

# *Plug-in marginal estimation under a general regression model with missing responses and covariates*

**Ana M. Bianco, Graciela Boente,  
Wenceslao González-Manteiga & Ana  
Pérez-González**

## **TEST**

An Official Journal of the Spanish  
Society of Statistics and Operations  
Research

ISSN 1133-0686

Volume 28

Number 1

TEST (2019) 28:106-146

DOI 10.1007/s11749-018-0591-5



**Your article is protected by copyright and all rights are held exclusively by Sociedad de Estadística e Investigación Operativa. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at [link.springer.com](http://link.springer.com)".**

## Plug-in marginal estimation under a general regression model with missing responses and covariates

Ana M. Bianco<sup>1</sup> · Graciela Boente<sup>2</sup> ·  
Wenceslao González-Manteiga<sup>3</sup> ·  
Ana Pérez-González<sup>4</sup>

Received: 29 November 2017 / Accepted: 3 May 2018 / Published online: 5 June 2018  
© Sociedad de Estadística e Investigación Operativa 2018

**Abstract** In this paper, we consider a general regression model where missing data occur in the response and in the covariates. Our aim is to estimate the marginal distribution function and a marginal functional, such as the mean, the median or any  $\alpha$ -quantile of the response variable. A missing at random condition is assumed in order to prevent from bias in the estimation of the marginal measures under a non-ignorable missing mechanism. We give two different approaches for the estimation of the responses distribution function and of a given marginal functional, involving inverse probability weighting and the convolution of the distribution function of the

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11749-018-0591-5>) contains supplementary material, which is available to authorized users.

---

✉ Graciela Boente  
gboente@dm.uba.ar

Ana M. Bianco  
abianco@dm.uba.ar

Wenceslao González-Manteiga  
wenceslao.gonzalez@usc.es

Ana Pérez-González  
anapg@uvigo.es

- <sup>1</sup> Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET Ciudad Universitaria, Pabellón 2, 1428 Buenos Aires, Argentina
- <sup>2</sup> Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IMAS, CONICET Ciudad Universitaria, Pabellón 1, 1428 Buenos Aires, Argentina
- <sup>3</sup> Departamento de Estadística, Análisis Matemática e Optimización, Facultad de Matemáticas, Universidad de Santiago de Compostela, Campus Sur., 15706 Santiago de Compostela, Spain
- <sup>4</sup> Departamento de Estadística e Investigación Operativa, Universidad de Vigo, Campus Orense. Campus Universitario As Lagoas s/n, 32004 Ourense, Spain

observed residuals and that of the observed estimated regression function. Through a Monte Carlo study and two real data sets, we illustrate the behaviour of our proposals.

**Keywords** Fisher consistency · Kernel weights ·  $L$ -estimators · Marginal functionals · Missing at random · Semiparametric models

**Mathematics Subject Classification** 62F10 · 62G08

## 1 Introduction

Missing observations are a challenge that data scientists have to face very often in many statistical applications in different areas, such as biology, genetics, chemistry, social and environmental sciences, among others. A wide literature has grown in the last decades addressing the estimation and inference problems that arise when missing data occur, due to the possible bias that they may introduce in different regression models, both parametric and nonparametric ones.

In this paper, we focus on the estimation of the marginal distribution function when we deal with a regression model where the responses and some of the covariates are subject to missingness. It is clear that once the responses distribution is obtained, estimators of a marginal functional, such as the mean, the median or any  $\alpha$ -quantile, can be obtained. As mentioned in Díaz (2017), estimation of quantiles in missing data models is an appealing statistical problem with applications in many areas. For instance, in medicine quantiles are of interest either when studying the effect of a treatment or when establishing the standard growth parameters of a given population. Furthermore, quantiles are also used for diagnostic purposes, for instance, through boxplots which are a useful tool to describe a data set and to identify possible atypical data. Díaz (2017) considers the situation in which missing data arise only on the responses. However, the situation of missing data on both responses and covariates arises in many biological experiments where some explanatory variables can be controlled and others not. In epidemiology and biomedical sciences, for example, missing values at the covariate data often occur in studies performing multivariate survival analyses. When treated inappropriately, the presence of missing covariate values inhibits the formulation of a reliable model, introducing a potential bias that may mislead to wrong conclusions. Indeed, the naive practice of simply excluding any case with missing values, either in the responses or in the covariates, known as the *complete-case analysis*, may seriously bias the estimation.

Let us consider a random sample  $(y_i, \mathbf{x}_i^T)^T$ ,  $1 \leq i \leq n$ , where  $y_i \in \mathbb{R}$  are the responses and  $\mathbf{x}_i \in \mathbb{R}^d$  the covariates or design points, satisfying

$$y_i = \mu(\mathbf{x}_i) + \epsilon_i, \quad (1)$$

with the errors  $\epsilon_i$  i.i.d., independent of  $\mathbf{x}_i$  and such that  $\mathbb{E}(\epsilon_i) = 0$  and  $\text{VAR}(\epsilon_i) = \sigma_0^2$ . Note that  $\mu(\mathbf{x}_i)$  stands for a general regression function which may relate the responses, for instance, linearly or nonparametrically with  $\mathbf{x}_i$  or either through a semiparametric

model. We assume that missing values may occur in the responses and some fixed components of  $\mathbf{x}_i$ , while other components of the covariate vector are always observed.

As it is well known, most of the statistical methods in parametric, nonparametric and semiparametric regression models are designed for complete data sets, but these procedures may be not valid when missing observations are present. There are several approaches to tackle missing data (see Little and Rubin 2002). As Burton and Altman (2004) notice these strategies include: (a) the mere deletion approach of complete-case analysis, where only the complete cases for all collected variables are included in the analysis, (b) the available case analysis, where the cases with complete data for the variables in the fitted model are analysed using the largest possible data set, and (c) the variable omission approach, where the incomplete variables are omitted in the model. Other possible methods to handle missing values involve imputation techniques. We refer to Chen et al. (2008) for an in-depth analysis of the effect of missing data on the bias and asymptotic efficiency of different estimators, including strategies (a)–(c), under a non-ignorable missing scheme. For instance, the complete data analysis may lead to biased estimators if the data are not missing completely at random. The usual missing at random (MAR) condition that is introduced to prevent from bias assumes that the missingness mechanism is related to the observed data, but not to the unobserved values. Furthermore, nowadays, when handling high-dimensional data it is very usual to deal with datasets where responses and covariates are observed only partially. The effect of missing values in this context can be emphasized due to the very nature of this kind of data, in terms of mass and sparsity.

The goal of this paper is to provide estimators of marginal quantities related to the distribution of  $y$ , say  $\theta$ , such as, the marginal mean or the marginal quantiles when there are missing values both in the responses and partially in the covariates  $\mathbf{x}_i$ . For that purpose, we introduce estimators for the marginal responses distribution using the predictive capability of the covariates at hand. The study of quantile estimators is motivated by the fact that quantiles may be used as an alternative to the mean when studying the causal effect of a treatment on an outcome whose distribution exhibits heavy tails or asymmetry. We introduce different methods for handling missing data under the missing model described in Chen et al. (2015). These procedures generalize several well-known estimators described in the literature for missing responses only. As mentioned above, to reduce bias and obtain more efficient estimators some strategies, as imputation or inverse probability weighting (IPW), have been developed in order to include the information of partially observed data. In order to estimate the responses distribution, our first method is based on the nonparametric IPW approach introduced in Horvitz and Thompson (1952), where observations are weighted according to the inverse of the estimated probability of dropouts. Our second method extends the approach given in Müller (2009) and Sued and Yohai (2013). For this purpose, we first estimate the regression function  $\mu$  given in (1) by a stepwise procedure based on a propensity weighting approach to define estimators of the parametric and nonparametric components in the partial linear model. With these estimators we estimate the errors distribution and the distribution of  $\mu(\mathbf{x}_1)$ . Finally, by considering their convolution, we obtain an estimator of the responses distribution function,  $F_y$ , that allows to estimate any marginal quantity  $T(F_y)$ , given through a functional  $T$ . In this sense, even when we are not interested in estimating the conditional quantiles, the available covariate

information is used to correctly identify the unconditional quantiles under the missing at random assumption. It is worth noticing that most of the literature on missing data focuses on estimating the mean or a linear functional of the responses. One exception is the paper by Zhang et al. (2012) that deals with the estimation of any quantile associated with the distribution of counterfactual variables. In this paper we follow a different point of view which takes into account that the study of an estimate of the responses distribution provides a more global picture of the effect of missing data. More precisely, the novelty of this paper is that marginal quantities estimators are obtained, through a plug-in approach, from an accurate consistent estimator of the responses distribution function when missing values arise on the responses and some of the covariates.

The paper is organized as follows. Section 2 presents some marginal measures of interest as functionals of the marginal distribution function. The estimators when missing data occur in the responses and some of the covariates are introduced in Sect. 3. Section 3.3 studies some particular cases of regression models. A numerical study is carried out in Sect. 4 to examine the small sample properties of the proposed procedures, under the partial linear model. Section 5 considers two real data examples, while some final comments are given in Sect. 6. All proofs are relegated to “Appendix”.

## 2 Marginal estimation

Let us denote by  $\theta = T(F_y)$  any marginal functional, where we will use indistinctly the notation  $T(F_y)$  or  $T(Q_y)$  with the related probability measure  $Q_y$ , i.e.,  $F_y(s) = Q_y((-\infty, s])$ . Some examples of usual interest are the marginal mean of  $y_1$ , the marginal quantile functional, in particular, any  $\alpha$ -quantile for fixed  $\alpha$ , the  $\alpha$ -trimmed mean or more generally, a given  $L$ -functional.

Recall that an  $L$ -functional is defined as

$$T(F_y) = T_M(F_y) = \int F_y^{-1}(s) \, dM(s),$$

where  $M$  is a signed measure on  $(0, 1)$  and  $F_y^{-1}(s) = \inf\{x : F_y(x) \geq s\}$ , for  $0 < s < 1$  is the quantile function. When  $M$  has a density  $m$ , the functional  $T_M$  can be written as

$$T(F_y) = T_M(F_y) = \int s \, m(F_y(s)) \, dF_y(s).$$

Interesting usual cases arise for different choices of  $M$  leading to some location functionals. In particular, the mean corresponds to the situation  $m \equiv 1$ , that is,  $M(s) = s$  the Lebesgue measure. On the other hand, the  $\alpha$ -quantile, including the median, corresponds to the point mass at  $\alpha$  denoted  $\Delta_\alpha$ , while the  $\alpha$ -trimmed mean is related to a measure with density  $m(s) = \mathbb{I}_{[\alpha, 1-\alpha]}(s)$ . An example of a signed measure  $M$  is given by  $m(s) = \cos(2\pi s) [\cos(2\pi s) - 1]$ , which is negative if  $|s - 0.5| > 0.25$ , that corresponds to the  $L$ -estimator with maximum efficiency when  $F_y$  is the Cauchy distribution.

Note that the mean, the median and the  $\alpha$ -quantiles also correspond to the minimizer of  $\mathbb{E}\rho((y_1 - a))$ , for some suitable  $\rho$ -function. For the  $\alpha$ -quantile, we have that  $\rho(u) = \rho_\alpha(u) = \alpha \max(u, 0) + (1 - \alpha) \max(-u, 0)$ , while  $\rho(u) = u^2$  for the mean. The resulting minimizer is also the solution of  $\lambda(a) = \mathbb{E}\psi((y_1 - a)) = 0$  for a suitable  $\psi$ -function and will be denoted as  $T_\psi(F_Y)$ . For instance, for the mean  $\psi(u) = u$ , while for the median,  $\psi(u) = \psi_{0.5}(u) = \text{sg}(u) = I_{(0,\infty)}(u) - I_{(-\infty,0)}(u)$ .

Given a sample  $y_1, \dots, y_n$ , estimators of  $\theta = T(F_Y)$  can be computed plugging-in an estimator  $\widehat{F}_Y$  of the marginal distribution function  $F_Y$ , such as the empirical distribution, denoted  $\widehat{F}_{y,n}$ , that is  $\widehat{\theta} = T(\widehat{F}_{y,n})$ . In particular, using the empirical distribution, the  $L$ -estimators of  $\int s m(F_Y(s)) dF_Y(s)$  can be written as  $\sum_{i=1}^n a_i y_{(i)}$ , where  $y_{(1)} \leq \dots \leq y_{(n)}$  is the ordered sample and  $a_i = \int_{(i-1)/n}^{i/n} m(s) ds$ . For the  $\alpha$ -sample quantile,  $y_{(\lfloor n\alpha \rfloor)}$ , with  $\lfloor u \rfloor$  the integer part of  $u$ , Bahadur (1966) provides a representation of  $y_{(\lfloor n\alpha \rfloor)}$  that is useful to obtain joint asymptotic normality of the sample quantiles. A smooth alternative to the conventional sample quantile function as a nonparametric estimator of a population quantile function is proposed in Yang (1985). The proposed estimator is essentially a kernel type estimator and corresponds to the choice  $m(s) = m_h(s) = K_h(s - \alpha)$ , where  $K_h(u) = K(u/h)/h$  with  $K$  a density function symmetric about zero and  $h = h_n \rightarrow 0$ . This estimator has the same asymptotic distribution as the conventional sample quantile. Note that for the smoothed  $L$ -estimators, the measure  $M$  is not fixed, but varies with the sample size. As mentioned in Yang (1985), an alternative smooth family of estimators for the quantiles can be defined as  $\widehat{F}_Y^{-1}(s)$  with

$$\begin{aligned} \widehat{F}_Y(s) &= \int_{-\infty}^s \frac{1}{n} \sum_{i=1}^n K_h(y_i - t) dt = \frac{1}{n} \sum_{i=1}^n \widetilde{K} \left( \frac{s - y_i}{h} \right) \\ &= \int \widetilde{K} \left( \frac{s - u}{h} \right) d\widehat{F}_{y,n}(u), \end{aligned} \tag{2}$$

where  $\widetilde{K}(u) = \int_{-\infty}^u K(t) dt$ . In this case, the measure  $M = \Delta_\alpha$  as above. This smoothed version of the empirical distribution function was studied in Fernholz (1993).

Let  $M = M^+ - M^-$  with  $M^+$  and  $M^-$  positive measures. If the supports of  $M^+$  and  $M^-$  are proper subsets of  $(0, 1)$ , that is, contained in a compact set  $[\eta, 1 - \eta]$ ,  $0 < \eta < 1$ , then the target  $L$ -functional  $T_M$  is weakly continuous at  $F_Y$  provided that  $M$  does not put any point mass on a discontinuity point of  $F_Y^{-1}$  (see Theorem 3.1 in Huber and Ronchetti (2009)). In particular, if  $M$  has a density  $m$  and a compact support  $[\eta, 1 - \eta] \subset (0, 1)$ ,  $T_M$  is weakly continuous, so that  $\widehat{\theta} = T_M(\widehat{F}_Y)$  will be consistent for any weakly consistent sequence of estimators  $\widehat{F}_Y$ . Analogously, a similar result is obtained if  $F_Y$  has a density and  $M$  has support strictly included in  $(0, 1)$ .

These results suggest that the estimation of a marginal quantile or any marginal continuous functional can be accomplished by plugging-in a prior estimator of the marginal distribution of the responses and this is the essence of our approach. However, when missing data arise either only in the responses or jointly in the responses and some covariates, the estimators described above cannot be computed in practice since the empirical distribution function or its smooth version given in (2) cannot be computed.

Section 3 describes some alternatives to solve this problem. It is worth noticing that even when we mainly focus on marginal location or on quantile parameters, scale or dispersion measures could also be considered, among others.

### 3 Marginal estimators under missing covariates and responses

This section focuses on estimation of the responses distribution function,  $F_y$ , and of a marginal parameter  $\theta = T(F_y)$  under the regression model (1), when both the responses and some of the covariates  $\mathbf{x}_i$  have missing observations. As mentioned above, our interest may be, for instance, the marginal location of  $y_1$  or the marginal  $\alpha$ -quantile, so among the missing variables we will always include the responses. When missing data arise only on the responses, a common practice in regression is to impute the incomplete observations and then proceed to carry out the estimation of the mean of the response variable with the complete sample. The methods considered include linear regression (Yates 1933), kernel smoothing (Cheng 1994; Cheng and Chu 1996) nearest neighbour imputation (Chen and Shao 2000), semiparametric estimation (Wang et al. 2004), nonparametric multiple imputation (Aerts et al. 2002), empirical likelihood over the imputed values (Wang and Rao 2002), among others. Wang et al. (2004) consider inference on the mean of  $y$  under regression imputation of missing responses based on a partly linear model regression model. Under a quantile regression model, Chen et al. (2015) consider a missing at random model in which the responses and/or some covariates are jointly missing, whereas previous proposals can handle only one of them but not both. On the other hand, for general regression models, Chen et al. (2008) provide a depth theoretical investigation for inference with missing responses and covariates. In particular, a careful analysis of the bias is given for the case in which only the complete cases are used to estimate the regression parameter. As is well known when only regression covariates are missing, the complete-case (CC) analysis is direct, but has two potential drawbacks: the loss of efficiency due to discarding data and the bias introduced in the parameter estimators. However, as mentioned in Little (1992), CC analysis gives valid inferences if the missingness depends only on  $\mathbf{x}_i$  and not on the responses. For regression models, Chen et al. (2008) study several missing models in which the missing variable has two components  $\delta_{i,x}$  and  $\delta_{i,y}$  indicating respectively, if the covariates or the responses are not available. However, in this paper we will consider marginal estimators only under the missingness model defined in Chen et al. (2015).

More precisely, throughout this paper, we deal with an incomplete data set  $(y_i, \mathbf{x}_i^T, \delta_i)^T$ ,  $1 \leq i \leq n$ , where  $(y_i, \mathbf{x}_i^T) = (\mathbf{z}_i^{(m)T}, \mathbf{z}_i^T)$  with  $\mathbf{z}_i$  the  $k$ -dimensional vector containing all the variables observed,  $0 < k \leq d$ , and  $\mathbf{z}_i^{(m)}$  the vector containing the variables subject to missingness. The binary variable  $\delta_i = 1$  if all the values in  $\mathbf{z}_i^{(m)}$  are observed and 0 otherwise, and the conditional expectation of  $\delta_i$  is modelled as

$$\mathbb{P}(\delta_i = 1 | (y_i, \mathbf{x}_i)) = \mathbb{P}(\delta_i = 1 | \mathbf{z}_i) = p(\mathbf{z}_i). \tag{3}$$

This missing model corresponds to that introduced in Chen et al. (2015). However, in this paper, instead of a quantile regression model that concerns the conditional distribu-



tion, we assume the regression model (1) and we aim to estimate a marginal parameter. For that reason, the natural situation of interest is when  $y_i$  is one of the components of  $\mathbf{z}_i^{(m)}$ , hence the probability of missingness  $p(\mathbf{z}_i)$  given in (3) does not depend on  $y_i$ . Model (3) includes the framework in which only responses are missing taking  $\delta_i = 1$  when  $y_i$  is observed and  $\mathbf{z}_i^{(m)} = y_i$ , since in this case, all the covariates  $\mathbf{x}_i$  are available. Thus, as done in Chen et al. (2015), using (3) we include in a unified approach the situations in which missing data occur only among the responses and those in which missing values are missing are in the responses and also in a subset (but not all) the covariates.

