (4), so that \( \hat{F}_{\text{RDSP}} = \hat{F}_1 - \hat{F}_2 + \hat{F}_3 \) and consider

\[
\hat{F}_1 = \frac{1}{C_n} \sum_{i=1}^n \frac{A_i \delta_{\hat{\epsilon}_i}}{\hat{\pi}_n(X_i)}, \quad \hat{F}_{2a} = \frac{1}{C_n} \sum_{i=1}^n \frac{A_i \delta_{\hat{g}_n(X_i)}}{\hat{\pi}_n(X_i)}, \quad \hat{F}_{3a} = \frac{1}{n} \sum_{i=1}^n \delta_{\hat{g}_n(X_i)} \quad \text{and} \quad \hat{G} = \frac{1}{m} \sum_{j=1}^m A_j \delta_{\hat{\epsilon}_j}.
\]

Thus, \( \hat{F}_{\text{RDSP}} = \hat{F}_1 - \hat{F}_{2a} + \hat{F}_{3a} \), where \( \hat{F}_2 = \hat{F}_{2a} \ast \hat{G} \) and \( \hat{F}_3 = \hat{F}_{3a} \ast \hat{G} \), and \( \ast \) stands
for the convolution operator between two distribution functions while \( C_n \) and \( m \) are
as in (5).

Theorem 2 states that, even though \( \hat{F}_{\text{RDSP}} \) is not necessarily a probability distribution
function, under assumptions A1–A3, it is a doubly protected estimator of \( F_0 \); that
is, it converges uniformly to \( F_0 \) almost surely (a.s.) if either \( \pi_\infty(X) = \pi(X) \) or \( g_\infty(X) = g(X) \). Moreover, it states that \( T_p(\hat{F}_{\text{RDP}}) \) is well defined, with \( T_p \) as in (2). Finally, the aforementioned theorem states that \( T_p(\hat{F}_{\text{RDP}}) \) converges to \( T_p(F_0) \) a.s. if either \( \pi_\infty(X) = \pi(X) \) or \( g_\infty(X) = g(X) \). Therefore, \( T_p(\hat{F}_{\text{RDP}}) \) is a doubly protected estimator of the \( p \)-quantile of the distribution of \( Y \). Robustness is achieved through a robust estimation of the regression function, as shown empirically in the simulation study presented in the following sections.

Let \( C_\infty = \mathbb{E}[\pi(X)/\pi_\infty(X)] \) and consider the probability distribution functions

\[
F_1(y) = \frac{1}{C_\infty} \mathbb{E} \left\{ \frac{\pi(X)}{\pi_\infty(X)} I_{\{Y \leq y\}} \right\}, \quad F_2_\alpha(y) = \frac{1}{C_\infty} \mathbb{E} \left\{ \frac{\pi(X)}{\pi_\infty(X)} I_{\{g_\infty(X) \leq y\}} \right\}, \quad (7)
\]

\[
F_3_\alpha(y) = F_{g_\infty(X)}(y), \quad G(y) = F_{\{Y - g_\infty(X)\}|A = 1}(y). \quad (8)
\]

Define \( F_\infty = F_1 - F_2_\alpha \ast G + F_3_\alpha \ast G \). Theorem 1 indicates in which circumstances \( F_\infty \) coincides with \( F_0 \); in Theorem 2 we prove that \( T_p(\hat{F}_{\text{RDP}}) \) is well defined and it is a doubly robust estimator of the \( p \)-quantile of \( Y \)

**Theorem 1** Assume that the propensity score \( \pi(X) = \mathbb{P}(A = 1 \mid X) \) is equal to \( \pi_\infty(X) \). Then, \( F_1 = F_0, F_2_\alpha = F_3_\alpha \) and therefore, \( F_\infty = F_0 \). Consider now the regression model \( Y = g(X) + u \), with \( u \) independent of \((A, X)\), and assume that \( g(X) = g_\infty(X) \). Then, \( F_1 = F_2_\alpha \ast G, F_0 = F_3_\alpha \ast G \) and consequently \( F_\infty = F_0 \).

**Theorem 2** Assume that \( Y = g(X) + u \), with \( u \) independent of \((A, X)\), and let \( \pi(X) \) denote the propensity score \( \pi(X) = \mathbb{P}(A = 1 \mid X) \). Let \( \{(X_i, A_i, Y_i)\}_{i \geq 1} \) be independent and identically distributed as \((X, A, Y)\). Assume that conditions A1–A3 are satisfied and that the probability distribution function \( G \) of \( \{Y - g_\infty(X)\} \mid (A = 1) \) is continuous. Assume also that either \( g(X) = g_\infty(X) \) or \( \pi(X) = \pi_\infty(X) \). Then,

\[
\sup_y |\hat{F}_{\text{RDP}}(y) - F_0(y)| \to 0 \quad \text{a.s.} \quad (9)
\]

Moreover, \( T_p(\hat{F}_{\text{RDP}}) \) is well defined for every \( p \in (0, 1) \) such that \( F_0 \) is strictly increasing in a neighborhood of \( T_p(F_0) \). Assume again that \( g(X) = g_\infty(X) \) or \( \pi(X) = \pi_\infty(X) \), then \( T_p(\hat{F}_{\text{RDP}}) \to T_p(F_0) \) a.s.

For the sake of simplicity, in what follows we will focus on a semiparametric setting in which a logistic regression model is postulated for the propensity score and a linear function is used to model the regression function. This particular version of our estimator will be denoted \( \hat{\beta}_{\text{mm}} \). The following theorem, derived from Theorem 2, establishes that \( T_p(\hat{F}_{\text{RDP}}) \) is a doubly protected estimator of \( T_p(F_0) \). We introduce the function \( \expit(t) = \{1 + \exp(-t)\}^{-1} \), which is used to define the logistic regression model.

