Computational Statistics and Data Analysis 54 (2010) 546–564

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Estimation of the marginal location under a partially linear model with missing responses

Ana Bianco [a,b], Graciela Boente [a,b,*], Wenceslao González-Manteiga [c], Ana Pérez-González [d]

[a] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina

[b] CONICET, Argentina

[c] Universidad de Santiago de Compostela, Spain

[d] Universidad de Vigo, Spain

ARTICLE INFO

Article history:
Received 4 April 2008
Received in revised form 12 September 2009
Accepted 18 September 2009
Available online 29 September 2009

ABSTRACT

In this paper, we consider a semiparametric partially linear regression model where there are missing data in the response. We propose robust Fisher-consistent estimators for the regression parameter, for the regression function and for the marginal location parameter of the response variable. A robust cross-validation method is briefly discussed, although, from our numerical results, the marginal estimators seem not to be sensitive to the bandwidth parameter. Finally, a Monte Carlo study is carried out to compare the performance of the robust proposed estimators among themselves and also with the classical ones, for normal and contaminated samples, under different missing data models. An example based on a real data set is also discussed.

1. Introduction

Partially linear regression models assume that the regression function can be modeled linearly on some covariates, while it depends nonparametrically on some others. To be more precise, assume that we have a response $y_i \in \mathbb{R}$ and covariates or design points $(\mathbf{x}_i^{\mathrm{T}}, t_i)$ such that $\mathbf{x}_i \in \mathbb{R}^p$, $t_i \in [0, 1]$ satisfying

$$y_i = m(\mathbf{x}_i, t_i) + \epsilon_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(t_i) + \sigma_0\epsilon_i \quad 1 \le i \le n, \tag{1}$$

with the errors ϵ_i i.i.d., independent of $(\mathbf{x}_i^{\mathrm{T}}, t_i)$ such that $E(\epsilon_i) = 0$ and VAR $(\epsilon_i) = 1$. Note that m stands for the regression function which is modeled linearly on $\mathbf{x}$ and nonparametrically on t.

As is well known, most of the statistical methods in nonparametric and semiparametric regression models are designed for complete data sets and problems arise when there are missing observations. This is a common situation in biomedical or socioeconomic studies, for instance. Typical examples are found in the field of social sciences where non-responses in sample surveys occur very often, and also in physics and genetics (Meng, 2000), among other areas. Although there are many situations in which both the response and the explanatory variables are missing, we will focus our attention on those cases where missing data occur only in the responses. This situation arises in many biological experiments where the explanatory variables can be controlled. This pattern is common, for example, in the scheme of double sampling proposed by Neyman (1938), where first a complete sample is obtained and then some additional covariate values are computed since perhaps this is less expensive than obtaining more response values. In this paper, we will thus assume that missing data occur only in the response variables.

* Corresponding address: Instituto de Cálculo, Facultad de Ciencias Exactas y Natural, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina. Tel.: +54 11 45763375; fax: +54 11 45763375.
E-mail address: gboente@dm.uba.ar (G. Boente).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.09.028

In the regression setting, a common method is to impute the incomplete observations and then proceed to carry out the estimation of the conditional or unconditional mean of the response variable with the complete sample. The methods considered include linear regression (Yates, 1933), kernel smoothing (Cheng, 1994; Chu and Cheng, 1995), nearest neighbor imputation (Chen and Shao, 2000), semiparametric estimation (Wang et al., 2004), nonparametric multiple imputation (Aerts et al., 2002), applying the empirical likelihood over the imputed values (Wang and Rao, 2002), among others. For a nonparametric regression model, González-Manteiga and Pérez-González (2004) considered an approach based on local polynomials to estimate the regression function when the response variable y is missing but the covariate $\mathbf{x}$ is totally observed. Wang et al. (2004) considered inference on the mean of y under regression imputation of missing responses based on the semiparametric regression model (1). Under the setting considered in this paper, the missingness of y is allowed to depend on $(\mathbf{x}^{\mathrm{T}}, t)$. All the proposals considered up to now are very sensitive to anomalous observations since they are based on a local least squares approach. Recently, Boente et al. (2009) introduced a robust proposal for estimating the regression function under missingness in the response.

The goal of this paper is to introduce resistant estimators for the marginal location of y, say θ, under the partially linear model (1), when the response variable has missing observations but the covariates $(\mathbf{x}^{\mathrm{T}}, t)$ are totally observed. The paper is organized as follows. Section 2 introduces the robust semiparametric estimators. An algorithm for computing the given estimators is described in Section 3, while their consistency is discussed in Section 4. A simulation study is described in Section 5 while an example based on a real data set is discussed in Section 6. Concluding remarks are provided in Section 7. Finally, technical proofs are given in the Appendix.