As mentioned above, we focus on the estimation of the responses distribution function that will allow to provide estimators of any marginal parameter  $\theta = T(F_y)$  as described in Sect. 2. Two families of estimators for the marginal probability  $Q_y$  can be considered: the weighted simplified which corrects the bias caused in the estimation by the missing mechanism using an estimator of the missingness probability  $\widehat{p}(\mathbf{z})$  and a convolution type estimator that uses the information given by the assumed regression model. The convolution-based estimators to be considered below generalize, to the setting in which missing covariates arise, the proposal given in Sued and Yohai (2013) for a fully parametric model with missing values only in the responses (see also Müller 2009).

### 3.1 The weighted simplified estimator

Using the complete sample and inverse probability weighting,  $Q_y$  can be estimated by

$$\widehat{Q}_{y,WS} = \left\{ \sum_{\ell=1}^n \frac{\delta_\ell}{\widehat{p}(\mathbf{z}_\ell)} \right\}^{-1} \sum_{i=1}^n \frac{\delta_i}{\widehat{p}(\mathbf{z}_i)} \Delta_{y_i} = \sum_{i=1}^n \tau_i \Delta_{y_i}, \tag{4}$$

where  $\Delta_a$  is the point mass at point  $a$  and  $\widehat{p}$  is an estimator of the missing probability  $p$ .

Theorem 3.1 ensures consistency of any weighted simplified marginal estimator  $T(\widehat{Q}_{y,WS})$  based on a continuous functional  $T$  if the estimator of the missingness probability is uniformly strong consistent. Its proof is given in ‘‘Appendix’’.

**Theorem 3.1** *Let  $(y_i, \mathbf{x}_i^T, \delta_i)$ ,  $1 \leq i \leq n$  be i.i.d. random vectors over  $(\Omega, \mathcal{A}, \mathbb{P})$ ,  $\mathbf{x}_i \in \mathbb{R}^d$ , such that  $(y_i, \mathbf{x}_i^T) = (\mathbf{z}_i^{(m)T}, \mathbf{z}_i^T)$ ,  $\mathbf{z}_i \in \mathbb{R}^k$ ,  $0 < k \leq d$ , and (3) holds. Assume that*

- (i)  $\inf_{\mathbf{z} \in \mathcal{S}_z} p(\mathbf{z}) = i_p > 0$ , where  $\mathcal{S}_z$  is the support of the distribution of  $\mathbf{z}_1$ .
- (ii)  $\sup_{\mathbf{z} \in \mathcal{S}_z} |\widehat{p}(\mathbf{z}) - p(\mathbf{z})| \xrightarrow{a.s.} 0$ .

*Then,  $\Pi_K(\widehat{Q}_{y,WS}, Q_y) \xrightarrow{a.s.} 0$  and  $\Pi(\widehat{Q}_{y,WS}, Q_y) \xrightarrow{a.s.} 0$ , where  $\Pi_K$  and  $\Pi$  stand for the Kolmogorov and Prohorov distance, respectively.*

**Remark 3.1** It is worth noticing that assumptions (i) and (ii) in Theorem 3.1 are standard conditions when dealing with missing data. In particular, assumption (i) ensures

that locally at least some responses and covariates, i.e., some values of the vector  $\mathbf{z}_i^{(m)}$  are observed, which is a common assumption in the literature. In fact, it was required, among others, in Zhou et al. (2008), Chen and Keilegom (2013), Bravo (2015) and Chen et al. (2015). As noted in Liang et al. (2004), who deal with inference under a partial linear model with missing covariates, condition (ii) is easily verified for a properly parametrized model with compact support for  $\mathbf{z}$ , when the parameter estimators are consistent. On the other hand, for purely nonparametric models, multidimensional smoothing of  $\delta_i$  is needed when  $k > 1$ . One possibility is to consider a kernel method, that is,

$$\hat{p}(\mathbf{z}) = \left\{ \sum_{j=1}^n L \left( \frac{\mathbf{z}_j - \mathbf{z}}{b} \right) \right\}^{-1} \sum_{i=1}^n L \left( \frac{\mathbf{z}_i - \mathbf{z}}{b} \right) \delta_i,$$

with  $L : \mathbb{R}^k \rightarrow \mathbb{R}$  a kernel function and  $b = b_n$  the smoothing parameter. In this case, uniform convergence may be obtained if  $L(\mathbf{z}) = \ell(\|\mathbf{z}\|)$  with  $\ell$  a monotone decreasing function on  $[0, \infty)$ , the propensity  $p(\mathbf{z})$  is uniformly continuous and  $b \rightarrow 0$  and  $n b^k / \log(n) \rightarrow \infty$  which are the usual conditions for kernel estimators (see Example 38 and Exercises 26 and 28 in Pollard 1984). A different point of view was considered in Hirano et al. (2003) who considered a sieve approach using a series logit estimator and derived uniform consistency in their Lemma 1.

*Remark 3.2* One of the most frequently used graphical techniques for analysing a univariate data set is the boxplot, proposed by Tukey (1977). It provides a graphical tool for visualizing information regarding the location, spread, skewness as well as the tails for continuous unimodal data. However, when missing data arise in the responses, the boxplot cannot be constructed from the data at hand without taking into account the missingness effect. For that reason, we recommend to use the defined inverse probability weighting quantiles, i.e., those defined as  $\widehat{F}_{y, \text{WS}}^{-1}(\alpha)$  for  $\alpha = 0.25, 0.50$  and  $0.75$  to construct a boxplot adapted to the presence of missing data. Note that using the weighted simplified estimators avoids imposing any regression model where the estimation of its parameters that may be influenced by atypical responses. We illustrate the benefits of the adapted boxplots in the first dataset analysed in Sect. 5.

### 3.2 The convolution-based estimator

Another estimator of  $F_y$  may be obtained using that  $F_y$  is the convolution of the distributions of the errors and of the regression function. Let  $F_\epsilon$  and  $Q_\epsilon$  denote the distribution and the related probability measure of the errors  $\epsilon_i$  and let  $F_\mu$  and  $Q_\mu$  correspond to those of the true regression function  $\mu(\mathbf{x}_1)$ . Using consistent estimators  $\widehat{F}_\epsilon$  and  $\widehat{F}_\mu$  of  $F_\epsilon$  and  $F_\mu$ , respectively, a consistent estimator for  $F_y$  can directly be constructed. In fact, the *convolution-based* estimator introduced in Sued and Yohai (2013) can be adapted to the present setting as follows. Let  $\widehat{\mu}(\mathbf{x})$  be a consistent estimator of  $\mu(\mathbf{x})$ .

As in the situation in which only responses are missing, define

$$\widehat{Q}_\mu = \left\{ \sum_{\ell=1}^n \frac{\delta_\ell}{\widehat{p}(\mathbf{z}_\ell)} \right\}^{-1} \sum_{i=1}^n \frac{\delta_i}{\widehat{p}(\mathbf{z}_i)} \Delta_{\widehat{\mu}(\mathbf{x}_i)} = \sum_{i=1}^n \tau_i \Delta_{\widehat{\mu}(\mathbf{x}_i)},$$

where the weights  $\tau_i$  are normalized to guarantee that  $\widehat{Q}_\mu$  is a probability measure. As shown in Theorem 3.2 below, under mild conditions,  $\widehat{Q}_\mu$  is a consistent estimator of  $Q_\mu$ , since condition (3) holds. It is worth noticing that, when missing data arise only in the responses, one may take  $\tau_i = 1$ , for any  $1 \leq i \leq n$ , and in this case, we recover the proposal considered in Sued and Yohai (2013). However, to the best of our knowledge the situation in which both responses and some covariates are missing has not been considered and the inverse probability weighting allows to construct consistent estimators of  $Q_\mu$ .

When  $\delta_i = 1$ , both the response and covariates are observed for the  $i$ th individual, so the residuals can be effectively computed as  $\widehat{\epsilon}_i = y_i - \widehat{\mu}(\mathbf{x}_i)$ , so that an estimator of  $Q_\epsilon$  can be computed as

$$\widehat{Q}_\epsilon = \left\{ \sum_{\ell=1}^n \delta_\ell \right\}^{-1} \sum_{i=1}^n \delta_i \Delta_{\widehat{\epsilon}_i} = \sum_{i=1}^n \kappa_i \Delta_{\widehat{\epsilon}_i}.$$

It is worth noticing that if we consider an inverse probability weighting to estimate  $Q_\epsilon$ , that is, if  $\kappa_i = \tau_i$ , the estimators of  $Q_\epsilon$  will still be consistent.

Finally, we estimate  $Q_y$  by

$$\widehat{Q}_{y,\text{CONV}} = \widehat{Q}_\epsilon * \widehat{Q}_\mu.$$

Note that  $\widehat{Q}_{y,\text{CONV}}$  is a weighted empirical distribution

$$\widehat{Q}_{y,\text{CONV}} = \sum_{i=1}^n \sum_{j=1}^n \kappa_i \tau_j \Delta_{\widehat{y}_{ij}}, \tag{5}$$

with

$$\widehat{y}_{ij} = \widehat{\mu}(\mathbf{x}_j) + \widehat{\epsilon}_i \quad i, j \in \{\delta_\ell = 1\}.$$

As in Theorem 3.1, in Theorem 3.2 we derive, under mild assumptions, the convergence in Prohorov distance of the convolution-based response distribution function estimator to the marginal distribution function, which entails the strong consistency of  $T(\widehat{F}_{y,\text{CONV}})$  when the functional  $T$  is continuous. In particular,  $M$ -location marginal estimators with bounded score function and  $\alpha$ -marginal quantile estimators based on  $\widehat{F}_{y,\text{CONV}}$  turn out to be strongly consistent. The proof of Theorem 3.2 is given in ‘‘Appendix’’.

**Theorem 3.2** Let  $(y_i, \mathbf{x}_i^T, \delta_i)$ ,  $1 \leq i \leq n$  be i.i.d. random vectors over  $(\Omega, \mathcal{A}, \mathbb{P})$ ,  $\mathbf{x}_i \in \mathbb{R}^d$ , such that  $(y_i, \mathbf{x}_i^T) = (\mathbf{z}_i^{(m)T}, \mathbf{z}_i^T)$ ,  $\mathbf{z}_i \in \mathbb{R}^k$ ,  $0 < k \leq d$ , and (1) and (3) hold. Assume that

- (i)  $\inf_{\mathbf{z} \in \mathcal{S}_{\mathbf{z}}} p(\mathbf{z}) = i_p > 0$ , where  $\mathcal{S}_{\mathbf{z}}$  is the support of the distribution of  $\mathbf{z}_1$ .
- (ii)  $\sup_{\mathbf{z} \in \mathcal{S}_{\mathbf{z}}} |\widehat{p}(\mathbf{z}) - p(\mathbf{z})| \xrightarrow{a.s.} 0$ .
- (iii)  $\sup_{\mathbf{x} \in \mathcal{K}} |\widehat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| \xrightarrow{a.s.} 0$ , for any compact set  $\mathcal{K} \in \mathbb{R}^d$ .

Then, we have that

- (a)  $\Pi(\widehat{Q}_\mu, Q_\mu) \xrightarrow{a.s.} 0$
- (b)  $\Pi(\widehat{Q}_\epsilon, Q_\epsilon) \xrightarrow{a.s.} 0$

implying that  $\Pi(\widehat{Q}_{y, \text{CONV}}, Q_y) \xrightarrow{a.s.} 0$ .

*Remark 3.3* Assumptions (i) and (ii) of Theorem 3.2 are also required in Theorem 3.2; some comments on them are given in Remark 3.1. On the other hand, condition (iii) allows to replace the estimated regression function by the true one. It is worth noticing that the regression function  $\mu(\mathbf{x})$  in model (1) is a very general one. The only restriction given in Theorem 3.2 to ensure consistency of the marginal distribution estimators is the uniform strong convergence of the regression estimator stated in (iii). A related requirement was considered in Chen and Keilegom (2013) who deal with missing values only in the responses (see their assumption A3 that also requires order of convergence). Conditions ensuring the validity of (iii) are discussed below.

### 3.3 Some examples of regression models

As mentioned above, one key point to compute the convolution estimators is to provide consistent estimators of the regression function under a given regression model. The considered regression model may be parametric, nonparametric or semiparametric and in the parametric case, linear or nonlinear models may be adequate. In this section, we illustrate the estimation of the regression function in some particular examples. Other possible models not considered here are the linear quantile regression model studied in Chen et al. (2015), the generalized partial linear varying coefficient model considered in Bravo (2015) when some of the covariates are missing at random, but the responses are totally observed, and in Bravo and Jacho-Chávez (2016) when only responses are missing. However, the development, under a partial linear varying coefficient model, of a modified version of these estimators adapted to missing responses and covariates and the study of their uniform strong consistency is beyond the scope of the paper.

Throughout this section, we deal with an incomplete data set  $(y_i, \mathbf{x}_i^T, \delta_i)^T$ ,  $1 \leq i \leq n$ , where  $(y_i, \mathbf{x}_i^T) = (\mathbf{z}_i^{(m)T}, \mathbf{z}_i^T)$ , that is,  $y_i$  is one of the components of  $\mathbf{z}_i^{(m)}$ . Furthermore,  $\mathbf{z}_i \in \mathbb{R}^k$  contains the coordinates of  $\mathbf{x}_i$  that are always observed, so  $k > 0$ . On the other hand,  $\delta_i = 1$  if all the values in  $\mathbf{z}_i^{(m)}$  are observed and 0 otherwise and the MAR assumption (3) holds.

It is worth noticing that the estimation of the regression function when missing values arise has its own interest and was extensively studied in the literature, under different frameworks including situations where missing values arise only in the responses or when only some covariates are subject to drop-out. As for the conditional quantile regression model studied in Chen et al. (2015), an appealing feature of the proposals given below is their skill to save the existing gap to handle missing response and/or partially missing covariates, whereas most existing techniques can manage only one or the other, but not both. Thus, we take advantage of this section to discuss in the framework of regression all these missingness possibilities.

To give this global approach, we introduce some notation. Let  $\xi_i = \delta_i/q(\mathbf{z}_i)$  and  $\hat{\xi}_i = \delta_i/\hat{q}(\mathbf{z}_i)$  be random variables. To obtain simplified estimators, the practitioner takes  $q(\mathbf{z}) = \hat{q}(\mathbf{z}) = 1$  and takes  $q(\mathbf{z}) = p(\mathbf{z})$  and  $\hat{q}(\mathbf{z}) = \hat{p}(\mathbf{z})$  to consider inverse probability weighting procedures. From now on, when considering the inverse probability weighting, we will assume that  $\sup_{\mathbf{z} \in \mathcal{S}_z} |\hat{p}(\mathbf{z}) - p(\mathbf{z})| \xrightarrow{a.s.} 0$  and  $\inf_{\mathbf{z} \in \mathcal{S}_z} p(\mathbf{z}) > 0$  which corresponds to assumptions (i) and (ii) of Theorems 3.1 and 3.2 discussed in Remark 3.1.

### 3.3.1 The fully parametric regression model

Our first example deals with a fully parametric model and our goal is to show that estimators can be provided to ensure that (iii) holds. Let us consider the situation of the nonlinear regression model

$$y_i = \mu(\mathbf{x}_i) + \epsilon_i = G(\mathbf{x}_i, \boldsymbol{\beta}_0) + \epsilon_i \quad 1 \leq i \leq n, \tag{6}$$

where the function  $G$  is known and  $\boldsymbol{\beta}_0$  is the parameter to be estimated. It is worth noticing that this model includes the usual linear regression model as well as the nonlinear ones.

Note that after some algebra and taking conditional expectation we have that

$$\begin{aligned} \mathbb{E} \left\{ \xi_1 [y_1 - G(\mathbf{x}_1, \mathbf{b})]^2 \right\} &= \mathbb{E} \xi_1 \epsilon_1^2 + \mathbb{E} \left[ \xi_1 \Delta^2(\mathbf{x}_1, \mathbf{b}) \right] \\ &\quad + 2 \mathbb{E} \left\{ \Delta(\mathbf{x}_1, \mathbf{b}) \mathbb{E} \left[ \frac{\delta_1}{q(\mathbf{z}_1)} \epsilon_1 | \mathbf{x}_1 \right] \right\}, \end{aligned}$$

where  $\Delta(\mathbf{x}_1, \mathbf{b}) = G(\mathbf{x}_1, \boldsymbol{\beta}_0) - G(\mathbf{x}_1, \mathbf{b})$ . The MAR condition entails that

$$\mathbb{E} \left( \frac{\delta_1}{q(\mathbf{z}_1)} \epsilon_1 | \mathbf{x}_1 \right) = \mathbb{E} \left\{ \mathbb{E} \left[ \frac{\delta_1}{q(\mathbf{z}_1)} \epsilon_1 | (\epsilon_1, \mathbf{x}_1) \right] | \mathbf{x}_1 \right\} = \mathbb{E} \left[ \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} \epsilon_1 | \mathbf{x}_1 \right]. \tag{7}$$

When  $y_1$  is one of the components of  $\mathbf{z}_1^{(m)}$ , for any function  $q(\mathbf{z})$  we have that

$$\mathbb{E} \left[ \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} \epsilon_1 | \mathbf{x}_1 \right] = \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} \mathbb{E}(\epsilon_1 | \mathbf{x}_1) = \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} \mathbb{E}(\epsilon_1) = 0,$$

since the error  $\epsilon_1$  and the vector of covariates  $\mathbf{x}_1$  are independent. On the other hand, when the function  $q(\mathbf{z})$  equals the propensity  $p(\mathbf{z})$ , from (7), using again the independence between  $\epsilon_1$  and  $\mathbf{x}_1$ , we get that

$$\mathbb{E} \left( \frac{\delta_1}{q(\mathbf{z}_1)} \epsilon_1 | \mathbf{x}_1 \right) = \mathbb{E} (\epsilon_1 | \mathbf{x}_1) = \mathbb{E} (\epsilon_1) = 0.$$

Summarizing, when  $y_1$  is one of the components of  $\mathbf{z}_1^{(m)}$  or when  $y_1$  is fully observed and the function  $q(\mathbf{z})$  equals the propensity  $p(\mathbf{z})$

$$\mathbb{E} \left( \frac{\delta_1}{q(\mathbf{z}_1)} \epsilon_1 | \mathbf{x}_1 \right) = 0. \tag{8}$$

In other words, when missing values occur only among some of the covariates, the practitioner is forced to use inverse probability weighting, otherwise the resulting estimator would not be in general consistent.