**Theorem 3** Assume that \( Y = g(X) + u \), with \( u \) independent of \((A, X)\). Let \( \{(X_i, A_i, Y_i)\}_{i \geq 1} \) be an i.i.d. sequence, distributed as \((X, A, Y)\). Denote with \( \hat{\gamma}_n \) the ML estimator assuming a logistic regression model \( \pi(X; \gamma) = \expit(\gamma'X) \) for the propensity score \( \mathbb{P}(A = 1 \mid X) \). Let \( \hat{\beta}_n \) be an MM estimator, under the linear
model \( g(X; \beta) = \beta'X \) for the regression function \( g(X) \). Assume that (i) there exist \( \gamma_{\infty} \) and \( \beta_{\infty} \) such that \( \hat{\gamma}_n \to \gamma_{\infty} \) a.s. and \( \hat{\beta}_n \to \beta_{\infty} \) a.s., (ii) \( S_x \) is compact, (iii) the distribution function \( G \) of \( \{Y - X'\beta_{\infty}^*\} \mid (A = 1) \) is continuous, (iv) either \( \mathbb{P}(A = 1 \mid X) = \expit(\gamma_0'X) \), for some \( \gamma_0 \), or \( g(X) = \beta_0'X \), for some \( \beta_0 \). Then, \( T_p(\hat{F}_{\text{RSDP}}) \to T_p(F_0) \) a.s., where

\[
\hat{F}_{\text{RSDP}} = \frac{1}{C_n} \sum_{i=1}^{n} \frac{A_i \delta_{Y_i}}{\pi(X_i; \hat{\gamma}_n)} - \frac{1}{C_n m} \sum_{i,j=1}^{n} \frac{A_j}{\pi(X_i; \hat{\gamma}_n)} A_j \delta_{g(X_j; \hat{\beta}_n) + Y_j - g(X_j; \hat{\beta}_n)} + \frac{1}{nm} \sum_{i,j=1}^{n} A_j \delta_{g(X_i; \hat{\beta}_n) + Y_j - g(X_j; \hat{\beta}_n)}.
\]

(10)

**Remark 1** Both MLE and MM estimators are particular cases of M estimators and therefore, under regularity conditions, their limit point can be characterized, regardless of the validity of the assumed model. In particular, under regularity conditions, the maximum likelihood estimator \( \hat{\gamma}_n \) converges a.s. to \( \gamma_{\infty} = \arg \max_{\gamma} \mathbb{E} \{\log p(X, A, \gamma)\} \), where \( p(X, A, \gamma) = \expit(\gamma'X)^A(1 - \expit(\gamma'X))^{1-A} \), whether or not the postulated model for the propensity score is correctly specified. Similarly, whether or not the regression model is correctly specified, MM estimators converge, under regularity conditions, to

\[
\beta_{\infty} = \arg \min_{\beta} \mathbb{E} \{\rho \{(Y - X\beta)/\sigma_{\infty}\} \mid A = 1\},
\]

for some \( \sigma_{\infty} > 0 \); see Theorems 2 and 3 in Fasano et al. (2012).

Recently, Bianco et al. (2018) have proposed an estimator of the marginal distribution of \( Y \) for the case in which missing values occur also in the vector of covariates \( X \). Unfortunately, it is not possible to generalize their argument for our doubly protected proposal.

Under the assumptions of Theorem 3 and further regularity conditions, it can be derived that \( \sqrt{n}(T_p(\hat{F}_{\text{RSDP}}) - T_p(F_0)) \) converges in distribution to a normal law assuming that both \( \hat{\gamma}_n \) and \( \hat{\beta}_n \) satisfy a linear expansion of the form

\[
\sqrt{n}(\hat{\gamma}_n - \gamma_{\infty}) = n^{-1/2} \sum_{i=1}^{n} IF_{\text{PROP}}(X_i, A_i),
\]

\[
\sqrt{n}(\hat{\beta}_n - \beta_{\infty}) = n^{-1/2} \sum_{i=1}^{n} IF_{\text{REG}}(X_i, A_i, Y_i)
\]

(11)

with finite-variance and zero-mean influence functions, that is \( \mathbb{E}\{IF_{\text{PROP}}(X_i, A_i)\} = 0 \), \( \mathbb{E}\{IF_{\text{REG}}(X_i, A_i, Y_i)\} = 0 \), \( \mathbb{V}\{IF_{\text{PROP}}(X_i, A_i)\} < \infty \) and \( \mathbb{V}\{IF_{\text{REG}}(X_i, A_i, Y_i)\} < \infty \).

Inference can be done using bootstrap techniques; asymptotic normal intervals with bootstrap estimated standard errors or quantile confidence intervals are among the most popular options.
A sketch of the proof of the asymptotic normality of $\sqrt{n} \{ TP(\hat{F}_{RSDP}) - TP(F_0) \}$ is presented in the Supplementary Material. A rigorous one is beyond the scope of this work. It is closely related to the proof of the asymptotic normality of the estimator presented in Sued and Yohai (2013), included in the Supplementary Material of the mentioned work.

### 4 Monte Carlo study

In this section we report the results of a Monte Carlo study where we analyze the performance of the estimators proposed in this work, as compared to the two other doubly protected estimators of the quantiles described in the introduction. In this and the following sections, we will note by $\text{RDP}$ the parametric doubly protected estimator proposed by Zhang et al. (2012); the targeted maximum Likelihood estimator proposed in Díaz (2017) will be denoted by $\text{TML}$. We also consider the estimator proposed in Sued and Yohai (2013) (SY), which is already robust, semiparametric and consistent when the regression model is well specified but inconsistent otherwise. This inconsistency motivates the present proposal that is robust, semiparametric and doubly protected.

As already mentioned, we use $\text{RSDP}$ to denote our proposed estimator $TP(\hat{F}_{\text{RSDP}})$, for $\hat{F}_{\text{RSDP}}$ defined in (10).

We consider a random vector $(X, A, Y) \sim H_0$, where $X = (1, X_1, X_2)$ with $(X_1, X_2)$ being a bivariate standard normal random vector of covariates, that is, $(X_1, X_2) \sim \mathcal{N}(0, I)$, $A$ is a binary variable following a Bernoulli distribution with $P(A = 1|X) = \pi(X) = \expit((1, X_1, X_2)\beta_0)\), with $\gamma_0 = (0, 0, 1, -1)$; the outcome $Y$ satisfies $Y = (1, X_1, X_2)\beta_0 + u$, where $\beta_0 = (0, -3, 2)$ and $u$ is independent of $(X, A)$. We consider different distributions of the error term $u$, namely standard normal and Student’s t distribution with one (Cauchy distribution) and three degrees of freedom to explore the capability of our $\text{RSDP}$ to adapt to any error distribution. In all three cases, this will be called the central model and samples $(X_i, A_i, Y_i), i = 1, \ldots, n$ generated in this way will be called clean samples.

To investigate the resistance of the estimators to the presence of atypical observations in the regression model, we only contaminate the central model with normal errors given that $\text{RDP}$ is designed for this situation. Let $K_0$ denote the marginal distribution of $(X, Y)$ when $(X, A, Y) \sim H_0$, with normal errors. We consider the outlier contamination model for the marginal law of $(X, Y)$ proposed by Tukey–Huber, with a mass point contamination. In our setting, to generate contaminated data we assume that $(X, Y)$ follows a distribution $K_{(X_0, y_0)} = 0.9K_0 + 0.1\delta_{(x_0, y_0)}$. This means that, with probability 0.9, $(X, Y)$ is sampled from $K_0$ while, with probability 0.1, $(X, Y)$ is equal to $(x_0, y_0)$. In our simulation we consider high leverage outliers of the form $(x_0, y_0)$ with $x_0 = (1, -5, 0)$ and $y_0$ in the grid $G = \{ y \in \mathbb{Z} : -50 \leq y \leq 50 \}$. The missing indicator $A$ is generated as in the case of clean data. We use $H_{(x_0, y_0)}$ to denote the joint distribution of $(X, A, Y)$ under the described data generating process.