Hence,  $\mathbb{E} \{ \xi_1 [y_1 - G(\mathbf{x}_1, \mathbf{b})]^2 \}$  is minimized at  $\beta_0$ . Furthermore, if  $\mathbb{P} (G(\mathbf{x}_1, \beta_0) - G(\mathbf{x}_1, \mathbf{b}) = 0) < 1$  for any  $\mathbf{b} \neq \beta_0$ , which is the usual identifiability condition in nonlinear models, the minimum is unique. Therefore, it is easy to see that the estimator defined as

$$\hat{\beta} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^n \hat{\xi}_i [y_i - G(\mathbf{x}_i, \mathbf{b})]^2$$

is strongly consistent to  $\beta_0$ . If the regression function  $G : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  is continuous, straightforward arguments allow to show that  $\sup_{\mathbf{x} \in \mathcal{K}} |\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| = \sup_{\mathbf{x} \in \mathcal{K}} |G(\mathbf{x}, \hat{\beta}) - G(\mathbf{x}, \beta_0)| \xrightarrow{a.s.} 0$ , for any compact set  $\mathcal{K} \in \mathbb{R}^d$ , showing that condition (iii) in Theorem 3.2 holds.

### 3.3.2 The fully nonparametric regression situation

Consider now the fully nonparametric regression model in which the regression function  $\mu(\mathbf{x})$  in (1) is a smooth, but otherwise unknown function. Note that the MAR assumption entails that

$$\mu(\mathbf{x}) = \frac{\mathbb{E} (\xi_1 y_1 | \mathbf{x}_1 = \mathbf{x})}{\mathbb{E} (\xi_1 | \mathbf{x}_1 = \mathbf{x})},$$

since (8) holds.

Let  $K$  a kernel function, i.e., a nonnegative integrable function on  $\mathbb{R}^d$  and  $h = h_n$  the bandwidth parameter. Based on the incomplete data set  $(y_i, \mathbf{x}_i^T, \delta_i)^T, 1 \leq i \leq n$  a consistent estimator of  $\mu(\mathbf{x})$  may be defined as

$$\widehat{\mu}(\mathbf{x}) = \left\{ \sum_{i=1}^n w_i(\mathbf{x}) \right\}^{-1} \sum_{i=1}^n w_i(\mathbf{x}) y_i,$$

where  $w_i(\mathbf{x}) = K((\mathbf{x}_i - \mathbf{x})/h) \widehat{\xi}_i$ , meaning that the estimator uses only the information at hand.

Let  $\mathcal{K} \subset \mathbb{R}^d$  be any compact set and  $\mathcal{K}_z$  its projection over  $\mathbb{R}^k$ . In this framework, the uniform convergence  $\sup_{\mathbf{x} \in \mathcal{K}} |\widehat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| \xrightarrow{a.s.} 0$  may be derived using analogous arguments to those considered in Proposition 2 of Collomb (1979) and in Proposition 3.2.1 of Boente et al. (2009). In fact, to derive condition (iii) in Theorem 3.2 we need that the covariates density  $f_{\mathbf{x}_1}$  and the propensity function  $p(\mathbf{z}_1)$  are continuous in a neighbourhood of  $\mathcal{K}$  and  $\mathcal{K}_z$ , respectively,  $\inf_{\mathbf{x} \in \mathcal{K}} f_{\mathbf{x}_1}(\mathbf{x}) > 0$ ,  $\inf_{z \in \mathcal{K}_z} p(z) > 0$ , the kernel  $K$  is a Lipschitz function such that  $\|\mathbf{u}\|^d K(\mathbf{u}) \rightarrow 0$  as  $\|\mathbf{u}\| \rightarrow \infty$  and the bandwidth  $h$  converges to 0 in such a way that  $nh^d/\log n \rightarrow \infty$ . This set of assumptions is standard when dealing with kernel estimators.

### 3.3.3 The partly linear regression model

In this section, we provide estimators for the regression function under a partly linear regression model that combines the parametric and nonparametric approaches given in the previous two subsections. Partial linear regression models provide a more flexible setting than the parametric ones by assuming that the regression function can be modelled linearly on some covariates, while it depends nonparametrically on some others. This model assume that

$$y_i = \mu(\mathbf{x}_i) + \epsilon_i = \mathbf{v}_i^T \boldsymbol{\beta}_0 + g_0(t_i) + \epsilon_i \quad 1 \leq i \leq n, \tag{9}$$

where  $\mathbf{x}_i = (\mathbf{v}_i^T, t_i)^T$  and  $\mathbf{v}_i \in \mathbb{R}^{d-1}$ . When missing values arise, the covariates  $t_i$  are usually assumed to be always observed, that is,  $t_i$  is one of the components of  $\mathbf{z}_i$ .

For fully observed data sets, two important viewpoints have been followed in the literature to supply estimators under the partly linear model (9), one is based on kernels and the other on  $B$ -splines, see for instance, Robinson (1988), Chen and Chen (1991), Härdle et al. (2000) and He et al. (2002).

*A kernel approach:* We first describe a three-step kernel-based procedure, where we combine a propensity estimator with a kernel smoothing.

**Step 1** Consider preliminary kernel estimators of the quantities

$$\eta(t) = \frac{\mathbb{E}(\xi_1 \mathbf{v}_1 | t_1 = t)}{\mathbb{E}(\xi_1 | t_1 = t)} \qquad \eta_0(t) = \frac{\mathbb{E}(\xi_1 y_1 | t_1 = t)}{\mathbb{E}(\xi_1 | t_1 = t)}$$

denoted  $\widehat{\boldsymbol{\eta}}(t) = (\widehat{\eta}_1(t), \dots, \widehat{\eta}_{d-1}(t))^T$  and  $\widehat{\eta}_0(t)$ , respectively. For instance, one may choose the kernel estimators

$$\widehat{\eta}_0(t) = \left\{ \sum_{i=1}^n w_i(t) \right\}^{-1} \sum_{i=1}^n w_i(t) y_i, \quad \widehat{\eta}_j(t) = \left\{ \sum_{i=1}^n w_i(t) \right\}^{-1} \sum_{i=1}^n w_i(t) v_{ij} .$$

where now  $w_i(t) = K((t_i - t)/h) \widehat{\xi}_i$ .

**Step 2** Note that using (3),  $\delta_1$  is conditionally independent of  $\mathbf{z}_1^{(m)}$  and  $\epsilon_1$ . As in (8), when  $p(\mathbf{z}) = q(\mathbf{z})$  or when  $y_1$  is one of the components of  $\mathbf{z}_1^{(m)}$ , this fact implies that  $\mathbb{E}(\xi_1 \epsilon_1 | t_1 = t) = 0$  since  $\mathbb{E}(\epsilon_1) = 0$ . Hence, taking conditional expectations in  $\xi_1 y_1 = \xi_1 \mathbf{v}_1^T \boldsymbol{\beta}_0 + \xi_1 g_0(t_1) + \xi_1 \epsilon_1$  we get that

$$\begin{aligned} \mathbb{E}(\xi_1 y_1 | t_1) &= \mathbb{E}(\xi_1 \mathbf{v}_1^T \boldsymbol{\beta}_0 | t_1) + \mathbb{E}[\xi_1 g_0(t_1) | t_1] + \mathbb{E}(\xi_1 \epsilon_1 | t_1) \\ &= \mathbb{E}(\xi_1 \mathbf{v}_1 | t_1)^T \boldsymbol{\beta}_0 + g_0(t_1) \mathbb{E}(\xi_1 | t_1), \end{aligned}$$

so  $\eta_0(t) = \boldsymbol{\eta}(t)^T \boldsymbol{\beta}_0 + g_0(t)$ . Replacing  $g_0(t)$  in the model we obtain that  $y_i - \eta_0(t_i) = [\mathbf{v}_i - \boldsymbol{\eta}(t_i)]^T \boldsymbol{\beta}_0 + \epsilon_i$ . Plugging-in the estimators  $\widehat{\eta}_0$  and  $\widehat{\boldsymbol{\eta}}$  of  $\eta_0$  and  $\boldsymbol{\eta}$  and using the approach given in Sect. 3.3.1 an estimator of  $\boldsymbol{\beta}_0$  may be defined as  $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} H_n(\boldsymbol{\beta})$  where

$$H_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \widehat{\xi}_i \left\{ [y_i - \widehat{\eta}_0(t_i)] - [\mathbf{v}_i - \widehat{\boldsymbol{\eta}}(t_i)]^T \boldsymbol{\beta} \right\}^2 = \frac{1}{n} \sum_{i=1}^n \widehat{\xi}_i \left( \widehat{r}_i - \widehat{\mathbf{r}}_{\mathbf{v},i}^T \boldsymbol{\beta} \right)^2 ,$$

with  $\widehat{r}_i = y_i - \widehat{\eta}_0(t_i)$  and  $\widehat{\mathbf{r}}_{\mathbf{v},i} = \mathbf{v}_i - \widehat{\boldsymbol{\eta}}(t_i)$ .

**Step 3** The nonparametric component may be estimated as  $\widehat{g}(t) = \widehat{\eta}_0(t) - \widehat{\boldsymbol{\eta}}(t)^T \widehat{\boldsymbol{\beta}}$  and the estimator of the regression function is then  $\widehat{\mu}(\mathbf{x}) = \mathbf{v}^T \widehat{\boldsymbol{\beta}} + \widehat{g}(t)$ .

It is worth noticing that, when missing observations arise only in the responses and  $q(\mathbf{z}) = \widehat{q}(\mathbf{z}) = 1$ , these estimators have been studied in Wang et al. (2004) who established their consistency. Our goal is to derive conditions under which assumption (iii) in Theorem 3.2 holds when missing responses and/or covariates arise according to (3). For that purpose, define  $H(\boldsymbol{\beta}) = \mathbb{E} \xi_1 \left( r_1 - \mathbf{r}_{\mathbf{v},1}^T \boldsymbol{\beta} \right)^2$  with  $r_1 = y_1 - \eta_0(t_1)$  and  $\mathbf{r}_{\mathbf{v},1} = \mathbf{v}_1 - \boldsymbol{\eta}(t_1)$ . Direct calculations entail that

$$H(\boldsymbol{\beta}) = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \mathbb{E} \left[ \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} \mathbf{r}_{\mathbf{v},1} \mathbf{r}_{\mathbf{v},1}^T \right] (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \mathbb{E} \left[ \epsilon_1^2 \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} \right] .$$

Hence,  $\boldsymbol{\beta}_0$  is the minimizer of  $H(\boldsymbol{\beta})$  that is unique when the matrix

$$\mathbf{A} = \mathbb{E} \left\{ \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} [\mathbf{v}_1 - \boldsymbol{\eta}(t_1)] [\mathbf{v}_1 - \boldsymbol{\eta}(t_1)]^T \right\}$$

is positive definite, which is a standard condition in partial linear models since it avoids any linear combination of  $\mathbf{v}_1$  from being perfectly predicted from  $t_i$ .

As when dealing with the fully nonparametric model, for any compact set  $\mathcal{T} \subset \mathbb{R}$  and any  $0 \leq j \leq d - 1$ , we have that  $\sup_{t \in \mathcal{T}} |\widehat{\eta}_j(t) - \eta_j(t)| \xrightarrow{a.s.} 0$  if



$h \rightarrow 0$  and  $nh/\log(n) \rightarrow \infty$ ,  $K$  is as in Sect. 3.3.2, the density of  $t_1$ ,  $f_{t_1}$ , and  $\mathbb{E}[p(\mathbf{z}_1)/q(\mathbf{z}_1)|t_1]$  are continuous in a neighbourhood of  $\mathcal{T}$ ,  $\inf_{t \in \mathcal{T}} f_{t_1}(t) > 0$  and  $\inf_{t \in \mathcal{T}} \mathbb{E}[p(\mathbf{z}_1)/q(\mathbf{z}_1)|t_1 = t] > 0$ . From this result, using standard arguments, it is easy to see that  $\widehat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}_0$ .

From the uniform strong consistency of  $\widehat{\eta}_j$  and the strong consistency of  $\widehat{\boldsymbol{\beta}}$ , we get that the estimator of the regression function  $\widehat{\mu}(\mathbf{x})$  will converge uniformly over compact sets to  $\mu(\mathbf{x}) = \mathbf{v}^T \boldsymbol{\beta}_0 + g_0(t)$ , showing that condition (iii) in Theorem 3.2 is fulfilled.

*A B-spline approach* We now focus on B-spline estimation of the function  $g_0$  that has been considered in the literature due to their good approximation properties. For that purpose, we assume that  $g_0$  has support on  $[0, 1]$  and that  $t_1$  has support on an interval  $\mathcal{J}$  within  $[0, 1]$ .

Let  $\ell \geq 1$  be the splines order and  $m_n$  the number of non-null knots. More precisely, denote as  $\mathcal{T}_n = \{\tau_i\}_{i=1}^{m_n+2\ell}$ , where  $0 = \tau_1 = \dots = \tau_\ell < \tau_{\ell+1} < \dots < \tau_{m_n+\ell+1} = \dots = \tau_{m_n+2\ell} = 1$  is a sequence of knots that partition the closed interval  $[0, 1]$  into  $m_n + 1$  subintervals  $\mathcal{I}_i = [\tau_{\ell+i}, \tau_{\ell+i+1})$ , for  $i = 0, \dots, m_n - 1$  and  $\mathcal{I}_{m_n} = [\tau_{m_n+\ell}, \tau_{m_n+\ell+1}]$ . A spline of order  $\ell \geq 1$  with knots  $\mathcal{T}_n$  is a polynomial of degree  $\ell - 1$  within the subintervals  $\mathcal{I}_i$ .

Denote as  $\mathcal{S}_n(\mathcal{T}_n, \ell)$  the class of splines of order  $\ell$  with knots  $\mathcal{T}_n$ . According to Corollary 4.10 of Schumaker (1981), for any  $g \in \mathcal{S}_n(\mathcal{T}_n, \ell)$ , there exists a class of B-spline basis functions  $\{B_j : 1 \leq j \leq k_n\}$ , with  $k_n = m_n + \ell$ , such that  $g = \sum_{j=1}^{k_n} \lambda_j B_j$ . This suggests that estimators of  $(\boldsymbol{\beta}_0, g_0)$  may be obtained as  $(\widehat{\boldsymbol{\beta}}, \widehat{g})$  where  $\widehat{g}(t) = \sum_{j=1}^{k_n} \widehat{\lambda}_j B_j(t)$  and

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{d-1}, \boldsymbol{\lambda} \in \mathbb{R}^{k_n}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \widehat{\xi}_i \left\{ y_i - \left[ \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^{k_n} \lambda_j B_j(t_i) \right] \right\}^2.$$

As above, assume that  $f_{t_1}$  is continuous in  $[0, 1]$ ,  $\inf_{t \in \mathcal{J}} f_{t_1}(t) > 0$ . Similar arguments to those considered in the fully observed data case would allow to show that condition (iii) in Theorem 3.2 holds, when the  $r$ -th derivative of the true function  $g_0$  is Lipschitz, the maximum spacing of the knots is of order  $O(n^{-\nu})$ , with  $1/(2r + 2) < \nu < 1/(2r)$  and the ratio between the maximum and the minimum spacings of knots is uniformly bounded. A complete study of B-spline estimators as well as the finite sample comparison with the kernel approach is beyond the scope of the paper.

### 3.3.4 The single-index regression model

One problem of the fully nonparametric model considered above is the well-known *curse of dimensionality*. Single-index models are a relevant topic in the broad class of semiparametric models. They reduce the dimensionality of the covariates through a suitable projection linked to the parametric component, while at the same time they capture a possible nonlinear relationship through an unknown smooth function. In this

section, we consider observations  $(y_i, \mathbf{x}_i^T, \delta_i)^T, 1 \leq i \leq n$ , such that

$$y_i = \mu(\mathbf{x}_i) + \epsilon_i = g_0(\boldsymbol{\beta}_0^T \mathbf{x}_i) + \epsilon_i, \tag{10}$$

where  $g_0$  and the single-index parameter  $\boldsymbol{\beta}_0$  are unknown. For identifiability purposes, it is assumed that  $\|\boldsymbol{\beta}_0\| = 1$  and the last component of  $\boldsymbol{\beta}_0$  is positive, while to obtain consistency results, it is usually assumed that the density of  $\boldsymbol{\beta}_0^T \mathbf{x}$  is uniformly bounded by below by a positive constant over its the compact support.

For complete data sets, several approaches have been considered (see Härdle et al. (2004). To face the problem of missing responses, Chen and Keilegom (2013) modify the semiparametric least squares approach. The semiparametric least squares approach can also be modified to deal with possible missing values also in some covariates by including the random variables  $\xi_i$ . More precisely, denote as

$$\widehat{g}_\beta(u) = \left\{ \sum_{i=1}^n w_{i,\beta}(u) \right\}^{-1} \sum_{i=1}^n w_{i,\beta}(u) y_i,$$

where  $w_{i,\beta}(u) = \widehat{\xi}_i K((\boldsymbol{\beta}^T \mathbf{x}_i - u)/h)$ . An estimator of  $\boldsymbol{\beta}_0$  adapted to the situation in which missing responses and/or some fixed covariates arise may be defined as

$$\widehat{\boldsymbol{\beta}} = \underset{\|\boldsymbol{\beta}\|=1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \widehat{\xi}_i [y_i - \widehat{g}_\beta(\boldsymbol{\beta}^T \mathbf{x}_i)]^2.$$

Hence, the final estimator of the regression function is obtained as  $\widehat{\mu}(\mathbf{x}) = \widehat{g}_{\widehat{\boldsymbol{\beta}}}(\widehat{\boldsymbol{\beta}}^T \mathbf{x})$ . As noted in Chen and Keilegom (2013), assumptions which guarantee that condition (iii) in Theorem 3.2 holds can be derived by adapting the arguments used in the fully nonparametric and the partly linear models to the context of the single-index one. We leave the complete study of the properties and small sample behaviour of these estimators for future research.

### 3.4 Some comments on the plug-in marginal estimators

The response probability measure estimators described above allow to estimate marginal quantities given through a functional of the marginal distribution. In particular, if the target is  $\theta = T(F_y)$ , the estimator equals  $\widehat{\theta} = T(\widehat{F}_y)$  where  $\widehat{F}_y$  is one of the estimators defined in Sects. 3.1 or 3.2.