For clean samples we vary the sample size in order to empirically assess the consistency of the estimators, taking $n = 100, 500$ and $1000$. Instead, in the case of contaminated data, the sample size is fixed in $n = 100$. 

 Springer
In all cases, we generate $N = 1000$ samples of size $n$ and we compute estimators of the first, second (i.e., the median) and third quartile, noted $\eta_p = T_p(F_0)$, for $p = 0.25, 0.50,$ and $0.75$, respectively. The population values of these quantiles were approximated by the Monte Carlo method, generating a sample of size $10^6$ and computing its corresponding quantile at the empirical distribution of the complete set of responses. In the case of the median or second quartile, instead of using this approximate value we used the exact one, $\eta_{0.5} = 0.$

We now specify three different situations that will allow us to compare the behavior of our estimator to other doubly protected proposals.

S.1 Both the model for the propensity score and the model for the regression function are well specified for the central model $H_0$.

S.2 The model for the propensity score is well specified, while the model for the regression function is misspecified for the central model $H_0$. More precisely, we fit an incorrect linear model for the regression function, using just the covariate $X_1$; that is, the covariate $X_2$ is omitted.

S.3 The model for the propensity score is misspecified, while the model for the regression function is well specified for the central model $H_0$. In this case, we fit an incorrect model for the propensity score, using an expit model where the covariate $X_2$ is omitted.

Estimators $\text{SY}$, $\text{PDP}$ and $\text{RSDP}$ are computed by evaluating $T_p$ at three different estimators of $F_0$: $\hat{F}_{\text{SY}}$, $\hat{F}_{\text{PDP}}$ and $\hat{F}_{\text{RSDP}}$, respectively, while estimator $\text{TML}$ is computed using the package causalquantile, downloaded from the author’s GitHub repository at https://github.com/idiazst. For the definitions of $\hat{F}_{\text{SY}}$, and $\hat{F}_{\text{PDP}}$, see Sued and Yohai (2013) and Zhang et al. (2012), respectively.

The MM estimator required for computing $\text{SY}$ and $\text{RSDP}$ uses $\rho_0$ and $\rho_1$ in the Tukey bisquare family with $k_0 = 1.57$ and $k_1 = 3.44$, respectively, and $\delta = 0.5$. See Maronna et al. (2018). The estimator of $\sigma$ necessary to compute $\text{PDP}$ is taken to be the classical estimator of error standard deviation $\hat{\sigma} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{\beta}'X_i)^2/(n-1)}$, where $\hat{\beta}$ is the least-squares estimator.

Empirical mean square errors are computed by

$$\text{MSE} = N^{-1} \sum_{j=1}^{N} \left\{ T_p(\hat{F}_{j,*}) - T_p(F_0) \right\}^2,$$

biases by

$$\text{BIAS} = N^{-1} \sum_{j=1}^{N} T_p(\hat{F}_{j,*}) - T_p(F_0),$$

and standard deviations by

$$\text{SD} = \sqrt{N^{-1} \sum_{j=1}^{N} \left\{ T_p(\hat{F}_{j,*}) - N^{-1} \sum_{i=1}^{N} T_p(\hat{F}_{i,*}) \right\}^2},$$

\( \text{\copyright } \) Springer
where \( j, j = 1, \ldots, N \), refers to the estimation based on the \( j \)th, replication while \( * \) stands for each possible procedure.

The results of the simulations are presented in Tables 1, 2, 3 and 4 and Fig. 1. These tables show that SY estimator is the most efficient one when the regression model is correctly specified but, as expected, it is not even consistent when it is not. This indicates the need of incorporating doubly protected estimators. On the other hand, we can see that, when the error term has a standard normal or a \( t_3 \) distribution, all the estimators have a similar performance in clean samples. However, when the error term has a Cauchy distribution, these results show that our proposed estimator (RSDP) behaves as consistent estimators, while PDP and TML do not. This is due to the semiparametric nature of RSDP estimator. Estimators PDP and TML can probably be adapted so that they are consistent in this simulation setting, using the correct distribution of the error term. However, in practice, the real distribution of the errors is unknown. Our proposed estimator adapts automatically to any error distribution.

Figure 1 gives the mean squared errors of the three doubly protected estimators for contaminated samples as a function of \( y_0 \) in situations S.1, S.2 and S.3. Note that, for contaminated samples, i.e., generated as \( H(x_0, y_0) \), the presence of outliers precludes the correct specification of the regression function. Nevertheless, in Fig. 1 and Tables 1, 2, 3 and 4, we use the legend “OR correct” to allude the case where the central distribution \( K_0 \) satisfies the model proposed for the regression function. This means that the regression function is correctly specified, except for presence of outliers.

We remark the robust behavior of our proposed estimator. This is evidenced by the fact that the MSE of RSDP is bounded, while the MSE of PDP and TML seem to grow indefinitely, especially in situation S.3. This is as expected, since in this situation the consistency of the procedure relies on the correct estimation of the regression function since the model for the propensity score is incorrect.

Table 4 gives the maximum mean square error under 10% of contamination for values of \( y_0 \) in the grid \( G \). We give this table as another measure of the robustness of the estimators. It shows that the estimator with the smallest maximum means squared error is RSDP in all the cases but one, in which the propensity score is consistently estimated.

Finally, we compare the performance of the three estimators in a causality framework. We generate samples following the simulation setup introduced in Kang and Schafer (2007) that was further considered by Porter et al. (2011) and used by Díaz (2017) to estimate the causal effect of a treatment variable on the median of the potential outcomes. In this setting, we also generate \( N = 1000 \) samples of size \( n = 100, 500 \) and 1000 and we consider two different distributions for the error term \( u \): standard normal and Cauchy. The results are summarized in Table 5. We again can see that all three estimators give similar results when the error term has a standard normal distribution but, when the error term has a Cauchy distribution, RSDP still behaves as a consistent estimator while the other estimators do not.

Similar figures and tables for the biases and standard deviations are relegated to the supplementary material due to space limitations.

As another measure of robustness, we compute the sensitivity curves. The sensitivity curve of an estimator \( \hat{\mu} \) for an observed sample \( \mathcal{Z}_n = (z_1, \ldots, z_n) \), where \( z_i = (x_i, a_i, y_i) \), is defined as \( c(z_0) = \hat{\mu}(z_1, \ldots, z_n, z_0) - \hat{\mu}(z_1, \ldots, z_n) \). It measures
| PS   | OR     | First quartile | Median       | Third quartile |
|------|--------|----------------|--------------|---------------|
|      |        | 100  | 500  | 1000 | 100  | 500  | 1000 | 100  | 500  | 1000 |
| RSDP | Correct | 0.941| 0.157| 0.072| 0.622| 0.125| 0.057| 1.114| 0.202| 0.096|
| PDP  | Correct | 1.444| 0.169| 0.080| 0.722| 0.139| 0.064| 1.475| 0.236| 0.115|
| TML  | Correct | 1.139| 1.293| 0.189| 0.735| 0.407| 0.789| 1.456| 0.281| 0.989|
| RSDP | Incorrect | 0.673| 0.136| 0.068| 0.508| 0.092| 0.050| 0.678| 0.125| 0.066|
| PDP  | Incorrect | 0.937| 0.484| 0.446| 0.850| 0.586| 0.584| 1.214| 0.698| 0.694|
| TML  | Incorrect | 0.963| 0.619| 1.294| 0.859| 1.001| 0.925| 1.174| 0.919| 2.644|
| RSDP | Correct | 1.059| 0.181| 0.081| 0.815| 0.147| 0.066| 1.605| 0.251| 0.120|
| PDP  | Incorrect | 1.604| 0.178| 0.083| 0.829| 0.145| 0.066| 1.622| 0.246| 0.117|
| TML  | Incorrect | 1.126| 1.418| 0.105| 0.745| 0.324| 0.675| 1.429| 0.787| 0.156|
| SY   | Correct  | 0.454| 0.079| 0.040| 0.321| 0.058| 0.031| 0.470| 0.089| 0.049|
| SY   | Incorrect | 1.106| 0.738| 0.684| 1.050| 0.838| 0.821| 1.346| 1.045| 1.032|

Table 1 Empirical MSE of estimators of the mean of Y for clean samples with $t_1$ errors
| PS | OR    | First quartile | Median | Third quartile |
|----|-------|----------------|--------|---------------|
|    |       | 100 500 1000   | 100 500 1000 | 100 500 1000 |
| RSDP Correct | Correct | 0.414 0.073 0.038 | 0.377 0.073 0.036 | 0.486 0.095 0.046 |
| RSDP Correct | Correct | 0.418 0.074 0.039 | 0.387 0.073 0.037 | 0.519 0.098 0.046 |
| RSDP Correct | Correct | 0.453 0.077 0.039 | 0.444 0.069 0.034 | 0.536 0.076 0.038 |
| RSDP Incorrect | Correct | 0.436 0.076 0.040 | 0.361 0.071 0.036 | 0.375 0.080 0.040 |
| RSDP Correct | Incorrect | 0.399 0.070 0.036 | 0.351 0.069 0.036 | 0.423 0.089 0.045 |
| SY Correct | Correct | 0.277 0.049 0.023 | 0.224 0.040 0.020 | 0.259 0.053 0.026 |
| SY Incorrect | Correct | 0.908 0.673 0.639 | 0.995 0.817 0.796 | 1.213 0.988 0.967 |
| PS   | OR       | First quartile |      |      |      |      |      |      |      |      |      |      |      |
|------|----------|----------------|------|------|------|------|------|------|------|------|------|------|------|
|      |          | 100  | 500  | 1000 | 100  | 500  | 1000 | 100  | 500  | 1000 | 100  | 500  | 1000 |
| RSDP | Correct  | 0.355 | 0.063 | 0.033 | 0.302 | 0.062 | 0.029 | 0.377 | 0.076 | 0.041 |      |      |      |
| PDP  | Correct  | 0.349 | 0.063 | 0.033 | 0.314 | 0.061 | 0.029 | 0.403 | 0.076 | 0.040 |      |      |      |
| TML  | Correct  | 0.340 | 0.064 | 0.033 | 0.312 | 0.061 | 0.029 | 0.394 | 0.077 | 0.041 |      |      |      |
| RSDP | Incorrect | 0.340 | 0.066 | 0.035 | 0.278 | 0.057 | 0.027 | 0.298 | 0.060 | 0.033 |      |      |      |
| PDP  | Incorrect | 0.336 | 0.066 | 0.035 | 0.281 | 0.056 | 0.027 | 0.299 | 0.061 | 0.033 |      |      |      |
| TML  | Incorrect | 0.315 | 0.059 | 0.032 | 0.279 | 0.055 | 0.027 | 0.328 | 0.068 | 0.036 |      |      |      |
| RSDP | Correct  | 0.478 | 0.092 | 0.046 | 0.450 | 0.093 | 0.042 | 0.774 | 0.124 | 0.071 |      |      |      |
| PDP  | Correct  | 0.523 | 0.095 | 0.047 | 0.529 | 0.098 | 0.044 | 0.956 | 0.130 | 0.075 |      |      |      |
| TML  | Correct  | 0.391 | 0.071 | 0.036 | 0.347 | 0.068 | 0.032 | 0.454 | 0.087 | 0.050 |      |      |      |
| SY   | Correct  | 0.238 | 0.044 | 0.022 | 0.198 | 0.037 | 0.020 | 0.223 | 0.048 | 0.024 |      |      |      |
| SY   | Incorrect | 0.844 | 0.663 | 0.643 | 0.977 | 0.819 | 0.806 | 1.214 | 1.002 | 0.984 |      |      |      |

Table 3 Empirical MSE of estimators of the mean of $Y$ for clean samples with normal errors.
Table 4  Maximum mean squared error of estimators of the mean of \( Y \) under 10% of outlier contamination and the regression model with normal errors

|       | PS Correct | OR Correct | First quartile | Median | Third quartile |
|-------|------------|------------|----------------|--------|---------------|
| RSDP  | Correct    | Correct    | 1.879          | 0.709  | 1.661         |
| PDP   | Correct    | Correct    | 2.275          | 1.030  | 2.257         |
| TML   | Correct    | Correct    | 2.084          | 0.927  | 2.142         |
| RSDP  | Incorrect  | Correct    | 2.083          | 0.758  | 1.646         |
| PDP   | Incorrect  | Correct    | 3.157          | 1.919  | 1.864         |
| TML   | Incorrect  | Correct    | 3.382          | 1.919  | 1.859         |
| RSDP  | Correct    | Incorrect  | 2.115          | 0.933  | 2.372         |
| PDP   | Correct    | Incorrect  | 2.333          | 1.138  | 2.735         |
| TML   | Correct    | Incorrect  | 2.171          | 0.940  | 2.173         |

The mean square errors of PDP and TML are unbounded, increasing as \( y_0 \) increases to \( \infty \) or decreases to \( -\infty \). The maximum attained in the considered grid is informed.