For instance, when the goal is to estimate the marginal mean value of  $y_1$ , the weighted simplified and the convolution-based are given by  $\widehat{\theta}_{\text{WS}} = \sum_{i=1}^n \tau_i y_i$  and  $\widehat{\theta}_{\text{CONV}} = \sum_{i=1}^n \kappa_i \widehat{\epsilon}_i + \sum_{i=1}^n \tau_i \widehat{\mu}(\mathbf{x}_i) = \sum_{i=1}^n \kappa_i y_i + \sum_{i=1}^n (\tau_i - \kappa_i) \widehat{\mu}(\mathbf{x}_i)$ , with  $\kappa_i = \delta_i \left\{ \sum_{\ell=1}^n \delta_\ell \right\}^{-1}$  and  $\tau_i = \left\{ \sum_{\ell=1}^n \delta_\ell / \widehat{p}(\mathbf{z}_\ell) \right\}^{-1} \delta_i / \widehat{p}(\mathbf{z}_i)$ . It is worth noticing that if we consider an inverse probability weighting to estimate  $Q_\epsilon$ , that is, if  $\kappa_i = \tau_i$ , then

$\widehat{\theta}_{\text{CONV}} = \widehat{\theta}_{\text{WS}}$ . Consistency of the marginal mean estimators is obtained using similar arguments to those used in the proof of Theorems 3.1 and 3.2

Note that if the target is the median or any other quantile, a simple expression is not available and the related quantity may be estimated using the median or quantile of  $\widehat{F}_y$ . Furthermore, in these last situations the functional  $T$  is continuous, so that Theorems 3.1 and 3.2 entail that the marginal estimators are consistent.

#### 4 Monte Carlo study: the particular case of the partial linear model

A simulation study was carried out to show the performance of the marginal estimators proposed in Sect. 3, under the partial linear regression model (9) when the regression parameter  $\beta \in \mathbb{R}^D$  has dimension  $D = d - 1 = 1$  and 2. We consider the kernel smoothing approach described in Sect. 3.3.3. To choose the smoothing parameter used in Step 1, a cross-validation procedure based on response predictions was implemented.

Since our target is to estimate marginal quantities it is sensible to consider the case in which missing values occur also among the responses. For that reason, either the simplified approach or the inverse probability weighting method can be considered to provide consistent estimators for  $\beta_0$  and  $g_0$ . We only report here the results corresponding to the simplified approach, i.e.,  $\widehat{q}(\mathbf{z}) = q(\mathbf{z}) = 1$ . Similar results are obtained when  $\widehat{q}(\mathbf{z}) = \widehat{p}(\mathbf{z})$  and  $q(\mathbf{z}) = p(\mathbf{z})$ .

The target marginal parameters are the mean, median and the quantiles 10, 25, 75 and 90%. For each of them, we computed the estimators defined as  $\widehat{\theta} = T(\widehat{F}_y)$  where  $\widehat{F}_y$  is the distribution function related to the probability measures defined in (4) and (5). As an illustration, in dimension  $D = 1$ , we have also computed the responses distribution function estimators  $\widehat{F}_{y,\text{WS}}$  and  $\widehat{F}_{y,\text{CONV}}$  for one of the considered settings, in order to have a deepest insight of the estimators performance.

We have chosen two errors distributions corresponding to symmetric and asymmetric errors with null expectations,

- $\mathcal{E}_1: \epsilon_i \sim N(0, \sigma^2)$  with  $\sigma^2 = 0.25$ .
- $\mathcal{E}_2: \epsilon_i \sim 0.25(\chi_2^2 - 2)$ .

To describe the missing models to be considered, denote  $v_1(t) = t - 0.5$ ,  $v_2(t) = (t - 0.5)^2$ ,

$$H_{\alpha, v_j}(t) = \frac{1}{1 + \exp(\alpha v_j(t))}, H_1 = H_{-2, v_1} \text{ and } H_2 = H_{-3, v_2}.$$

Missing responses and covariates were generated, both when  $D = 1$  and 2, according to the following models

- $\mathcal{M}(H_1): \mathbf{z}_i^{(m)\text{T}} = (y_i, \mathbf{v}_i^{\text{T}})^{\text{T}}$  and  $\mathbb{P}(\delta_i = 1|t_i) = H_1(t_i)$
- $\mathcal{M}(H_2): \mathbf{z}_i^{(m)\text{T}} = (y_i, \mathbf{v}_i^{\text{T}})^{\text{T}}$  and  $\mathbb{P}(\delta_i = 1|t_i) = H_2(t_i)$ .

In all cases,  $\mathcal{M}(1)$  stands for the situation in which no missing data arise, that is,  $\mathbb{P}(\delta_i = 1|(y_i, \mathbf{x}_i)) = 1$  and is a benchmark that allows to study the loss of the different

procedures on the selected missing schemes. Finally,  $\mathcal{M}(H_1)$  and  $\mathcal{M}(H_2)$  correspond to the model described in Sect. 3 in which the responses and covariates corresponding to the linear component are missing and the procedures to be compared are those described therein, that is, those based on  $\widehat{Q}_{y,WS}$  and  $\widehat{Q}_{y,CONV}$ .

Three settings for the propensity are considered. In the first one, the propensity is assumed to be known so that  $\widehat{p} = p$ , in the second one it is estimated using the true logistic model which generates the missing observation. Finally, in the last setting, the missingness probability is estimated using a kernel estimator based on the covariates  $t_i$ . When computing the kernel estimators a cross-validation criterion was used.

In all figures, the blue-filled circle corresponds to the estimators based on  $Q_{y,WS}$ , while the green star to those based on the convolution-based probability measure  $\widehat{Q}_{y,CONV}$ . Besides, first tick corresponds to  $\mathcal{M}(1)$ , the next two ticks on the horizontal axis indicate that  $\widehat{p}(t) = p(t)$ , while the two subsequent ones correspond to  $\widehat{p} = \widehat{p}_{log}$  and the last two to the kernel fit  $\widehat{p} = \widehat{p}_K$ . Note that under  $\mathcal{M}(1)$  the missing probability is not estimated. The label indicating the fit for the propensity model is followed either by  $\mathcal{M}(H_2)$  or  $\mathcal{M}(H_1)$ , in this order.

To evaluate the performance of the estimators, we performed  $nr = 1000$  replications with samples of size  $n = 100$ .

### 4.1 Simulation study when $\beta \in \mathbb{R}$

When the regression parameter is one-dimensional,  $D = 1$ , we conduct a simulation study based on the following partly linear regression model

$$y_i = \beta v_i + 2 \sin(4\pi(t_i - 0.5)) + \epsilon_i \quad 1 \leq i \leq n, \tag{11}$$

where  $\beta = 2$ . The covariates  $(v_i, t_i)^T$  were generated under four different distributions

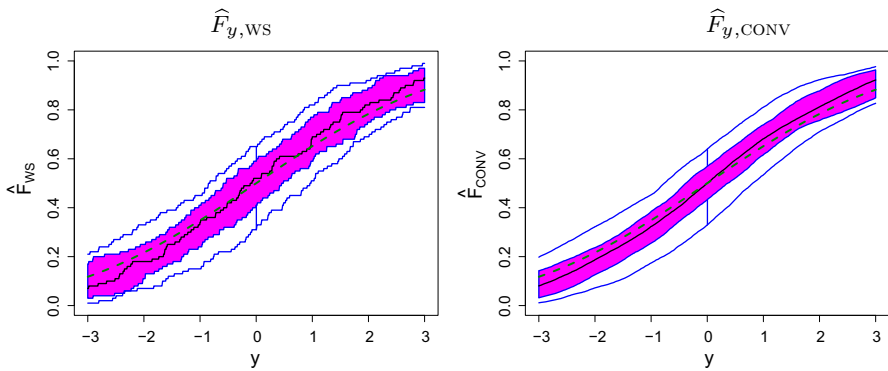
- $\mathcal{D}_1$ :  $(v_i, t_i)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\mu} = (0, \frac{1}{2})^T$  and  $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1/(6\sqrt{3}) \\ 1/(6\sqrt{3}) & 1/36 \end{pmatrix}$ .
- $\mathcal{D}_2$ :  $v_i \sim N(0, 1)$  and  $t_i \sim \mathcal{U}(0, 1)$ .
- $\mathcal{D}_3$ : The distributions of  $v_1$  and  $t_1$  were  $\mathcal{U}(-0.5, 0.5)$  and  $\mathcal{U}(0, 1)$ , respectively, but in this setting,  $\text{cor}(v_1, t_1) = 0.25$ .
- $\mathcal{D}_4$ : The distributions of  $v_1$  and  $t_1$  were  $\mathcal{U}(-0.5, 0.5)$  and  $\mathcal{U}(0, 1)$ , respectively, but in this setting, unlike  $\mathcal{D}_3$ ,  $\text{cor}(v_1, t_1) = 0.10$ .

Note that under  $\mathcal{D}_3$  the correlation between the covariates is stronger than under  $\mathcal{D}_2$  and  $\mathcal{D}_4$ . To generate the correlated uniform distributions given in  $\mathcal{D}_3$  and  $\mathcal{D}_4$ , we use copulas as follows. Let  $\mathbf{Z} \sim N_2(0, \boldsymbol{\Sigma})$  be a bivariate Gaussian random vector where  $\boldsymbol{\Sigma}_{1,1} = \boldsymbol{\Sigma}_{2,2} = 1$  and  $\boldsymbol{\Sigma}_{1,2} = \boldsymbol{\Sigma}_{2,1} = 2 \sin(\rho \pi/6)$ , with  $\rho \in (-1, 1)$ . Let  $v_1 = v_{1,0} - 0.5, v_{1,0} = \Phi(Z_1)$  and  $t_1 = \Phi(Z_2)$ , where  $\Phi$  is the cumulative distribution function of a standard normal distribution. It follows that the marginal distribution of both  $v_{1,0}$  and  $t_1$  is  $U(0, 1)$  with correlation  $\rho$ .

As mentioned above, besides the marginal quantile estimators, to illustrate we have also computed the responses distribution function estimators  $\widehat{F}_{y,WS}$  and  $\widehat{F}_{y,CONV}$  for

**Table 1** Target marginal values: mean values over replications of the estimated parameters, when  $D = 1$

	Mean	Median	$F^{-1}(0.10)$	$F^{-1}(0.25)$	$F^{-1}(0.75)$	$F^{-1}(0.90)$
$\mathcal{E}_1$						
$\mathcal{D}_1$	0.0000	-0.0003	-3.2104	-1.7328	1.7325	3.2105
$\mathcal{D}_2$	0.0001	-0.0002	-2.7952	-1.4725	1.4726	2.7954
$\mathcal{D}_3$	0.0000	-0.0001	-1.4668	-0.8279	0.8279	1.4669
$\mathcal{D}_4$	0.0000	0.0001	-1.4137	-0.7746	0.7747	1.4138
$\mathcal{E}_2$						
$\mathcal{D}_1$	-0.0000	-0.0053	-3.2069	-1.7356	1.7276	3.2103
$\mathcal{D}_2$	0.0001	-0.0082	-2.7885	-1.4750	1.4660	2.7978
$\mathcal{D}_3$	-0.0000	-0.0182	-1.4682	-0.8428	0.8112	1.4279
$\mathcal{D}_4$	0.0000	-0.0274	-1.4111	-0.7793	0.7567	1.3865

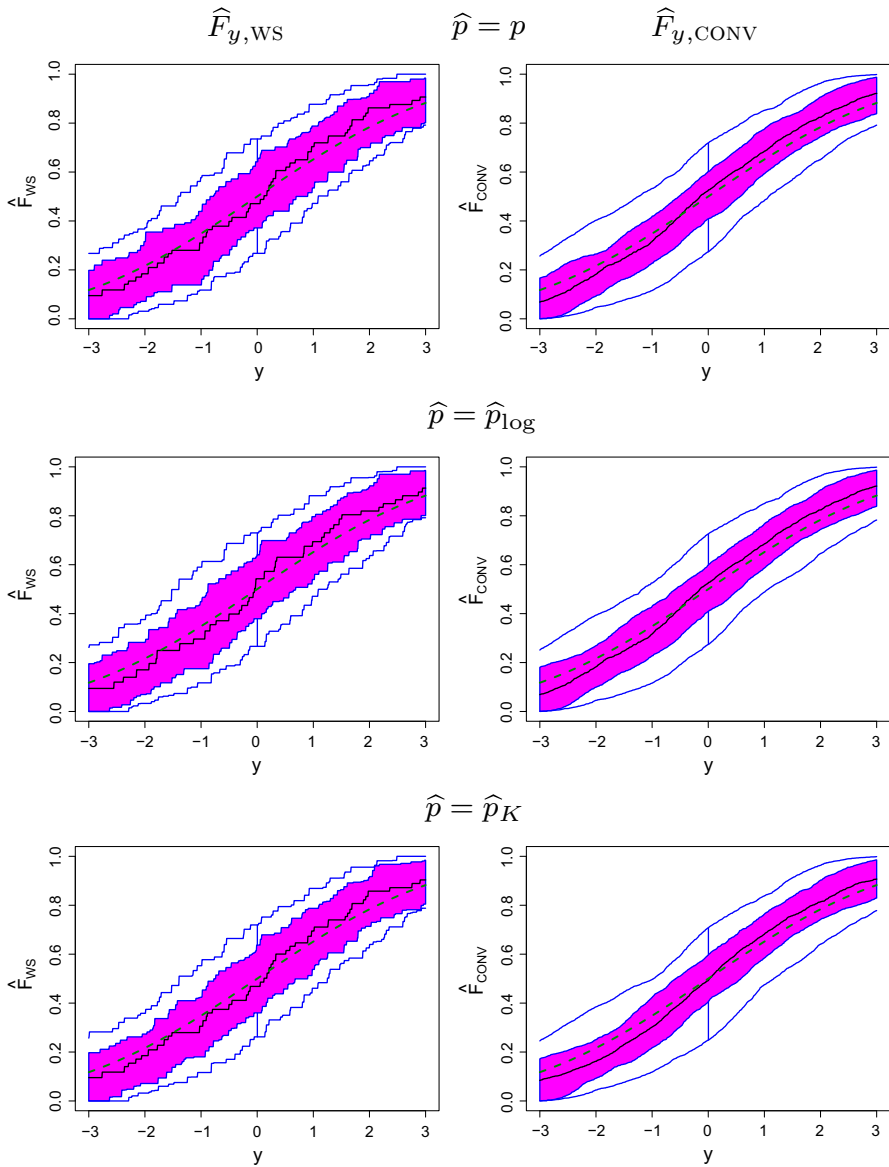


**Fig. 1** Functional boxplots of the responses distribution function estimators under  $\mathcal{E}_1, \mathcal{D}_2$  when  $D = 1$  and no missing data arise. The dashed line corresponds to  $F_y$

the special case of symmetric errors, that is, under  $\mathcal{E}_1$  and when the covariates have a distribution given by  $\mathcal{D}_2$ . In this case the true distribution function  $F_y$  is given by

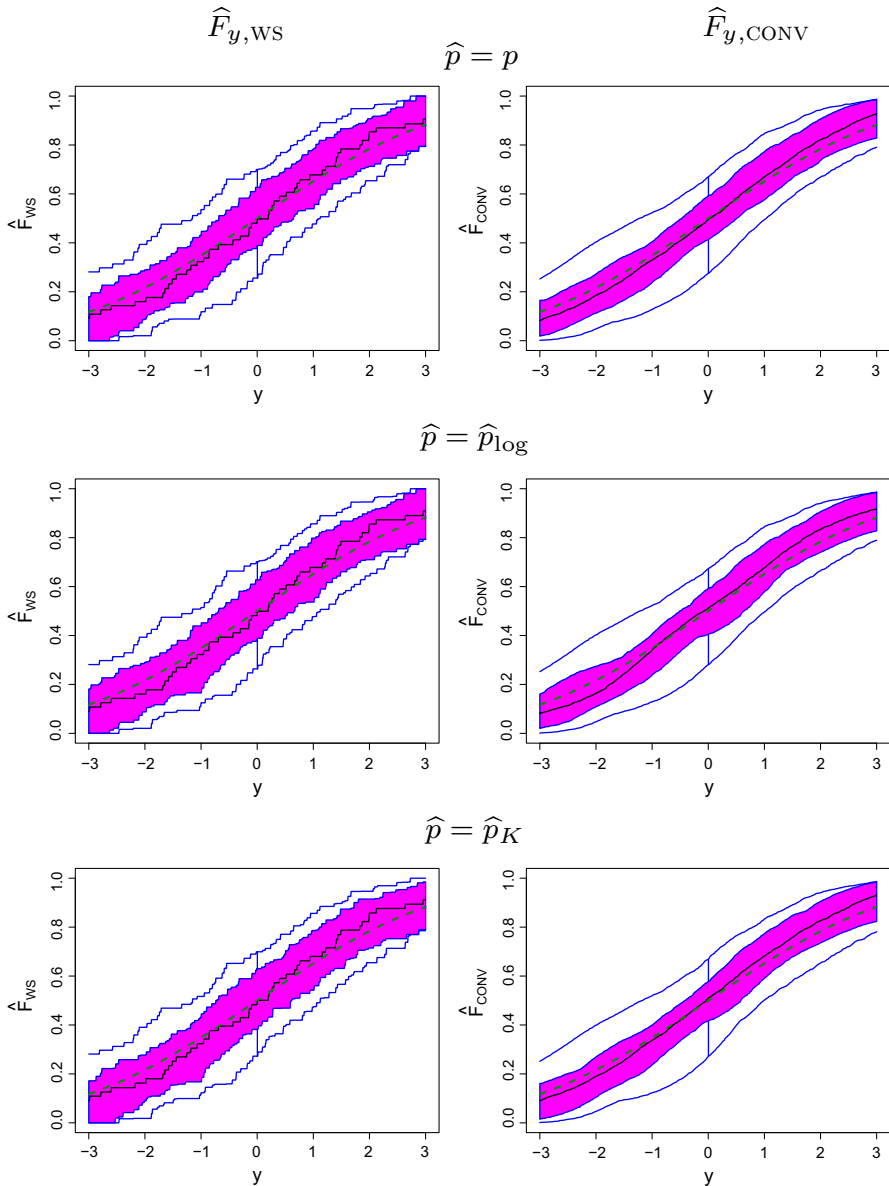
$$F_y(y) = 0.5 + \frac{1}{4\pi} \int_{-2\pi}^{2\pi} \left\{ \Phi \left[ \frac{y - 2 \sin(x)}{\sqrt{(\beta^2 + \sigma^2)}} \right] - \Phi \left[ \frac{-2 \sin(x)}{\sqrt{(\beta^2 + \sigma^2)}} \right] \right\} dx \quad (12)$$

for  $y > 0$  and  $F_Y(y) = 1 - F_Y(-y)$  for  $y < 0$  which can be computed numerically. We evaluated  $F_y$  and its estimates over an equally spaced grid of 1000 points between  $-3$  and  $3$ , that correspond approximately to the 10 and 90% of  $F_y$  (see Table 1). Figures 1, 2 and 3 give the functional boxplots of the obtained estimates under  $\mathcal{M}(1)$ ,  $\mathcal{M}(H_1)$  and  $\mathcal{M}(H_2)$ , respectively. The deepest curve is plotted in black, while the true distribution function  $F_y$  is given in dashed lines. As seen in these plots, an advantage of the convolution-based responses distribution estimators is that they are smoothers than those obtained with the weighted simplified procedure. Besides, the band defined by the 50% deepest curves is narrower for  $\hat{F}_{y,conv}$  than when using  $\hat{F}_{y,ws}$ , showing



**Fig. 2** Functional boxplots of the responses distribution function estimators under  $\mathcal{E}_1, D_2$  when  $D = 1$  and missing data arise according to  $\mathcal{M}(H_1)$ . The dashed line corresponds to  $F_y$

that the convolution method provides more accurate estimates, a result that was already known for the marginal mean estimators when only responses are missing. Finally, for this sample size and the considered missing schemes, these plots suggest that  $\hat{F}_{y,WS}$  provides a better fit for large absolute values of  $y$ , while  $\hat{F}_{y,CONV}$  has a better



**Fig. 3** Functional boxplots of the responses distribution function estimators under  $\mathcal{E}_1, \mathcal{D}_2$  when  $D = 1$  and missing data arise according to  $\mathcal{M}(H_2)$ . The dashed line corresponds to  $F_y$

performance in a neighbourhood of 0. Hence, one may guess that the extreme quantiles based on  $\hat{F}_{y,WS}$  will be less biased than those based on  $\hat{F}_{y,CONV}$ .