Fig. 1  Estimated mean squared error of doubly robust estimators of the quantiles. The rows correspond to the first, second and third quartile and the columns to situations S.1, S.2 and S.3, respectively. Black lines correspond to PDP, red to RSDP and blue to TML (color figure online).

the effect a single outlier on the estimator \( \hat{\mu} \) based on an observed sample. This function, properly standardized, can be considered an empirical analogue of the influence function. See Maronna et al. (2018). Figure 2 gives plots of the sensitivity curves of estimators RSDP, PDP and TML for the first, second and third quartiles in situations S.1, S.2 and S.3 for a fixed data set \( Z_n \), with \( n = 1000, z_0 = (x_0, a_0, y_0) \) for \( x_0 = (1, -5, 0) \), \( a_0 = 1 \) and \( y \) in the grid \( G \) described before.
The closer the sensitivity curve is to the horizontal line $y = 0$, the less sensitive the estimator is to a single outlier. Figure 2 shows that the sensitivity curve of RSDP is much closer to zero than the other two. We remark that, in contaminated samples, the OR model can never be completely correctly specified, since the presence of the outlier means that there is one observation that does not follow the assumed model. The legend “OR correct” in Fig. 2 means that the OR model is correctly specified except for the outliers.

5 Example: hospital data

We consider a sample of $n = 100$ patients hospitalized in a Swiss hospital during 1999 for medical back problems. We study the relationship between $Y$, the cost of stay (Cost, in thousands of Swiss francs) and some explanatory variables that are available on 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), discharge destination ($1 =$ home; $0 =$ another health institution). This information plus an intercept constitute the vector of covariates $X$. This data set has been analyzed in Marazzi and Yohai (2004) and has no missing values. In order to study the performance of our proposed estimators, we artificially delete some of the responses and compute the estimators in the sample with missing values. We repeat this procedure $N = 1000$ times. In each replication we generate a sample of dichotomous variables $A_1, \ldots, A_n$ according to the mechanism $\mathbb{P}(A_i = 1 \mid X_i) = \expit(0.1 \times \text{LOS}_i - 1.1)$. The responses with corresponding $A_i = 0$ are deleted from the sample and considered missing. In this way, the proportion of missing responses is approximately 0.5.

For each replication, we compute estimators of the median cost of stay and its first and third quartiles by methods PDP, RSDP and TML. To evaluate the performance of
Fig. 2 Sensitivity curves. The rows correspond to the first, second and third quartile and the columns to situations S.1, S.2 and S.3, respectively. Black lines correspond to PDP, red to RSDP and blue to TML.

Each method, we compare these estimates to the corresponding quantiles computed at the empirical distribution of the entire set of responses (Cost); i.e., the quantiles of $y_1, \ldots, y_n$. The values of these quantiles are $\eta_{0.25} = 5.13$, $\eta_{0.5} = 9.69$ and $\eta_{0.75} = 14.12$. We compute the MSE of each procedure as in Eq. (12), but replacing $\eta_p$ for $T_p$. 
| PS          | OR       | First quartile |          | Median       |          | Third quartile |          |
|------------|----------|----------------|----------|--------------|----------|----------------|----------|
| Correct    | Correct  | 0.1320         | 0.3157   | 0.3011       | 0.1251   | 0.1727         | 0.1703   |
| Correct    | Incorrect| 0.1435         | 0.1599   | 0.2100       | 0.1276   | 0.1462         | 0.1408   |
| Incorrect  | Correct  | 0.1419         | 0.3040   | 0.2101       | 0.1172   | 0.2144         | 0.1670   |

Table 6: MSE of estimators for hospital data
An analysis of the linear regression fit with the complete data set shows that all six available variables are relevant to predict cost and that no transformations are necessary; for this reason we consider this the “correct” model, for both the PS and the OR. To compare the fit with the one obtained if either model is misspecified, we also consider “incorrect” models, which include all six covariates, but LOS is transformed to logLOS.

The results are summarized in Table 6. \text{RSDP} gives much better results than \text{PDP} and \text{TML} in most of the situations considered. The exception is the estimator of the third quartile when the propensity score is correctly specified. Note that in this example the propensity score is known exactly, since the missing values are generated. In real data sets, however, it often happens that neither of the two models is exactly correct and in these cases using our proposed \text{RSDP} estimator is recommended. The largest difference in favor of \text{RSDP} is obtained whenever the propensity score is incorrectly specified, since in these cases, all three estimators rely on the outcome regression model, which is not known exactly.

6 Discussion

We have presented an estimator of the quantiles of the distribution of a variable \( Y \) in the missing at random setting. The estimator is doubly protected, in the sense of being consistent as far as either the propensity score or the regression function is consistently estimated. In particular, we have considered a generalized linear model for the propensity score and a linear model for the regression function, but no assumptions have been made regarding the distribution of the error term. Besides, a robust fit for the regression model has been used. In this way, we have obtained a procedure that mitigates the impact of atypical observations regarding the regression model. These two last properties constitute a novelty with respect to other estimators reported in the literature.

Combining double protection and robustness implies studying situations in which neither the model for the propensity score nor the outcome regression model is completely correct due to the presence of outlier contamination. In this situation, one cannot expect to obtain consistent estimators. Our robust procedure, besides being consistent when either the propensity score or the regression function is consistently estimated and there are not any outliers, it provides estimators of the quantiles that are reliable even if there are outliers.

Outlier contamination is quite frequent in real data sets, and this is why, we think that the methods introduced in this work will be very useful for scientists with missing data or causality problems.

Acknowledgements We would like to thank Dr. Alfio Marazzi for the data set in the example and the editor and referees for their comments and suggestions which have helped us to improve this paper.

Funding This work was supported by Secretaria de Ciencia y Tecnica, Universidad de Buenos Aires (Grant Nos. 20020150200110BA, 20020130100279BA), Agencia Nacional de Promoción Científica y Tecnológica (Grant No. pict 2014-0351).
Appendix

Proof of Theorem 1 If $\pi(X) = \pi_\infty(X)$, then $\pi(X)/\pi_\infty(X) = 1$ and $C_\infty = 1$. Therefore, $F_1 = F_0$, $F_{2a} = F_{3a}$ and $F_\infty = F_0$.