To summarize the simulation results, the true marginal values are needed. Except for the marginal mean which is equal to 0 for all the considered cases, the quantiles should be evaluated numerically. In order to obtain an accurate value, we generate 100

**Table 2** Bias for the marginal estimates computed with the observations at hand, i.e., obtained from  $Q_n = \{\sum_{i=1}^n \delta_i\}^{-1} \sum_{i=1}^n \delta_i \Delta_{y_i}$ , when  $D = 1$

	$\mathcal{E}_1$		$\mathcal{E}_2$		$\mathcal{E}_1$		$\mathcal{E}_2$	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
	$\mathcal{D}_1$				$\mathcal{D}_2$			
$\mathcal{M}(1)$	-0.010	-0.012	-0.013	-0.013	-0.006	-0.008	-0.007	-0.005
$\mathcal{M}(H_1)$	0.278	0.304	0.276	0.302	0.145	0.151	0.145	0.147
	$\mathcal{D}_3$				$\mathcal{D}_4$			
$\mathcal{M}(1)$	-0.001	-0.003	-0.004	-0.004	-0.002	-0.004	-0.004	-0.015
$\mathcal{M}(H_1)$	0.195	0.250	0.194	0.257	0.173	0.205	0.171	0.200

samples of size  $10^6$  and for each sample we compute the considered marginal quantiles. With this choice of the sample size and the number of replications, the standard errors of the computed marginal measures reported in Table 1 do not exceed  $6 \times 10^{-4}$ . The average values over the 100 samples are given in Table 1 and are considered as the target quantities when computing the bias and the mean square error of our estimators. It is worth noting that under  $\mathcal{E}_1$  both the median and the mean of the responses are 0; however, the values obtained for the median present some disagreement. In this case, in the comparisons the value 0 was taken as the true one, the same decision was made when comparing the estimators of the mean under  $\mathcal{E}_2$ .

For the missing model  $\mathcal{M}(H_1)$ , the proportion of observed data is around a 50%. More precisely, to estimate the proportion of observed data we compute the mean over 100 replications of the average number of observed data  $\sum_{i=1}^n \delta_i/n$  with samples of size 100,000. Under  $\mathcal{M}(H_1)$  this mean value was 0.5 for all the distributions.

Once the target values were computed, as mentioned above, in our simulations we performed  $nr = 1000$  replications with samples of size  $n = 100$  to evaluate the performance of the estimators. It is well known that computing the marginal estimators with the observations at hand may produce biased estimators. Table 2 illustrates this fact reporting the biases of the marginal estimators  $\tilde{\theta}$  based on the naive estimator  $\tilde{Q}_n = \{\sum_{i=1}^n \delta_i\}^{-1} \sum_{i=1}^n \delta_i \Delta_{y_i}$ . The obtained values show the high bias of both the mean and the median when missing observations are generated according to  $\mathcal{M}(H_1)$ . The aim is to show that our proposals improve the bias behaviour of the naive approach.

Tables S.1 to S.6 in the supplementary file report the bias (BIAS), standard deviation (SD) and mean square errors (MSE) of the estimators mentioned above, that is, for mentioned above for the mean, the median and the 10, 25, 75 and 90% quantiles. On the other hand, Figures S.1 to S.6 in the supplementary file plot the absolute value of the bias and the mean square errors under  $\mathcal{E}_1$ , where also similar plots under  $\mathcal{E}_2$  are given.

It is clear from all the tables and figures that the convolution estimator based on  $\hat{Q}_{y,CONV}$  is the one with the best behaviour in MSE. This behaviour is coherent with the described performance of the functional boxplots of the responses distribution function estimators, i.e., the band of the 50% deepest functions are less spread when using  $\hat{F}_{y,CONV}$ , under  $\mathcal{D}_2$ . In all cases, the MSE of all estimators are larger when



the covariate  $x$  is normal than when it is uniform. This enlargement is in most cases due to an increase of the standard deviation under  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . Neither the different correlations considered nor the asymmetry of the errors affect the performance of the estimators. Regarding the bias, the weighted simplified quantile estimators have smaller biases at a cost of a larger standard deviation. However, as shown in Figure S.1,  $\widehat{Q}_{y, \text{CONV}}$  has similar biases and MSE for the mean, while for the median, the biases are slightly increased when using the convolution estimator, but the MSE are always smaller (see Figure S.1) showing the advantage of this proposal.

### 4.2 Simulation study when $\beta \in \mathbb{R}^2$

The model used for bivariate covariates ( $D = 2$ ) in the linear regression component is

$$y_i = \beta^T \mathbf{v}_i + 2 \sin(4\pi(t_i - 0.5)) + \epsilon_i \quad 1 \leq i \leq n,$$

where  $\beta^T = (2, 2)$ . For the covariates  $(\mathbf{v}_i^T, t_i)^T$  we choose the following two distributions, for  $\rho = 0$  and  $0.25$ ,

-  $\mathcal{D}_{1,\rho}: \mathbf{x}_i = (\mathbf{v}_i^T, t_i)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\mu} = (0, 0, \frac{1}{2})^T$  and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & 1/(6\sqrt{3}) \\ \rho & 1 & 1/(6\sqrt{3}) \\ 1/(6\sqrt{3}) & 1/(6\sqrt{3}) & 1/36 \end{pmatrix}$$

-  $\mathcal{D}_{2,\rho}: \mathbf{v}_i^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\mu} = (0, 0)^T$  and  $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$  while  $t_i \sim \mathcal{U}(0, 1)$ .

As in dimension 1, the proportion of missing was 0.5 for all the distributions, under  $\mathcal{M}(H_2)$ , since it only depends on the distribution of  $t$ , while under  $\mathcal{M}(H_1)$ , the proportion was 0.521 under  $\mathcal{D}_{1,\rho}$ , and 0.561 for the uniform covariates  $t$  generated by model  $\mathcal{D}_{2,\rho}$ .

As in dimension  $D = 1$ , we compute the true marginal values by simulating 100 samples of size  $10^6$  to ensure that the standard errors of the computed marginal measures used when the real ones are unknown do not exceed  $6 \times 10^{-4}$ . The average values are given in Table 3 and are taken as the target values when computing the mean square error and the bias of the different proposals. Again, it is worth noting that under  $\mathcal{E}_1$  both the median and the mean of the responses are 0; however, the values obtained for the median present some disagreement. In this case, in the comparisons the value 0 was taken as the true one, the same decision was made when comparing the estimators of the mean under  $\mathcal{E}_2$ .

Tables S.7 to S.12 in the supplementary file report the bias (BIAS), standard deviation (SD) and mean square errors (MSE) of the estimators mentioned above computed using

**Table 3** Target marginal values: mean values over replications of the estimated parameters, when  $D = 2$

	Mean	Median	$F^{-1}(0.10)$	$F^{-1}(0.25)$	$F^{-1}(0.75)$	$F^{-1}(0.90)$
$\mathcal{E}_1$						
$\mathcal{D}_{1,0}$	-0.0001	-0.0004	-4.4452	-2.4260	2.4257	4.4455
$\mathcal{D}_{1,0.25}$	-0.0001	-0.0005	-4.7849	-2.5822	2.5818	4.7851
$\mathcal{D}_{2,0}$	-0.0000	-0.0008	-3.7907	-1.9960	1.9957	3.7913
$\mathcal{D}_{2,0.25}$	-0.0001	-0.0009	-4.2012	-2.2121	2.2116	4.2019
$\mathcal{E}_2$						
$\mathcal{D}_{1,0}$	-0.0001	-0.0026	-4.4441	-2.4280	2.4233	4.4445
$\mathcal{D}_{1,0.25}$	-0.0001	-0.0026	-4.7834	-2.5837	2.5796	4.7848
$\mathcal{D}_{2,0}$	-0.0001	-0.0051	-3.7871	-1.9981	1.9923	3.7930
$\mathcal{D}_{2,0.25}$	-0.0001	-0.0042	-4.1985	-2.2137	2.2090	4.2033

samples of size  $n = 100$ , over  $nr = 1000$  replications. On the other hand, Figures S.13 to S.18 in the supplementary file plot the absolute value of the bias and the mean square errors under  $\mathcal{E}_1$ . The figures for the asymmetric distribution  $\mathcal{E}_2$  are given in Figures S.14 to S.24 in the supplementary file.

It is worth noticing that, for any marginal quantity, in most situations, the MSE of both estimators are larger under  $\mathcal{D}_{1,\rho}$  than under  $\mathcal{D}_{2,\rho}$ . As in the univariate case, this enlargement is mainly caused by an increase of the standard deviation even though in some situations larger biases are obtained under  $\mathcal{D}_{2,\rho}$ . The different correlations considered nor the asymmetry of the errors affect notably the performance of the estimators, although under  $\mathcal{D}_{2,\rho}$  slightly larger MSE values are obtained when  $\rho = 0.25$ . When looking at the bias, for the marginal mean and median, the convolution procedure is competitive or even better than that based on  $\widehat{Q}_{y,WS}$ . However, it should be noted that the estimators based on  $\widehat{Q}_{y,CONV}$  show an increased bias when estimating the 75% quantile, in particular, under  $\mathcal{D}_{1,\rho}$ . Furthermore, as shown for instance in Figure S.14, the estimators based on  $\widehat{Q}_{y,CONV}$  have smaller mean square errors than those based on the weighted simplified when estimating the median and the quantiles, while similar ones are also obtained when estimating the mean. This shows the advantage of this proposal. The main reason for the decrease in mean square error is that smaller standard errors are obtained with the convolution-based estimators. Besides, under  $\mathcal{D}_{2,\rho}$ , there is an important gain in mean square error when using  $\widehat{Q}_{y,CONV}$ .

### 4.3 Sensitivity analysis when $\beta \in \mathbb{R}$

A numerical experiment was carried out to study the sensitivity of the marginal estimators proposed in Sect. 3 to the percentage of missingness. The study was performed under the partial linear regression model (11) considered in Sect. 4.1. As above, the smoothing parameters involved were chosen through a cross-validation procedure.

For brevity, we only report the results when the errors distribution is  $\mathcal{E}_1$ , i.e.,  $\epsilon_i \sim N(0, \sigma^2)$  with  $\sigma^2 = 0.25$  and when the covariates distribution is  $\mathcal{D}_2$ , that is,  $v_i \sim$

$N(0, 1)$  and  $t_i \sim \mathcal{U}(0, 1)$ . In this case, the true distribution function  $F_y$  is given by (12), and therefore, the target values can be found in Table 1. Similar results are obtained in other situations. We performed  $nr = 1000$  replications with samples of size  $n = 100$ .

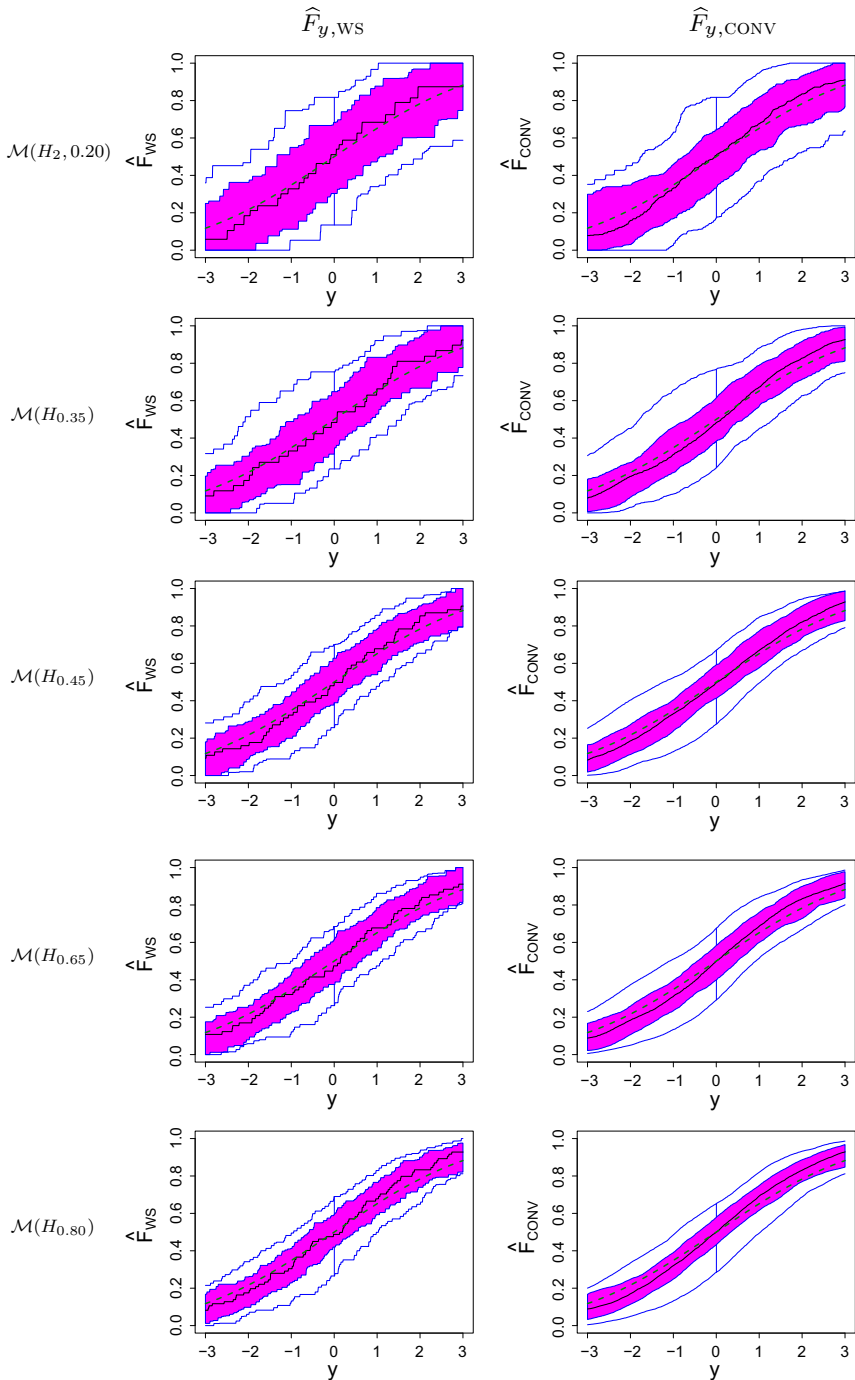
To conduct the sensitivity analysis, we generate samples with different proportions of missing data. To describe the missing models, let us consider the logistic function

$$\mathcal{L}_{a,b}(t) = \frac{1}{1 + \exp(-a - b(t - 0.5)^2)}.$$

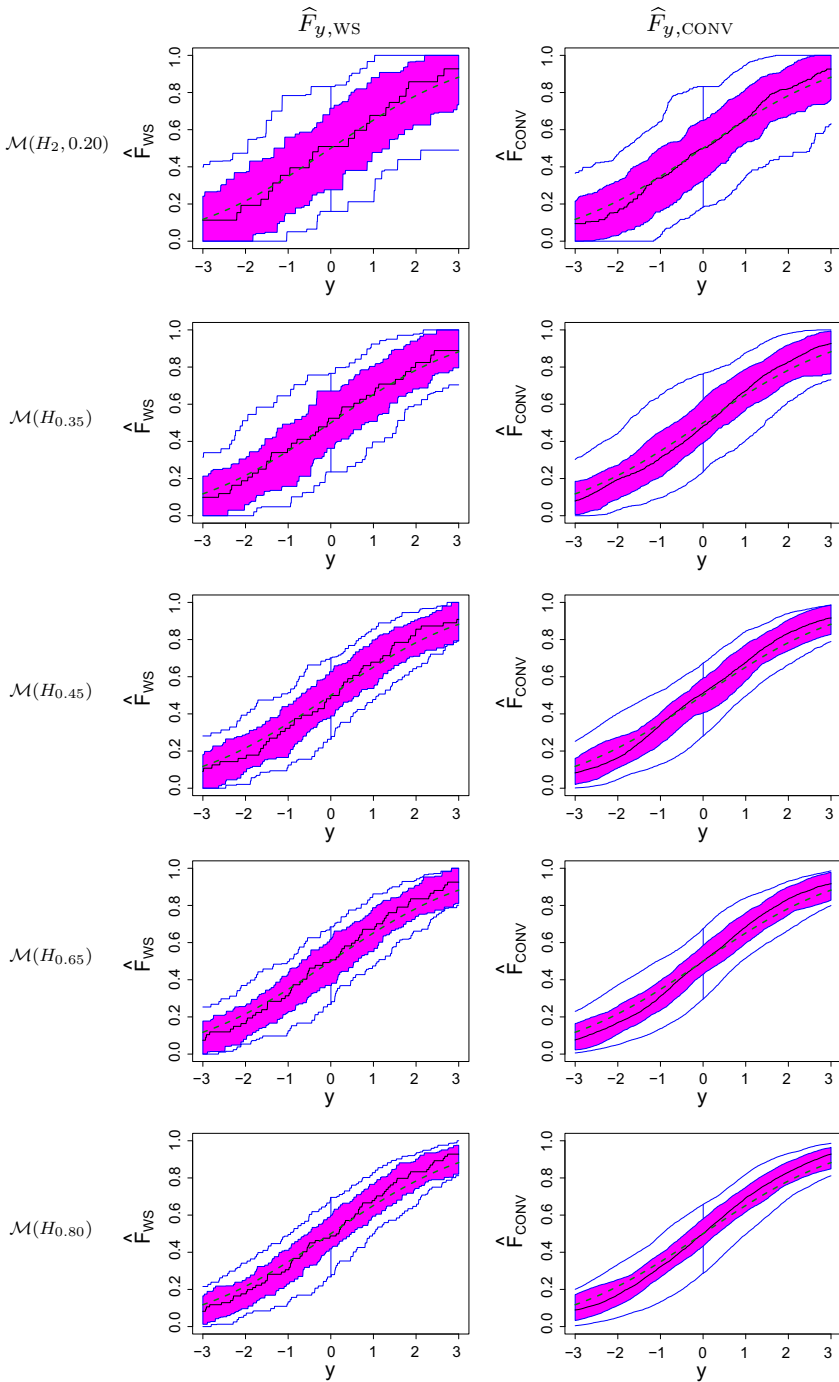
Missing responses and covariates were generated for different values of the parameters  $a$  and  $b$  yielding to a range of presence probabilities  $\pi = \mathbb{E}(\delta)$ . The selected pairs are  $(a, b) = (-1, -4.5), (-0.3, -4), (0, -3), (0.2, 5)$  and  $(1, 4.5)$  corresponding approximately to  $\pi = 0.20, 0.35, 0.45, 0.65$  and  $0.80$ . For the selected values of  $(a, b)$ , missing responses and covariates corresponding to the linear component are introduced according to  $\mathbb{P}(\delta_i = 1|t_i) = \mathcal{L}_{a,b}(t_i)$ . Henceforth, these propensity models will be denoted as  $\mathcal{M}(H_\pi)$  for  $\pi = 0.20, 0.35, 0.45, 0.65, 0.80$  with  $\pi = \mathbb{P}(\delta = 1)$ . Note that model  $\mathcal{M}(H_{0.45})$  coincides with  $\mathcal{M}(H_2)$ . Besides, as above, in all tables and figures  $\mathcal{M}(1)$  stands for the situation in which no missing data arise, that is,  $\mathbb{P}(\delta_i = 1|(y_i, \mathbf{x}_i)) = 1$  and is a benchmark that allows to study the loss of the different procedures on the selected missing schemes.

As mentioned in Sect. 4.1, besides the marginal quantile estimators, the responses distribution function has its own interest, since it gives a whole picture of the target variable behaviour. For this reason, we evaluated  $F_y$  and its estimates  $\widehat{F}_{y,WS}$  and  $\widehat{F}_{y,CONV}$  over an equally spaced grid of 1000 points between  $-3$  and  $3$ , that correspond approximately to the 10 and 90% of  $F_y$ . Figures 4, 5, 6 give the functional boxplots of the obtained estimates under  $\mathcal{M}(H_\pi)$ . The deepest curve is plotted in black, while the true distribution function  $F_y$  is given in dashed lines. These plots show that, as expected, when the percentage of missing data increases, i.e., when  $\pi$  decreases, the central band is enlarged making evident the higher variability of the estimation procedures. As noted in Sect. 4.1, the band defined by the 50% deepest curves corresponding to the convolution-based responses distribution estimators are narrower and smoother than those obtained with the weighted simplified procedure in all cases. Furthermore,  $\widehat{F}_{y,CONV}$  has a better performance in a neighbourhood of 0, while  $\widehat{F}_{y,WS}$  provides a better fit of  $F_y$  for large absolute values of  $y$ . This behaviour suggests that the extreme quantiles based on  $\widehat{F}_{y,WS}$  will be less biased, but with an increased standard deviation than those based on  $\widehat{F}_{y,CONV}$  as it can be observed in Tables 5, 6 and 9. It is worth noticing that this behaviour leads to larger values of the mean square error when considering the marginal estimates  $\widehat{\theta}_{WS}$ .

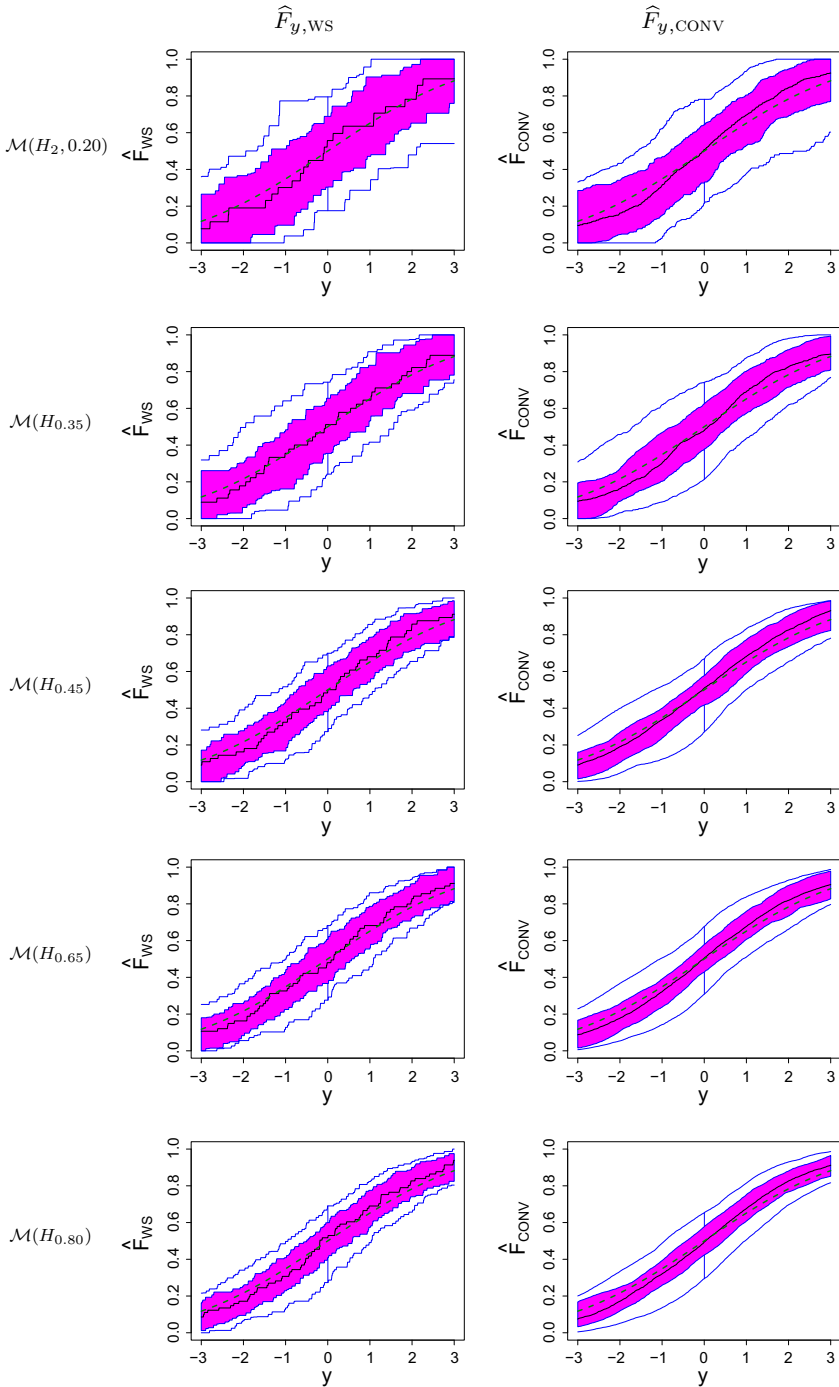
Tables 4, 5, 6, 7, 8 and 9 exhibit that, as expected, when the percentage of missing values increases (that is  $\pi$  decreases), the mean squared error and the standard deviation of the considered marginal estimators tend to raise up. The behaviour of the bias is slightly more erratic. However, when the propensity is estimated through a kernel fit, the bias tends to increase with the proportion of missing data. Regarding the behaviour of the mean squared errors of the estimators  $\widehat{\theta}_{WS}$  and  $\widehat{\theta}_{CONV}$ , it becomes evident that when  $\pi$  varies the estimations based on the convolution procedure attain the lowest mean square error values. Indeed, the mean squared error of the estimators based on



**Fig. 4** Functional boxplots of the responses distribution function estimators under  $\mathcal{E}_1, \mathcal{D}_2$  when  $\hat{p} = p$  and missing data arise according to  $\mathcal{M}(H_2, \pi)$ , where  $\pi = \mathbb{E}(\delta)$ . The dashed line corresponds to  $F_y$



**Fig. 5** Functional boxplots of the responses distribution function estimators under  $\mathcal{E}_1, \mathcal{D}_2$  when  $\hat{p} = \hat{p} \log$  and missing data arise according to  $\mathcal{M}(H_2, \pi)$ , where  $\pi = \mathbb{E}(\delta)$ . The dashed line corresponds to  $F_y$



**Fig. 6** Functional boxplots of the responses distribution function estimators under  $\mathcal{E}_1, \mathcal{D}_2$  when  $\hat{p} = \hat{p}K$  and missing data arise according to  $\mathcal{M}(H_2, \pi)$ , where  $\pi = \mathbb{E}(\delta)$ . The dashed line corresponds to  $F_y$

**Table 4** Summary measures for the marginal mean under  $\mathcal{D}_2$  and  $\mathcal{E}_1$ , when  $D = 1$

$\hat{p}$		BIAS		SD		MSE	
		$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$
$p$	$\mathcal{M}(1)$	-0.006	-0.006	0.227	0.227	0.052	0.052
	$\mathcal{M}(H_{0.20})$	-0.000	-0.001	0.491	0.492	0.242	0.242
	$\mathcal{M}(H_{0.35})$	-0.003	-0.003	0.385	0.385	0.148	0.148
	$\mathcal{M}(H_{0.45})$	-0.011	-0.011	0.291	0.291	0.085	0.085
	$\mathcal{M}(H_{0.65})$	-0.010	-0.010	0.279	0.279	0.078	0.078
$\hat{p}_{log}$	$\mathcal{M}(H_{0.80})$	-0.009	-0.009	0.253	0.253	0.064	0.064
	$\mathcal{M}(H_{0.20})$	0.005	0.004	0.513	0.515	0.263	0.265
	$\mathcal{M}(H_{0.35})$	-0.002	-0.003	0.392	0.393	0.154	0.154
	$\mathcal{M}(H_{0.45})$	-0.012	-0.012	0.291	0.291	0.085	0.085
	$\mathcal{M}(H_{0.65})$	-0.011	-0.011	0.279	0.279	0.078	0.078
$\hat{p}_K$	$\mathcal{M}(H_{0.80})$	-0.009	-0.009	0.253	0.253	0.064	0.064
	$\mathcal{M}(H_{0.20})$	-0.002	-0.004	0.470	0.470	0.221	0.221
	$\mathcal{M}(H_{0.350})$	-0.003	-0.004	0.372	0.372	0.138	0.138
	$\mathcal{M}(H_{0.45})$	-0.008	-0.008	0.285	0.285	0.081	0.081
	$\mathcal{M}(H_{0.65})$	-0.009	-0.009	0.274	0.274	0.075	0.075
	$\mathcal{M}(H_{0.80})$	-0.008	-0.008	0.251	0.251	0.063	0.063

**Table 5** Summary measures for the marginal median under  $\mathcal{D}_2$  and  $\mathcal{E}_1$ , when  $D = 1$

$\hat{p}$		BIAS		SD		MSE	
		$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$
$p$	$\mathcal{M}(1)$	-0.009	-0.008	0.281	0.257	0.079	0.066
	$\mathcal{M}(H_{0.20})$	-0.007	-0.012	0.623	0.573	0.388	0.328
	$\mathcal{M}(H_{0.35})$	-0.007	-0.007	0.486	0.440	0.237	0.194
	$\mathcal{M}(H_{0.45})$	-0.018	-0.018	0.366	0.333	0.134	0.111
	$\mathcal{M}(H_{0.65})$	-0.019	-0.014	0.350	0.320	0.123	0.103
$\hat{p}_{log}$	$\mathcal{M}(H_{0.80})$	-0.016	-0.012	0.314	0.287	0.099	0.082
	$\mathcal{M}(H_{0.20})$	-0.001	-0.008	0.653	0.602	0.426	0.362
	$\mathcal{M}(H_{0.35})$	-0.001	-0.007	0.490	0.449	0.240	0.201
	$\mathcal{M}(H_{0.45})$	-0.018	-0.019	0.363	0.332	0.132	0.110
	$\mathcal{M}(H_{0.65})$	-0.018	-0.015	0.349	0.320	0.122	0.103
$\hat{p}_K$	$\mathcal{M}(H_{0.80})$	-0.017	-0.012	0.314	0.287	0.099	0.082
	$\mathcal{M}(H_{0.20})$	-0.007	-0.019	0.605	0.555	0.366	0.308
	$\mathcal{M}(H_{0.35})$	-0.015	-0.011	0.477	0.429	0.227	0.184
	$\mathcal{M}(H_{0.45})$	-0.012	-0.015	0.361	0.326	0.130	0.106
	$\mathcal{M}(H_{0.65})$	-0.016	-0.013	0.344	0.315	0.118	0.099
	$\mathcal{M}(H_{0.80})$	-0.016	-0.010	0.311	0.285	0.097	0.081

**Table 6** Summary measures for the marginal 10% quantile  $\mathcal{D}_2$  and  $\mathcal{E}_1$ , when  $D = 1$

$\hat{p}$		BIAS		SD		MSE	
		$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$
$p$	$\mathcal{M}(1)$	0.004	0.019	0.378	0.342	0.143	0.117
	$\mathcal{M}(H_{0.20})$	0.025	0.070	0.882	0.796	0.778	0.639
	$\mathcal{M}(H_{0.35})$	0.019	0.037	0.663	0.594	0.440	0.355
	$\mathcal{M}(H_{0.45})$	0.013	0.033	0.524	0.455	0.275	0.208
	$\mathcal{M}(H_{0.65})$	0.001	0.027	0.491	0.425	0.241	0.181
$\hat{p}_{log}$	$\mathcal{M}(H_{0.80})$	0.010	0.025	0.448	0.383	0.201	0.147
	$\mathcal{M}(H_{1,0.20})$	0.043	0.080	0.892	0.804	0.798	0.652
	$\mathcal{M}(H_{1,0.35})$	0.020	0.030	0.675	0.610	0.456	0.373
	$\mathcal{M}(H_{1,0.50})$	0.008	0.030	0.526	0.458	0.276	0.211
	$\mathcal{M}(H_{1,0.65})$	-0.005	0.025	0.491	0.426	0.241	0.182
$\hat{p}_K$	$\mathcal{M}(H_{0.80})$	0.011	0.025	0.445	0.383	0.198	0.147
	$\mathcal{M}(H_{0.20})$	0.030	0.075	0.872	0.772	0.761	0.602
	$\mathcal{M}(H_{0.35})$	0.012	0.027	0.658	0.588	0.433	0.347
	$\mathcal{M}(H_{0.45})$	0.014	0.032	0.522	0.453	0.272	0.207
	$\mathcal{M}(H_{0.65})$	-0.001	0.027	0.480	0.420	0.230	0.177
	$\mathcal{M}(H_{0.80})$	0.010	0.026	0.447	0.382	0.200	0.147

**Table 7** Summary measures for the marginal 25% quantile  $\mathcal{D}_2$  and  $\mathcal{E}_1$ , when  $D = 1$

$\hat{p}$		BIAS		SD		MSE	
		$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$
$p$	$\mathcal{M}(1)$	-0.000	0.010	0.306	0.282	0.093	0.080
	$\mathcal{M}(H_{0.20})$	0.019	0.039	0.681	0.632	0.464	0.400
	$\mathcal{M}(H_{0.35})$	-0.009	0.013	0.521	0.484	0.272	0.235
	$\mathcal{M}(H_{0.45})$	-0.002	0.011	0.401	0.372	0.160	0.139
	$\mathcal{M}(H_{0.65})$	-0.007	0.011	0.379	0.347	0.144	0.120
$\hat{p}_{log}$	$\mathcal{M}(H_{0.80})$	-0.004	0.011	0.340	0.311	0.116	0.097
	$\mathcal{M}(H_{1,0.20})$	0.027	0.041	0.709	0.650	0.503	0.424
	$\mathcal{M}(H_{1,0.35})$	-0.011	0.007	0.528	0.489	0.279	0.239
	$\mathcal{M}(H_{1,0.50})$	-0.007	0.008	0.401	0.372	0.161	0.138
	$\mathcal{M}(H_{1,0.65})$	-0.007	0.009	0.380	0.346	0.144	0.120
$\hat{p}_K$	$\mathcal{M}(H_{0.80})$	-0.005	0.011	0.339	0.311	0.115	0.097
	$\mathcal{M}(H_{0.20})$	0.026	0.034	0.672	0.616	0.453	0.380
	$\mathcal{M}(H_{0.35})$	-0.009	0.009	0.509	0.472	0.259	0.223
	$\mathcal{M}(H_{0.45})$	0.001	0.013	0.398	0.368	0.158	0.136
	$\mathcal{M}(H_{0.65})$	-0.002	0.013	0.375	0.342	0.141	0.117
	$\mathcal{M}(H_{0.80})$	-0.001	0.013	0.336	0.310	0.113	0.096