If $Y = g(X) + u$, with $u$ independent of $(A, X)$ and $g(X) = g_\infty(X)$, then $F_{3a}$ is the distribution function of $g(X)$ and $G$ is the distribution function of $u$. Therefore, $F_{3a} \ast G$ is the distribution function of $g(X) + u = Y$, that is to say $F_{3a} \ast G = F_0$. On the other hand, let $Z$ be a random variable, independent of $u$, with distribution function $F_{2a}$, then $F_{2a} \ast G$ is the distribution function of $Z + u$, which, by definition, is equal to

$$P(Z + u \leq y) = P(Z \leq y - u) = \frac{1}{C_\infty} \mathbb{E} \left\{ \pi(X) \pi_\infty(X) I_{[g(X) \leq y - u]} \right\}$$

$$= \frac{1}{C_\infty} \mathbb{E} \left\{ \frac{\pi(X)}{\pi_\infty(X)} I_{[g(X) + u \leq y]} \right\} = F_1(y).$$

Thus, $F_\infty = F_0$ also in this case. $\square$

The following five lemmas will be used to prove Theorem 2. Recall that $\tilde{F}_1$ and $\tilde{F}_{2a}$, defined in (6), are indeed random sequences of cumulative distribution functions based on sample of size $n$ (which we omit in the notation).

Lemma 1 Consider $\tilde{F}_1$ and $F_1$, defined in (6) and (7), respectively. Under assumptions A1 and A2, $\tilde{F}_1$ converges to $F_1$ uniformly, a.s., that is $\mathbb{P} \left( \sup_y |\tilde{F}_1(y) - F_1(y)| \to 0 \right) = 1$

Proof We show first that $C_n/n \to C_\infty$ a.s. To do so, note that we can write

$$\frac{C_n}{n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{A_i}{\pi_\infty(X_i)} - \frac{A_i}{\pi_\infty(X_i)} \right\} + \frac{1}{n} \sum_{i=1}^{n} A_i \pi_\infty(X_i).$$

(13)

By the law of large numbers, the second term in (13) converges a.s. to

$$\mathbb{E} \left\{ \frac{A}{\pi_\infty(X)} \right\} = \mathbb{E} \left\{ \frac{1}{\pi_\infty(X)} \mathbb{E} (A \mid X) \right\} = \mathbb{E} \left\{ \frac{\pi(X)}{\pi_\infty(X)} \right\} = C_\infty.$$

It remains to prove that the first term in (13) converges to zero a.s. Now, under conditions A1 and A2, given $\varepsilon \in (0, 1)$ there exists $n_0$ such that $|\pi_\infty(X) - \pi_n(X)| < \varepsilon \pi_\infty$ for all $n \geq n_0$, and therefore, $(1 - \varepsilon) \pi_\infty \leq \pi_n(X)$ for such $n$, implying that

$$\frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{\pi_\infty(X_i) \pi_n(X_i)} \leq \frac{1}{n} \frac{1}{(1 - \varepsilon) \pi_\infty} \sum_{i=1}^{n} A_i |\pi_\infty(X_i) - \pi_n(X_i)| < \frac{\varepsilon}{(1 - \varepsilon) \pi_\infty}.$$  

(14)

and then we obtain the announced result.
Second, we prove that

$$
P \left\{ \lim_{n \to \infty} \sup_y \left| \tilde{F}_1(y) - \frac{1}{C_\infty} \frac{1}{n} \sum_{i=1}^{n} A_i I\{Y_i \leq y\} \pi_\infty(X_i) \right| = 0 \right\} = 1. \quad (15)$$

To prove (15), notice that adding and subtracting \( \left( \frac{nC_\infty}{n} - \frac{1}{n} \sum_{i=1}^{n} A_i I\{Y_i \leq y\} / \hat{\pi}_n(X_i) \right) \), we get

$$
\left| \tilde{F}_1(y) - \frac{1}{C_\infty} \frac{1}{n} \sum_{i=1}^{n} A_i I\{Y_i \leq y\} \pi_\infty(X_i) \right| \leq \left| \left( \frac{C_n}{n} - 1 \right) \frac{1}{n} \sum_{i=1}^{n} A_i I\{Y_i \leq y\} / \hat{\pi}_n(X_i) \right| + \frac{1}{C_\infty} \frac{1}{n} \sum_{i=1}^{n} A_i |\tilde{\pi}(X_i) - \pi_\infty(X_i)|. \quad (16)
$$

Neither of the two terms in (16) depend on \( y \), and they both converge to zero under A1-A2; the convergence of the first term follows from the convergence of \( C_n/n \) to \( C_\infty \) a.s., while the convergence of the second one has already been proved in (14). This proves (15).

Finally, using arguments similar to those in the proof of the Glivenko–Cantelli theorem (see, for instance, Theorem 19.1 in Van der Vaart 2000), it can be shown that

$$
P \left\{ \lim_{n \to \infty} \int f \, d\tilde{F}_2a = \int f \, dF_2a, \quad \forall f \in \mathcal{C}_{\text{buc}} \right\} = 1. \quad (17)
$$

The result follows combining (15) and (17). \( \square \)

Henceforth, we use \( G_n \xrightarrow{w} G \) to denote weak convergence of cumulative distribution functions.

**Lemma 2** Consider \( \tilde{F}_{2a} \) and \( F_{2a} \), defined in (6) and (7), respectively. Under assumptions A1–A3, it holds that \( \tilde{F}_{2a} \) converges weakly to \( F_{2a} \) a.s., i.e.,

$$
P \left( \tilde{F}_{2a} \xrightarrow{w} F_{2a} \right) = 1. \quad (18)
$$

**Proof** Let \( \mathcal{C}_{\text{buc}} \) denote the set of functions \( f : \mathbb{R} \to \mathbb{R} \) bounded and uniformly continuous. In order to prove the lemma, we will show that

$$
P \left( \lim_{n \to \infty} \int f \, d\tilde{F}_{2a} = \int f \, dF_{2a}, \quad \forall f \in \mathcal{C}_{\text{buc}} \right) = 1. \quad (19)
$$

Let

$$
\tilde{F}_3(y) = \frac{1}{C_n} \sum_{i=1}^{n} A_i \delta_{g_a}(X_i)(y) \pi_\infty(X_i) \quad \text{and} \quad \tilde{F}_4(y) = \frac{1}{C_n} \sum_{i=1}^{n} A_i \delta_{g_\infty}(X_i)(y) \pi_\infty(X_i).
$$
Note that both $\tilde{F}_3$ and $\tilde{F}_4$ defined above are sequences of random functions; however, we omit $n$ in the notation for simplicity.

Fix $f \in C_{buc}$. Defining $I_1(f) = \left| \int f d\tilde{F}_2 - \int f d\tilde{F}_3 \right|$, $I_2(f) = \left| \int f d\tilde{F}_3 - \int f d\tilde{F}_4 \right|$, and $I_3(f) = \left| \int f d\tilde{F}_4 - \int f dF_2 \right|$, we get that

$$\left| \int f d\tilde{F}_2 - \int f dF_2 \right| \leq I_1(f) + I_2(f) + I_3(f).$$

(20)

Let us now consider each of these three terms. Since $f$ is bounded, using arguments similar to those in the proof of Lemma 1, we have that under A1 and A2

$$\mathbb{P} \left( \lim_{n \to \infty} \left| \int f d\tilde{F}_2 - \int f dF_2 \right| = 0, \forall f \in C_{buc} \right) = 1.$$

(21)

To deal with $I_2(f)$, notice that

$$I_2(f) = \left| \frac{1}{C_n} \sum_{i=1}^{n} A_i f(\hat{g}_n(X_i)) - \frac{1}{C_n} \sum_{i=1}^{n} A_i f(g_{\infty}(X_i)) \right| \leq \frac{n}{C_n n \pi_\infty} \sum_{i=1}^{n} \left| f(\hat{g}_n(X_i)) - f(g_{\infty}(X_i)) \right|.$$

Since $f$ is uniformly continuous, given $\varepsilon > 0$, there exists $\delta$ such that $|u_1 - u_2| < \delta$ implies $|f(u_1) - f(u_2)| < \varepsilon$. Take $K$ large and consider the compact set $\mathcal{K} = \{||X|| \leq K\}$. For $n$ large enough, invoking now A3, we get that $\sup_{X \in \mathcal{K}} |\hat{g}_n(X) - g_{\infty}(X)| < \delta$ and therefore, the right-hand side of (22) is smaller than

$$\frac{n}{C_n} \left( \varepsilon + \frac{1}{n \pi_\infty} \sum_{i=1}^{n} \left| f(\hat{g}_n(X_i)) - f(g_{\infty}(X_i)) \right| \right),$$

(22)

which implies that

$$\mathbb{P} \left( \lim_{n \to \infty} \left| \int f d\tilde{F}_3 - \int f d\tilde{F}_4 \right| = 0, \forall f \in C_{buc} \right) = 1.$$

(23)

It remains to show that

$$\mathbb{P} \left( \lim_{n \to \infty} \int f d\tilde{F}_4 = \int f dF_2, \forall f \in C_{buc} \right) = 1$$

(24)

Notice that, as in Lemma 1, using arguments similar to those in the proof of the Glivenko–Cantelli theorem, we have that

$$\mathbb{P} \left( \limsup_{n \to \infty} \left| \int f d\tilde{F}_4 - \int f dF_2 \right| = 0 \right) = 1,$$

(25)
and therefore
\[ P\left( \lim_{n \to \infty} \frac{1}{C_n} \sum_{i=1}^{n} \frac{A_i \delta_{g_{\infty}(X_i)}(y)}{\pi_{\infty}(X_i)} = \frac{1}{C_{\infty}} \mathbb{E} \left\{ \frac{A I_{[g_{\infty}(X) \leq y]}(X)}{\pi_{\infty}(X)} \right\}, \forall y \in \mathbb{R} \right) = 1, \] (26)

where \( \tilde{C}_n = \sum_{i=1}^{n} \frac{A_i}{\pi_{\infty}(X_i)} \). Both of the sequences as the limit function presented in (26) are cumulative distribution functions. By the MAR assumption,
\[ \frac{1}{C_{\infty}} \mathbb{E} \left\{ \frac{A I_{[g_{\infty}(X) \leq y]}(X)}{\pi_{\infty}(X)} \right\} = F_{2a}(y) \] (27)
and, therefore, (26) implies that
\[ P\left( \lim_{n \to \infty} \frac{1}{C_n} \sum_{i=1}^{n} \frac{A_i f_{(g_{\infty}(X_i))}}{\pi_{\infty}(X_i)} = \int f dF_{2a}, \forall f \in C_{buc} \right) = 1. \] (28)

Finally, since \( \tilde{C}_n/C_n \to 1 \), we conclude that (24) holds. The result stated in the lemma follows from combining (20), (21), (23) and (24).

The following lemma was proved in Sued and Yohai (2013), as a part of Theorem 1.

**Lemma 3** Consider \( \tilde{F}_{3a} \) and \( \tilde{G} \), defined in (6) and \( F_{3a} \) and \( G \) defined in (8). Under assumption A3, \( \tilde{F}_{3a} \) converges weakly to \( F_{3a} \) a.s. and also \( \tilde{G} \) converges weakly to \( G \) a.s., i.e.,
\[ P\left( \tilde{F}_{3a} \overset{w}{\to} F_{3a} \right) = 1 \text{ and } P\left( \tilde{G} \overset{w}{\to} G \right) = 1. \]

As announced in Sect. 3, we will now show that the functional \( T_p \), presented in (2), can be defined over an enlarged family of functions, which includes cumulative distribution functions, preserving its continuity.

**Lemma 4** Consider a distribution function \( F : \mathbb{R} \to [0, 1] \) and \( p \in (0, 1) \) such that there exists a unique value \( y_p \) with \( F(y_p) = p \), and so \( T_p(F) = y_p \), for \( T_p \) defined in (2). Let \( F_n : \mathbb{R} \to \mathbb{R}, n \geq 1 \), be a sequence of functions such that
1. \( \lim_{y \to -\infty} F_n(y) = 0 \) and \( \lim_{y \to +\infty} F_n(y) = 1 \).
2. \( F_n \) converges uniformly to \( F \).

Then \( T_p \) can be defined at \( F_n \) and \( \lim_{n \to \infty} T_p(F_n) = T_p(F) \).

**Proof** Let \( A_{n,p} = \{ y \in \mathbb{R} : F_n(y) \geq p \} \). By the assumptions of the lemma, \( \lim_{y \to +\infty} F_n(y) = 1 \), and therefore, \( A_{n,p} \) is not empty. Since \( \lim_{y \to -\infty} F_n(y) = 0 \) we conclude that \( A_{n,p} \) is bounded from below, and therefore \( T_p(F_n) = \inf A_{n,p} \) is well defined.

Given \( \varepsilon > 0 \), let \( \delta = \min \left\{ (F(y_p + \varepsilon) - F(y_p))/2, (F(y_0) - F(y_p - \varepsilon))/2 \right\} \). By the assumptions of the lemma, \( \delta > 0 \). Now, the uniform convergence of \( F_n \) to \( F \)
guarantees that there exists \( n_0 \) such that \( \sup_{y \in \mathbb{R}} |F_n(y) - F(y)| \leq \delta \), for all \( n \geq n_0 \). In particular, the following inequalities hold

\[
\sup_{y < y_p - \varepsilon} F_n(y) < F(y_p - \varepsilon) + \delta \leq F(y_p) - 2\delta + \delta \leq p - \delta
\]

(29)

\[
F_n(y_p + \varepsilon) \geq F(y_p + \varepsilon) - \delta \geq F(y_p) + 2\delta - \delta = p + \delta > p.
\]

(30)

From (29) and (30) we conclude that for all \( n \geq n_0 \) we have \( |y_n - y_p| \leq \delta \), and therefore, \( y_n \to y_p \) a.e. This concludes the proof.