**Table 8** Summary measures for the marginal 75% quantile  $\mathcal{D}_2$  and  $\mathcal{E}_1$ , when  $D = 1$

$\hat{p}$		BIAS		SD		MSE	
		$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$
$p$	$\mathcal{M}(1)$	-0.017	-0.021	0.292	0.273	0.086	0.075
	$\mathcal{M}(H_{0.20})$	-0.017	-0.028	0.656	0.603	0.431	0.364
	$\mathcal{M}(H_{0.35})$	-0.008	-0.015	0.511	0.469	0.262	0.221
	$\mathcal{M}(H_{0.45})$	-0.025	-0.033	0.385	0.354	0.149	0.126
	$\mathcal{M}(H_{0.20})$	-0.021	-0.032	0.365	0.336	0.133	0.114
$\hat{p}_{log}$	$\mathcal{M}(H_{0.80})$	-0.021	-0.031	0.326	0.302	0.107	0.092
	$\mathcal{M}(H_{1,0.20})$	-0.009	-0.022	0.704	0.643	0.496	0.413
	$\mathcal{M}(H_{1,0.35})$	0.002	-0.008	0.536	0.488	0.287	0.239
	$\mathcal{M}(H_{1,0.50})$	-0.024	-0.032	0.383	0.354	0.147	0.126
	$\mathcal{M}(H_{1,0.65})$	-0.023	-0.032	0.366	0.336	0.134	0.114
$\hat{p}_K$	$\mathcal{M}(H_{0.80})$	-0.021	-0.031	0.326	0.302	0.106	0.092
	$\mathcal{M}(H_{0.20})$	-0.014	-0.031	0.640	0.577	0.409	0.334
	$\mathcal{M}(H_{0.35})$	-0.001	-0.010	0.495	0.454	0.245	0.206
	$\mathcal{M}(H_{0.45})$	-0.018	-0.027	0.377	0.347	0.143	0.121
	$\mathcal{M}(H_{0.65})$	-0.021	-0.031	0.361	0.330	0.131	0.110
	$\mathcal{M}(H_{0.80})$	-0.019	-0.030	0.326	0.301	0.107	0.092

**Table 9** Summary measures for the marginal 90% quantile  $\mathcal{D}_2$  and  $\mathcal{E}_1$ , when  $D = 1$

$\hat{p}$		BIAS		SD		MSE	
		$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$	$\hat{Q}_{y,WS}$	$\hat{Q}_{y,CONV}$
$p$	$\mathcal{M}(1)$	-0.013	-0.028	0.363	0.330	0.132	0.110
	$\mathcal{M}(H_{0.20})$	-0.049	-0.054	0.834	0.745	0.698	0.558
	$\mathcal{M}(H_{0.35})$	-0.033	-0.042	0.631	0.574	0.399	0.331
	$\mathcal{M}(H_{0.45})$	-0.025	-0.045	0.500	0.438	0.250	0.194
	$\mathcal{M}(H_{0.65})$	-0.028	-0.042	0.460	0.412	0.212	0.171
$\hat{p}_{log}$	$\mathcal{M}(H_{0.80})$	-0.020	-0.038	0.420	0.365	0.177	0.135
	$\mathcal{M}(H_{1,0.20})$	-0.047	-0.053	0.885	0.788	0.785	0.624
	$\mathcal{M}(H_{1,0.35})$	-0.020	-0.033	0.658	0.596	0.433	0.356
	$\mathcal{M}(H_{1,0.50})$	-0.018	-0.044	0.499	0.441	0.250	0.196
	$\mathcal{M}(H_{1,0.65})$	-0.029	-0.043	0.460	0.412	0.212	0.171
$\hat{p}_K$	$\mathcal{M}(H_{0.80})$	-0.021	-0.037	0.422	0.367	0.179	0.136
	$\mathcal{M}(H_{0.20})$	-0.038	-0.061	0.818	0.732	0.670	0.540
	$\mathcal{M}(H_{0.35})$	-0.007	-0.029	0.645	0.574	0.416	0.330
	$\mathcal{M}(H_{0.45})$	-0.020	-0.038	0.495	0.439	0.245	0.194
	$\mathcal{M}(H_{0.65})$	-0.021	-0.040	0.457	0.410	0.209	0.170
	$\mathcal{M}(H_{0.80})$	-0.020	-0.036	0.416	0.365	0.174	0.134

**Table 10** Refinery data: marginal estimators with the full data set

$\theta$	$\mathbb{E}(y)$	$\theta_{0.5}$	$\theta_{0.10}$	$\theta_{0.25}$	$\theta_{0.75}$	$\theta_{0.90}$
$\hat{\theta}$	91.542	91.610	90.320	90.710	92.300	92.680
$\widehat{s}_{\hat{\theta},J}$	0.103	0.000	0.106	0.187	0.299	0.186
$\widehat{s}_{\hat{\theta},B}$	0.104	0.180	0.134	0.186	0.118	0.090
$\widehat{s}_{\hat{\theta},AS}$	0.103	0.175	0.142	0.176	0.129	0.112

the weighted simplified procedure may be up to 25% larger than the corresponding counterpart based on the convolution method.

## 5 Real data analysis

In this section, we consider two data sets. In the first one, missing data are introduced using a logistic model, while in the second one, the original data set already includes missing responses and covariates.

### 5.1 Refinery data

Daniel and Wood (1980) considered a data set obtained in a process variable study of a refinery unit. The response variable  $y$  is the octane number of the final product, while the covariates  $\mathbf{v} = (v_1, v_2, v_3)^T$  represent the feed compositions and the covariate  $t$  is the logarithm of a combination of process conditions scaled to  $[0, 1]$ . Bianco et al. (2010) fitted to this data set the partial linear model (9).

Daniel and Wood (1980) discussed the presence of three anomalous observations (labelled 75–77) which correspond to high values of octanes associated with high leverage points. Besides these three observations that influence the mean estimates, we have also exclude the observations labelled 71–74 since they correspond to isolated values of  $t$  enlarging the value of the cross-validation bandwidth. Hence, the whole data set has  $n = 75$  observations. We used the Epanechnikov kernel when considering semiparametric estimators.

We first compute the estimates of the marginal parameter  $\theta$  for this data set. The obtained values  $\hat{\theta}$  are reported in Table 10 together with the standard deviations  $\widehat{s}_{\hat{\theta}}$  evaluated using jackknife (with a subscript J), bootstrap (with a subscript B) and estimated using its asymptotic variance (with a subscript AS). For simplicity in the tables we denote as  $\theta_\alpha$  the marginal  $\alpha$ -quantile. It is worth noticing that the standard error of the median estimated using jackknife equals 0 since three observations in this data set are equal to the median. For the mean and the 0.25-quantile estimators, the jackknife and the asymptotic standard deviations are quite close. On the other hand, as expected the bootstrap and the asymptotic standard errors are close to each other. For that reason, when missing data are introduced we only report bootstrap standard deviations. Due to the asymmetry of the marginal distribution, the quantiles are not

**Table 11** Refinery data: marginal k-estimates when missing values are generated in the responses and covariates according to the model  $p(v_2, v_3, t) = 1/(1 + \exp(-1.3(t - 1.58)))$  and the propensity is estimated using a kernel estimator

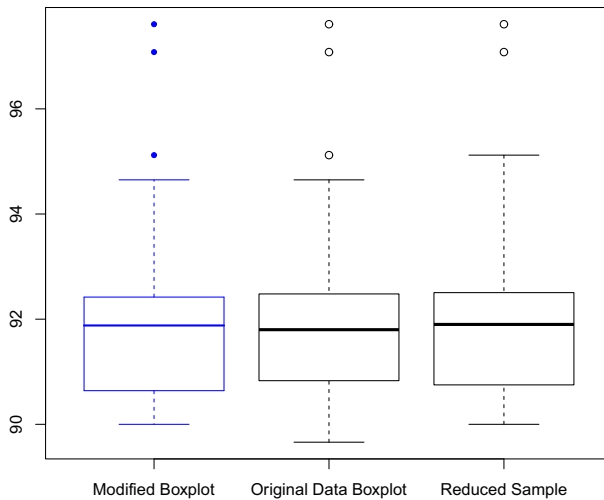
	Mean	$\hat{\theta}_{0.5}$	$\hat{\theta}_{0.10}$	$\hat{\theta}_{0.25}$	$\hat{\theta}_{0.75}$	$\hat{\theta}_{0.90}$
$\hat{\theta}_{WS}$	91.722 (0.122)	91.880 (0.181)	90.570 (0.229)	91.100 (0.255)	92.220 (0.129)	92.680 (0.167)
$\hat{\theta}_{CONV}$	91.734 (0.059)	91.730 (0.083)	90.798 (0.153)	91.203 (0.102)	92.274 (0.099)	92.692 (0.139)
$r_{WS}$	1.002	1.003	1.003	1.004	0.999	1
$r_{CONV}$	1.002	1.001	1.005	1.005	1.000	1

Bootstrap standard deviations are reported in the first two lines between brackets

symmetric with respect to the median and this asymmetry also influences the standard deviations estimators.

Missing observations in the responses and the covariate  $v_1$  were artificially introduced at random according to the following model  $p(v_2, v_3, t) = 1/(1 + \exp(-1.3(t - 1.58)))$ . Under this propensity model 42 observations are missing. The marginal estimators described in Sect. 3 were computed using a partial linear model when the propensity is computed through a kernel estimator. Their values are reported in Table 11 together with their standard deviations computed using the bootstrap procedure, between brackets. With respect to the bootstrap method, we split the sample in two groups corresponding to  $\delta = 0$  and  $\delta = 1$  and we resample within each group. Furthermore, we also report the values  $r_{WS}$  and  $r_{CONV}$  that correspond to the ratio between the marginal estimators computed with the full data set and those obtained with the reduced sample using the weighted simplified and the convolution procedure, respectively. Except for the 10 and 25%—quantiles the ratios  $r_{CONV}$  are closer to 1 than those based on the weighted simplified estimators. This fact together with the smaller standard errors obtained makes the convolution-based estimator a preferable choice.

As an illustration on the usefulness of the proposed estimators, over those obtained considering just the observed responses, Fig. 7 shows the boxplot obtained with the original 82 observations, labelled *Original Data Boxplot*, the corresponding one with the observed responses after introducing missing responses according to the missing probability  $p(t) = 1/(1 + \exp(-1.3(t - 1.58)))$ , labelled *Reduced Sample* and in blue the boxplot obtained using the weighted simplified quantiles. With this missing probability almost half of the data were missing. The missing probability was estimated through a kernel estimator. Note that the modified boxplot taking into account the propensity detects the three atypical observations, while the naive one based only on the reduced sample without any weighting does not point out the observation labelled as 77, even when the box is slightly enlarged when missing responses arise. The obtained values for the weighted simplified first and third quartiles are 90.64 and 92.42, while for the original sample the first and third quartile equal 90.83 and 92.48. On the other hand, the weighted simplified median equals 91.88 and that of the original data set is 91.80. Note that the boxplot may be constructed also considering the convolution-based quantiles. However, taking into account that outliers arise in the responses, robust estimators of the regression parameter and the regression function



**Fig. 7** Refinery data: The boxplot on the left corresponds to that based on the weighted simplified quantiles, the central one is the boxplot of the original sample and the right hand side one corresponds to the boxplot obtained with the data set  $\{y_i : \delta_i = 1\}$

need to be considered. When missing data arise in the responses, robust proposals in partial linear models have been considered in Bianco et al. (2010), the extension to the situation of missing responses and covariates is beyond the scope of this paper.

## 5.2 Air quality data

In this second application we illustrate the performance of the proposed methods on a real data set with missing values at the responses and one of the covariates. Cleveland (1985) analyses a set of environmental data corresponding to air quality measurements registered at New York metropolitan area between May 1, 1973 and September 30, 1973. The data consist in 153 observations that record daily readings of four air quality variables. Cleveland finds a nonlinear trend between Ozone (ppb) and Wind Speed (mph), where Ozone ( $y$ ) decreases as Wind Speed increases. This behaviour is explained by the effect of ventilation that higher wind speeds produce.

Bianco and Spano (2017) propose an exponential growth model between Ozone and Wind Speed and apply a weighted  $MM$ -estimators of the regression coefficients. With this robust fit, they identify five outliers corresponding to observations labelled as 86, 100, 101, 121 and 126. These five observations are excluded from our analysis since they influence the least squares estimators. We also include in our analysis as a linear component the covariate Solar Radiation. An interesting feature of these data is that about 24% of the responses are missing, more precisely, 37 values of Ozone are missing, 7 values of the variable Solar Radiation are also missing, while the variable Wind Speed is completely observed. Hence, we fit a partly nonlinear model with an exponential component based on Wind Speed ( $x_1$ ) and that depends linearly on Solar

**Table 12** Air quality data: marginal estimates and its bootstrap standard deviations, in brackets

	Mean	$\hat{\theta}_{0.5}$
$\hat{\theta}_{WS}$	39.339 (2.847)	29.000 (3.821)
$\hat{\theta}_{CONV}$	39.337 (1.707)	32.960 (1.981)

Radiation ( $x_2$ ), given by

$$y = \theta_1 \exp(\theta_2 x_1) + \theta_3 + \theta_4 x_2 + \epsilon .$$

We fit the regression parameters using the classical least squares estimator and then, we compute the marginal estimators described in Sect. 3 using a kernel estimator of the propensity model based on the Epanechnikov function, where the bandwidth is chosen through a cross-validation criterion.

In this example we focus on the central location marginal parameters, that is, the marginal mean and median. Table 12 summarizes the obtained estimators of the marginal mean and median, together with their bootstrap standard deviations, between brackets. The standard deviations were computed using bootstrap. In the bootstrap procedure, we use 1000 bootstrap samples obtained by resampling separately the missing and the completely observed data. Furthermore, when using a regression model the bootstrap was performed through resampled residuals.

Table 12 shows the advantages of the convolution estimators in terms of accuracy. Regarding the estimators of the marginal mean, the weighted simplified and the convolution estimators are very close; however, the convolution estimator is much less deviated; indeed, the ratio between their standard deviations is almost 1.67. Concerning the marginal median, the standard deviation of the convolution estimator is almost the half of the deviation of the weighted estimator. In this case, the two estimators of the median look further away; however, the standard deviation of the difference computed by bootstrap is 3.295, showing that the difference between these estimators is not significant which suggests that the fitted regression model is accurate. Otherwise, significant differences between the weighted simplified and the convolution-based procedures are expected. Taking into account the standard deviations, the convolution-based estimators seem more reliable.

## 6 Final remarks

One advantage of the marginal estimators proposed in Sect. 3 is their capability to deal with missing data both in responses and on some of the covariates under a general missing scheme. Besides, the plug-in approach given allows to handle any marginal quantity  $T(F_y)$ , given through a continuous functional  $T$ . In particular, the convergence of the marginal distribution estimators given in Sect. 3 allows to obtain consistent estimators for marginal quantities other than those considered in our simulation study, for instance, for dispersion measures. As shown in our simulation study, when estimating the mean and  $\alpha$ -quantiles the convolution-based estimators are more accurate,

a fact that has been noticed by Müller (2009) for linear functionals, such as the mean, when missing data arise only in the responses.

### 7 Appendix

We begin by stating the following Lemma that will be useful in the proofs. Its proof follows similar arguments to those considered in Appendix III from Billingsley (1968). It has been derived in a more general setting for Polish spaces in Varadarajan (1958), see also Theorem 3.3.1 in Bali (2012). We include its proof for completeness.

**Lemma 7.1** *Let  $(\Omega, \mathcal{A}, \mathbb{P})$  be a probability space and  $Q_n$  and  $Q$  induced probability measures over  $\mathbb{R}$  such that for any borelian set  $A$ ,  $Q_n(A) \xrightarrow{a.s.} Q(A)$ , that is,  $Q_n(A) \rightarrow Q(A)$  except for a set  $\mathcal{N}_A \subset \Omega$  such that  $\mathbb{P}(\mathcal{N}_A) = 0$ . Then,  $\Pi(Q_n, Q) \xrightarrow{a.s.} 0$ .*

*Proof* Let us show that given  $j \in \mathbb{N}$ , there exists  $\mathcal{N}_j \subset \Omega$  such that  $\mathbb{P}(\mathcal{N}_j) = 0$  and, for any  $\omega \notin \mathcal{N}_j$ , there exists  $n_j(\omega) \in \mathbb{N}$  such that if  $n \geq n_j(\omega)$ , then  $\Pi(Q_n, Q) < 1/j$ .

Let  $\{A_i, 1 \leq i \leq k\}$  be a finite class of disjoint sets with diameter smaller than  $\frac{1}{2j}$  such that

$$Q\left(\bigcup_{i=1}^k A_i\right) > 1 - \frac{1}{2j}. \tag{13}$$

Denote by  $\mathcal{A}$  the class of all the sets that are obtained as a finite union of the  $A_i$ , i.e.,  $B \in \mathcal{A}$  if and only if there exists  $A_{i_1}, \dots, A_{i_\ell}$  such that  $B = \bigcup_{j=1}^\ell A_{i_j}$ . Note that  $\mathcal{A}$  has a finite number of elements  $s$ . For each  $1 \leq i \leq s$ , and  $B_i \in \mathcal{A}$ , let  $\mathcal{N}_{B_i} \subset \Omega$  with  $\mathbb{P}(\mathcal{N}_{B_i}) = 0$  such that if  $\omega \notin \mathcal{N}_{B_i}$ , then  $|Q_n(B_i) - Q(B_i)| \rightarrow 0$ . Define  $\mathcal{N}_j = \bigcup_{i=1}^s \mathcal{N}_{B_i}$ , then we get  $\mathbb{P}(\mathcal{N}_j) = 0$ .

Fix  $\omega \notin \mathcal{N}_j$ , then  $|P_n(B_i) - P(B_i)| \rightarrow 0$ , for  $1 \leq i \leq s$ . Hence, there exists  $n_j(\omega) \in \mathbb{N}$  such that for  $n \geq n_j(\omega)$  we have that  $|Q_n(B) - Q(B)| < \frac{1}{2j}$  for any  $B \in \mathcal{A}$ . Let us show that if  $n \geq n_j(\omega)$ , then  $\Pi(Q_n, Q) < 1/j$ .

Consider  $B$  a borelian set and let  $A$  be the union of all the sets  $A_i$  that intersect  $B$ . Note that  $A \in \mathcal{A}$  and so  $|Q_n(A) - Q(A)| < \frac{1}{2j}$ . Therefore,  $B \subset A \cup \left(\bigcup_{i=1}^k A_i\right)^c$  and  $A \subset B^{1/j}$ , where the last inclusion holds since the sets  $A_i$  have diameter smaller than  $\frac{1}{2j}$ . Thus, using (13), we get that

$$Q(B) \leq Q(A) + Q\left[\left(\bigcup_{i=1}^k A_i\right)^c\right] = Q(A) + 1 - Q\left[\left(\bigcup_{i=1}^k A_i\right)\right] < Q(A) + \frac{1}{2j},$$

which together with the fact that  $|Q_n(A) - Q(A)| < \frac{1}{2j}$  implies that  $Q(B) \leq Q(A) + \frac{1}{2j} < Q_n(A) + 1/j$ . Using that  $A \subset B^{1/j}$ , we get that  $Q_n(A) + 1/j \leq Q_n(B^{1/j}) + 1/j$ , so  $Q(B) < Q_n(B^{1/j}) + 1/j$  which implies that  $\Pi(Q_n, Q) < 1/j$ , as desired.