\( \square \)

**Proof of Theorem 2** The continuity of \( G \) implies that \( F_{2a} \ast G \) and \( F_{3a} \ast G \) are both continuous cumulative distribution functions. Since weak convergence to a continuous limit distribution function implies uniform convergence (see, for example, Lemma 2.11 in Van der Vaart (2000)), Lemmas 2 and 3 imply that \( \hat{F}_{2a} \ast \hat{G} \) and \( \hat{F}_{3a} \ast \hat{G} \) converge uniformly to \( F_{2a} \ast G \) and \( F_{3a} \ast G \), respectively, a.s.

Combining these results with Theorem 1, we obtain (9). From Lemma 4, we conclude that \( T_p(\hat{F}_{\operatorname{RSEP}}) \) is well defined. Moreover, Lemma 4 and the uniform convergence proved below imply that \( T_p(\hat{F}_{\operatorname{RSEP}}) \) converges to \( T_p(F_0) \) a.s.

\( \square \)

**Proof of Theorem 3** We will show that A1–A3 are satisfied, with \( \hat{\pi}_n(X) = \expit(\hat{\gamma}'_nX) \), \( \pi_\infty(X) = \expit(\gamma'_\infty X) \), \( \hat{\beta}_n(X) = \beta'_nX \) and \( g_\infty(X) = \beta'_\inftyX \). To prove A1, note that

\[
|\hat{\pi}_n(X) - \pi_\infty(X)| = |\pi(X; \hat{\gamma}_n) - \pi(X; \gamma'_\infty)| = |\expit'(\hat{\gamma}'_nX)X'(\hat{\gamma}_n - \gamma'_\infty)|,
\]

where \( \hat{\gamma}_n \) is an intermediate point between \( \hat{\gamma}_n \) and \( \gamma'_\infty \). The convergence of \( \hat{\gamma}_n \) to \( \gamma'_\infty \) a.s. combined with the assumed compactness for the support of \( X \) implies the validity of A1.

A2 is satisfied since \( \expit(\gamma'_\infty X) \) is continuous and \( X \) has a compact support.

To prove the validity of A3, observe that \( |\hat{\beta}_n(X) - g_\infty(X)| = |(\hat{\beta}_n - \beta'_\infty)'X| \). The convergence of \( \hat{\beta}_n \) to \( \beta'_\infty \) a.s. guarantees that A3 is also satisfied.

Finally, note that if \( \mathbb{P}(A = 1 \mid X) = \expit(\gamma'_0X) \), then \( \gamma'_\infty = \gamma'_0 \), and so \( \pi_\infty(X) = \mathbb{P}(A = 1 \mid X) = \mathbb{P}(A = 1 \mid X) \). Also, if \( g(X) = \beta'_0X \), then \( \beta'_\infty = \beta'_0 \) implying that \( g_\infty(X) = g(X) \). We can now invoke Theorem 2 to conclude the proof of the theorem.

\( \square \)

**References**

Agostinelli C, Bianco AM, Boente G (2017) Robust estimation in single index models when the errors have a unimodal density with unknown nuisance parameter. arXiv preprint arXiv:1709.05422

Bianco AM, Spano PM (2019) Robust inference for nonlinear regression models. TEST 28(2):369–398

Bianco A, Boente G (2004) Robust estimators in semiparametric partly linear regression models. J Stat Plan Inference 122(1–2):229–252

Bianco A, Boente G, González-Manteiga W, Pérez-González A (2010) Estimation of the marginal location under a partially linear model with missing responses. Comput Stat Data Anal 54(2):546–564

Bianco AM, Boente G, González-Manteiga W, Pérez-González A (2011) Asymptotic behavior of robust estimators in partially linear models with missing responses: the effect of estimating the missing probability on the simplified marginal estimators. TEST 20(3):524–548

Bianco AM, Boente G, González-Manteiga W, Pérez-González A (2018) Plug-in marginal estimation under a general regression model with missing responses and covariates. TEST 28(1):1–41
Boente G, Fraiman R (1989) Robust nonparametric regression estimation. J Multivar Anal 29(2):180–198
Boente G, Rodriguez D (2008) Robust bandwidth selection in semiparametric partly linear regression models: Monte Carlo study and influential analysis. Comput Stat Data Anal 52(5):2808–2828
Boente G, Rodriguez D (2012) Robust estimates in generalized partially linear single-index models. Test 21(2):386–411
Cheng PE (1994) Nonparametric estimation of mean functionals with data missing at random. J Am Stat Assoc 89(425):81–87
Díaz I (2017) Efficient estimation of quantiles in missing data models. J Stat Plan Inference 141(2):711–724
Fasano MV, Maronna Ricardo RA, Sued M, Víctor J et al (2012) Continuity and differentiability of regression M functionals. Bernoulli 18(4):1284–1309
Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4):1161–1189
Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47(260):663–685
Kang JDY, Schafer JL (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci 22(4):523–539
Little RJA, Rubin DB (2014) Statistical analysis with missing data. Wiley, New York
Marazzi A, Yohai VJ (2004) Adaptively truncated maximum likelihood regression with asymmetric errors. J Stat Plan Inference 122(1–2):271–291
Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M (2018) Robust statistics: theory and methods (with R). Wiley, New York
Porter KE, Gruber S, Van Der Laan MJ, Sekhon JS (2011) The relative performance of targeted maximum likelihood estimators, Int J Biostat 7(1):1–34
Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89(427):846–866
Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90(429):106–121
Rotnitzky, A, Robins, J, Babino, L (2017) On the multiply robust estimation of the mean of the g-functional. arXiv preprint arXiv:1705.08582
Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592
Statti F, Sued M, Yohai VJ (2018) High breakdown point robust estimators with missing data. Commun Stat Theory Methods 47(21):5145–5162
Sued M, Yohai VJ (2013) Robust location estimation with missing data. Can J Stat 41(1):111–132
Van der Vaart AW (2000) Asymptotic statistics, vol 3. Cambridge University Press, Cambridge
Wang Q, Linton O, Härdle W (2004) Semiparametric regression analysis with missing response at random. J Am Stat Assoc 99(466):334–345
Yohai VJ et al (1987) High breakdown-point and high efficiency robust estimates for regression. Ann Stat 15(2):642–656
Zhang Z, Chen Z, Troendle JF, Zhang J (2012) Causal inference on quantiles with an obstetric application. Biometrics 68(3):697–706

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.