To conclude the proof, we will show that  $\Pi(Q_n, Q) \rightarrow 0$  except for a zero  $\mathbb{P}$ -measure set. Consider the previously defined sets  $\mathcal{N}_j$  and let  $\mathcal{N} = \bigcup_{j \in \mathbb{N}} \mathcal{N}_j$ . It is

clear that  $\mathbb{P}(\mathcal{N}) = 0$ . Thus, for any  $\omega \notin \mathcal{N}$ , we will have that for each  $j$  there exists  $n_j = n_j(\omega)$  such that  $\Pi(Q_n, Q) < 1/j$  if  $n \geq n_j$ , concluding the proof.  $\square$

*Proof of Theorem 3.1* We will begin by deriving that  $\Pi_K(\widehat{Q}_{y,WS}, Q_y) \xrightarrow{a.s.} 0$ . Using analogous arguments to those considered in Theorem 3 in Pollard (1984), it is enough to show that for any borelian set  $B$ ,  $\widehat{Q}_{y,WS}(B) \xrightarrow{a.s.} Q_y(B)$ . Denote as

$$\widetilde{Q}_{y,WS} = \frac{1}{n} \sum_{i=1}^n \zeta_i \Delta_{y_i},$$

with  $\zeta_i = \delta_i / \widehat{p}(\mathbf{z}_i)$ . Note that  $\widehat{Q}_{y,WS}(B) = \widetilde{Q}_{y,WS}(B) / \widetilde{Q}_{y,WS}(\mathbb{R})$ . We will show that

$$\widehat{Q}_{y,WS}(B) \xrightarrow{a.s.} Q_y(B). \tag{14}$$

Using (14) with  $B = \mathbb{R}$ , we get that  $\widetilde{Q}_{y,WS}(\mathbb{R}) = (1/n) \sum_{i=1}^n \zeta_i \xrightarrow{a.s.} 1$  which leads to  $\widehat{Q}_{y,WS}(B) \xrightarrow{a.s.} Q_y(B)$ .

Note that  $\widehat{Q}_{y,WS}(B) = S_{1n} + S_{2n}$  where

$$S_{1n} = \frac{1}{n} \sum_{i=1}^n \left[ \frac{1}{\widehat{p}(\mathbf{z}_i)} - \frac{1}{p(\mathbf{z}_i)} \right] \delta_i I_B(y_i)$$

$$S_{2n} = \frac{1}{n} \sum_{i=1}^n \frac{1}{p(\mathbf{z}_i)} \delta_i I_B(y_i).$$

Using assumptions i) and ii), we have that  $|S_{1n}| \xrightarrow{a.s.} 0$ . On the other hand, using the strong law of large numbers and the MAR assumption (3), we have that  $S_{2n} \xrightarrow{a.s.} Q_y(B)$ , since  $\mathbb{E}\delta_1/p(\mathbf{z}_1) = 1$ , concluding the proof of (14).

The proof of  $\Pi(\widehat{Q}_{y,WS}, Q_y) \xrightarrow{a.s.} 0$  is a direct consequence of Lemma 7.1 and the fact that for any borelian set  $B$ ,  $\widehat{Q}_{y,WS}(B) \xrightarrow{a.s.} Q_y(B)$ .  $\square$

*Proof of Theorem 3.2* We will begin by proving a). As in the proof of Theorem 3.1, to derive a) it is enough to show that  $\Pi(\widetilde{Q}_\mu, Q_\mu) \xrightarrow{a.s.} 0$ , where

$$\widetilde{Q}_\mu = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\widehat{p}(\mathbf{z}_i)} \Delta_{\widehat{\mu}(\mathbf{x}_i)}$$

The proof will be carried out in several steps by showing that

$$\Pi(\widetilde{Q}_\mu, \widetilde{Q}_\mu) \xrightarrow{a.s.} 0 \tag{15}$$

$$\Pi(\widetilde{Q}_\mu, Q_{n,\mu}) \xrightarrow{a.s.} 0 \tag{16}$$

$$\Pi(Q_{n,\mu}, Q_\mu) \xrightarrow{a.s.} 0, \tag{17}$$

where

$$\tilde{Q}_\mu = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{p(\mathbf{z}_i)} \Delta_{\hat{\mu}(\mathbf{x}_i)} \quad Q_{n,\mu} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{p(\mathbf{z}_i)} \Delta_{\mu(\mathbf{x}_i)} .$$

Denote as  $\mathcal{N} \subset \Omega$  the null-probability set such that for  $\omega \notin \mathcal{N}$ ,  $\sup_{\mathbf{z} \in \mathcal{S}_z} |\hat{p}(\mathbf{z}) - p(\mathbf{z})| \rightarrow 0$ , then using assumption i), we have that  $\inf_{\mathbf{z} \in \mathcal{S}_z} \hat{p}(\mathbf{z}) > i_p/2$ , for  $n \geq n_0$ .

Fix  $\omega \notin \mathcal{N}$ ,  $n \geq n_0$  and a bounded and continuous function  $f : \mathbb{R} \rightarrow \mathbb{R}$ . Therefore, we have that

$$\begin{aligned} \left| \mathbb{E}_{\tilde{Q}_\mu}(f) - \mathbb{E}_{Q_{n,\mu}}(f) \right| &= \left| \frac{1}{n} \sum_{i=1}^n \frac{\hat{p}(\mathbf{z}_i) - p(\mathbf{z}_i)}{\hat{p}(\mathbf{z}_i)p(\mathbf{z}_i)} \delta_i f(\hat{\mu}(\mathbf{x}_i)) \right| \\ &\leq \|f\|_\infty \frac{1}{n} \sum_{i=1}^n \frac{|\hat{p}(\mathbf{z}_i) - p(\mathbf{z}_i)|}{\hat{p}(\mathbf{z}_i)p(\mathbf{z}_i)} \\ &\leq \|f\|_\infty \frac{2}{i_p^2} \sup_{\mathbf{z} \in \mathcal{S}_z} |\hat{p}(\mathbf{z}) - p(\mathbf{z})| \end{aligned}$$

hence  $\mathbb{E}_{\tilde{Q}_\mu}(f) - \mathbb{E}_{Q_{n,\mu}}(f) \rightarrow 0$ , concluding the (15).

For any  $\epsilon > 0$ , there exists a compact sets  $\mathcal{K} \subset \mathbb{R}^d$  such that,  $\mathbb{P}(\mathbf{x}_1 \in \mathcal{K}) > 1 - i_p \epsilon/8$ . The Strong Law of Large Numbers and assumption (iii) entail that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\mathcal{K}^c}(\mathbf{x}_i) \xrightarrow{a.s.} \mathbb{P}(\mathbf{x}_1 \in \mathcal{K}^c) \quad \text{and} \quad \sup_{\mathbf{x} \in \mathcal{K}} |\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| \xrightarrow{a.s.} 0 .$$

Hence, there exists a set  $\mathcal{N}_1 \subset \Omega$  such that  $\mathbb{P}(\mathcal{N}_1) = 0$  and  $(1/n) \sum_{i=1}^n \mathbb{I}_{\mathcal{K}^c}(\mathbf{x}_i) \rightarrow \mathbb{P}(\mathbf{x}_1 \in \mathcal{K}^c)$  and  $\sup_{\mathbf{x} \in \mathcal{K}} |\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| \rightarrow 0$  for any  $\omega \notin \mathcal{N}_1$ .

Fix  $\omega \notin \mathcal{N} \cup \mathcal{N}_1$  and  $n \geq n_0$ . To derive (16), without loss of generality assume that  $\|f\|_\infty = 1$ . Note that

$$\begin{aligned} \left| \mathbb{E}_{\tilde{Q}_\mu}(f) - \mathbb{E}_{Q_{n,\mu}}(f) \right| &= \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{p(\mathbf{z}_i)} [f(\hat{\mu}(\mathbf{x}_i)) - f(\mu(\mathbf{x}_i))] \right| \\ &\leq \frac{1}{i_p} \frac{1}{n} \sum_{i=1}^n |f(\hat{\mu}(\mathbf{x}_i)) - f(\mu(\mathbf{x}_i))| \\ &\leq \frac{2}{i_p} \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\mathcal{K}^c}(\mathbf{x}_i) + \frac{1}{i_p} \frac{1}{n} \sum_{i=1}^n |f(\hat{\mu}(\mathbf{x}_i)) \\ &\quad - f(\mu(\mathbf{x}_i))| \mathbb{I}_{\mathcal{K}}(\mathbf{x}_i) = S_{1,n} + S_{2,n} \end{aligned}$$

As  $(1/n) \sum_{i=1}^n \mathbb{I}_{\mathcal{K}^c}(\mathbf{x}_i) \rightarrow \mathbb{P}(\mathbf{x}_1 \in \mathcal{K}^c) < i_p \epsilon/4$ , there exists  $n_1 \geq n_0$  such that for  $n \geq n_1$ ,  $(1/n) \sum_{i=1}^n \mathbb{I}_{\mathcal{K}^c}(\mathbf{x}_i) < i_p \epsilon/4$ , hence  $S_{1,n} \leq \epsilon/2$ .



On the other hand, the uniform continuity of  $f$  over  $\mathcal{K}$  implies that there exists  $\delta$  such that  $|u - v| < \delta$ ,  $u, v \in \mathcal{K}$  entails  $|f(u) - f(v)| < i_p \epsilon/2$ . As  $\sup_{\mathbf{x} \in \mathcal{K}} |\widehat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| \rightarrow 0$ , there exists  $n_2 \geq n_1$  such that  $\sup_{\mathbf{x} \in \mathcal{K}} |\widehat{\mu}(\mathbf{x}) - \mu(\mathbf{x})| < \delta$ , so  $\sup_{\mathbf{x} \in \mathcal{K}} |f(\widehat{\mu}(\mathbf{x})) - f(\mu(\mathbf{x}))| < i_p \epsilon/2$  leading to  $S_{2,n} < \epsilon/2$ . Therefore,  $\left| \mathbb{E}_{\widetilde{Q}_\mu} (f) - \mathbb{E}_{Q_{n,\mu}} (f) \right| < \epsilon$  for any  $n \geq n_2$  and  $\omega \notin \mathcal{N} \cup \mathcal{N}_1$ , concluding the proof of (16).

The proof of (17) follows immediately from Lemma 7.1 and the fact that the MAR assumption yields

$$\begin{aligned} \mathbb{E} \frac{\delta_1}{p(\mathbf{z}_1)} \mathbb{I}_A(\mu(\mathbf{x}_1)) &= \mathbb{E} \left\{ \mathbb{E} \frac{\delta_1}{p(\mathbf{z}_1)} \mathbb{I}_A(\mu(\mathbf{x}_1)) | (y_1, \mathbf{x}_1) \right\} \\ &= \mathbb{E} \left\{ \frac{1}{p(\mathbf{z}_1)} \mathbb{I}_A(\mu(\mathbf{x}_1)) \mathbb{E} \delta_1 | (y_1, \mathbf{x}_1) \right\} \\ &= \mathbb{E} \left\{ \frac{1}{p(\mathbf{z}_1)} \mathbb{I}_A(\mu(\mathbf{x}_1)) \mathbb{P}(\delta_1 = 1 | \mathbf{z}_1) \right\} = Q_\mu(A). \end{aligned}$$

To derive b), it is enough to prove that

$$\Pi(\widetilde{Q}_\epsilon, \widetilde{Q}_\epsilon) \xrightarrow{a.s.} 0 \tag{18}$$

$$\Pi(\widetilde{Q}_\epsilon, Q_\epsilon) \xrightarrow{a.s.} 0 \tag{19}$$

where  $\widetilde{Q}_\epsilon = \sum_{i=1}^n \kappa_i \Delta_{\epsilon_i}$ . The proof of (18) is analogous to that of (16).

The proof of (19) follows from Lemma 7.1 using that for any borelian set  $A$ ,  $\widetilde{Q}_\epsilon(A) \xrightarrow{a.s.} Q_\epsilon(A)$  since

$$\begin{aligned} \mathbb{E} \delta_1 \mathbb{I}_A(\epsilon_1) &= \mathbb{E} (\mathbb{E} \{ \delta_1 \mathbb{I}_A(\epsilon_1) | (y_1, \mathbf{x}_1) \}) = \mathbb{E} [\mathbb{I}_A(\epsilon_1) \mathbb{E} \{ \delta_1 | (y_1, \mathbf{x}_1) \}] \\ &= \mathbb{E} \mathbb{I}_A(\epsilon_1) p(\mathbf{z}_1) = \mathbb{E} \mathbb{I}_A(\epsilon_1) \mathbb{E} p(\mathbf{z}_1) \end{aligned}$$

since the errors are independent of the covariates. □

**Acknowledgements** The authors wish to thank two anonymous referees for valuable comments which led to an improved version of the original paper. This work was partially developed while Ana M. Bianco and Graciela Boente were visiting the Departamento de Estadística, Análisis Matemática e Optimización de la Universidad de Santiago de Compostela, Spain under the bilateral agreement between the Universidad de Buenos Aires and the Universidad de Santiago de Compostela. This research was partially supported by Grants PICT 2014-0351 from ANPCYT and 20020130100279BA from the Universidad de Buenos Aires, Argentina and also by the Spanish Projects MTM2013-41383P and MTM2016-76969P from the Ministry of Science and Innovation, Spain. A. Bianco and G. Boente also wish to thank the Minerva Foundation for its support to present some of this paper results at the International Conference on Robust Statistics 2017.

## References

Aerts M, Claeskens G, Hens N, Molenberghs G (2002) Local multiple imputation. *Biometrika* 89:375–388  
 Bahadur RR (1966) A note on quantiles in large samples. *Ann Math Stat* 37:577–580  
 Bali L (2012) Métodos robustos de estimación de componentes principales funcionales y el modelo de componentes principales comunes. Ph. Thesis. Universidad de Buenos Aires (in spanish). Available at <http://cms.dm.uba.ar/academico/carreras/doctorado/2012/tesisBali.pdf>.

- 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* 54:546–564
- Bianco A, Spano P (2017) Robust inference for nonlinear regression models. <https://doi.org/10.1007/s11749-017-0570-2>
- Billingsley P (1968) *Convergence of probability measures*. Wiley, New York
- Boente G, González-Manteiga W, Pérez-González A (2009) Robust nonparametric estimation with missing data. *J Stat Plan Inference* 139:571–592
- Bravo F (2015) Semiparametric estimation with missing covariates. *J Multivar Anal* 139:329–346
- Bravo F, Jacho-Chávez D (2016) Semiparametric quasi-likelihood estimation with missing data. *Commun Stat Theory Methods* 45:1345–1369
- Burton A, Altman DG (2004) Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* 91:4–8
- Chen H, Chen K (1991) Selection of the splined variables and convergence rates in a partial spline model. *Can J Stat* 19:323–339
- Chen Q, Ibrahim J, Chen M, Senchaudhuri P (2008) Theory and inference for regression models with missing responses and covariates. *J Multivar Anal* 99:1302–1331
- Chen J, Shao J (2000) Nearest neighbor imputation for survey data. *J Off Stat* 16:113–131
- Chen S, Van Keilegom I (2013) Estimation in semiparametric models with missing data. *Ann Inst Math Stat* 65:785–805
- Chen X, Wan A, Zhou Y (2015) Efficient quantile regression analysis with missing observations. *J Am Stat Assoc* 110:723–741
- Cheng PE (1994) Nonparametric estimation of mean functionals with data missing at random. *J Am Stat Assoc* 89:81–87
- Cheng PE, Chu CK (1996) Kernel estimation of distribution functions and quantiles with missing data. *Stat Sinica* 6:63–78
- Cleveland W (1985) *The elements of graphing data*. Bell Telephone Laboratories Inc., New Jersey
- Collomb G (1979) Conditions nécessaires et suffisantes de convergence uniforme d'un estimateur de la régression, estimation des dérivées de la régression. *Comptes Rendus Academie de Sciences de Paris* 228:161–163
- Daniel C, Wood F (1980) *Fitting equations to data: computer analysis of multifactor data*. Wiley, New York
- Díaz I (2017) Efficient estimation of quantiles in missing data models. *J Stat Plan Inference* 190:39–51
- Fernholz L (1993) Smoothed versions of statistical functionals. In: Morgenthaler S, Ronchetti E, Stahel W (eds) *New directions in statistical data analysis and robustness*. Birkhauser, Basel, pp 61–72
- Härdle W, Liang H, Gao J (2000) *Partially linear models*. Springer, Heidelberg
- Härdle W, Müller M, Sperlich S, Werwatz A (2004) *Nonparametric and semiparametric models*. Springer, Heidelberg
- He X, Zhu Z, Fung W (2002) Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* 89:579–590
- Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71:1161–1189
- Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 47:663–685
- Huber P, Ronchetti E (2009) *Robust statistics*. Wiley, New York
- Liang H, Wang S, Robins J, Carroll R (2004) Estimation in partially linear models with missing covariates. *J Am Stat Assoc* 99:357–367
- Little R (1992) Regression with missing X's: a review. *J Am Stat Assoc* 87:1227–1237
- Little R, Rubin D (2002) *Statistical analysis with missing data*. Wiley, New York
- Müller U (2009) Estimating linear functionals in nonlinear regression with responses missing at random. *Ann Stat* 37:2245–2277
- Pollard D (1984) *Convergence of stochastic processes*. Springer, New York
- Robinson P (1988) Root-n-consistent semiparametric regression. *Econometrica* 56:931–954
- Schumaker L (1981) *Spline functions: basic theory*. Wiley, New York
- Sued M, Yohai V (2013) Robust location estimation with missing data. *Can J Stat* 41:111–132
- Tukey JW (1977) *Exploratory data analysis*. Addison-Wesley, Reading
- Varadarajan VS (1958) On the convergence of sample probability distributions. *Sankyā Indian J Stat* 19:23–26

- Wang Q, Linton O, Härdle W (2004) Semiparametric regression analysis with missing response at random. *J Am Stat Assoc* 99:334–345
- Wang W, Rao J (2002) Empirical likelihood-based inference under imputation for missing response data. *Ann Stat* 30:896–924
- Yang SS (1985) A smooth nonparametric estimator of a quantile function. *J Am Stat Assoc* 80:1004–1011
- Yates F (1933) The analysis of replicated experiments when the field results are incomplete. *Empire J Exp Agric* 1:129–142
- Zhang Z, Chen Z, Troendle JF, Zhang J (2012) Causal inference on quantiles with an obstetric application. *Biometrics* 68:697–706
- Zhou Y, Wan ATK, Wang X (2008) Estimating equation inference with missing data. *J Am Stat Assoc* 103:1187–1199