2. Proposals

We will consider inference with an incomplete data set $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i, \delta_i\right)$, $1 \leq i \leq n$, where $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if y_i is missing and

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + g_0\left(t_i\right) + \sigma_0\, \epsilon_i \quad 1 \leq i \leq n, \tag{2}$$

with errors ϵ_i independent and identically distributed with symmetric distribution $F_0(\cdot)$.

Let $(Y, \mathbf{X}^{\mathrm{T}}, T, \delta)$ be a random vector with the same distribution as $(y_i, \mathbf{x}_i, t_i, \delta_i)$. Our aim is to estimate, with the data set at hand, the regression parameter and the regression function robustly to provide a robust estimator for the marginal location parameter. An ignorable missing data mechanism will be imposed by assuming that Y is missing at random (MAR), i.e., δ and Y are conditionally independent given $(\mathbf{X}, T)$, i.e.,

$$P\left(\delta = 1 | (Y, \mathbf{X}, T)\right) = P\left(\delta = 1 | (\mathbf{X}, T)\right) = p\left(\mathbf{X}, T\right). \tag{3}$$

We will consider kernel smoother weights for the nonparametric component given by

$$w_i(t) = \frac{K\left(\frac{t_i - t}{h_n}\right) \delta_i}{\sum_{j=1}^{n} K\left(\frac{t_j - t}{h_n}\right) \delta_j}, \tag{4}$$

with K a kernel function, i.e., a nonnegative integrable function on $\mathbb{R}$, and h_n the bandwidth parameter. Note that the kernel weights are modified, multiplying by the indicator of the missing variables, in order to adapt to the complete sample and avoid bias.

For the sake of completeness, we recall the classical proposals. The least squares regression estimators are defined by considering preliminary kernel estimators, $\widehat{\boldsymbol{\eta}}_n(t)$ and $\widehat{\eta}_{0,n}(t)$, of the quantities $\boldsymbol{\eta}(t) = E(\delta \mathbf{X} | T = t)/E(\delta | T = t)$ and $\eta_0(t) = E(\delta Y | T = t)/E(\delta | T = t)$, respectively. Note that using (3), δ is conditionally independent of Y and so we have that $\eta_0(t) = E(Y | T = t)$. Since $\delta Y = \delta \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta}_0 + \delta g_0(T) + \delta \sigma_0 \epsilon$, taking conditional expectation, we get $\eta_0(t) = \boldsymbol{\eta}(t)^{\mathrm{T}} \boldsymbol{\beta}_0 + g_0(t)$ if $E\left(\epsilon | (\mathbf{x}, t)\right) = 0$ and so $\delta_i \left(y_i - \eta_0(t_i)\right) = \delta_i \left(\mathbf{x}_i - \boldsymbol{\eta}(t_i)\right)^{\mathrm{T}} \boldsymbol{\beta}_0 + \delta_i \sigma_0 \epsilon_i$, $1 \leq i \leq n$. Then, the estimator of the regression parameter $\boldsymbol{\beta}_0$, introduced by Wang et al. (2004), is defined as the value minimizing $\sum_{i=1}^{n} \delta_i \left\{ \left(y_i - \widehat{\eta}_{0,n}(t_i)\right) - \left(\mathbf{x}_i - \widehat{\boldsymbol{\eta}}_n(t_i)\right)^{\mathrm{T}} \boldsymbol{\beta} \right\}^2$. This estimator is based on weighted means of the response variables and so it is highly sensitive to anomalous data. This suggests that some resistant estimation procedure needs to be considered.

It is worth noticing that Wang et al. (2004) assumed only that $E(\epsilon_i | (\mathbf{x}_i, t_i)) = 0$ instead of the independence between the errors and the covariates. However, the stronger independence assumption stated in (1) will be needed to obtain robust consistent estimators of $\boldsymbol{\beta}$ as in linear regression models.

2.1. Robust estimators of the regression parameter and regression function

The estimation of the robust location conditional functional related to each component of $\mathbf{x}_i$ causes no problem since all covariates are complete, while that of the response y_i is problematic since there are missing responses. We will consider the approach given in Boente et al. (2009) for estimating the regression functions. The simplified local M-smoother defined therein uses the information at hand and defines the estimator with the complete observations only. The main problem

is that if we proceed as in Bianco and Boente (2004) with the complete sample, the conditions needed to ensure Fisher-consistency entail that $p(\mathbf{X}, T) = p(T)$, which eliminates many situations arising in practice. Thus, to guarantee Fisher-consistency, a robust profile-likelihood approach will be considered.

Let ψ_1 be an odd and bounded score function and ρ be a *rho*-function as defined in Maronna et al. (2006, Chapter 2), i.e., a function ρ such that

- $\rho(x)$ is a nondecreasing function of $|x|$,
- $\rho(0) = 0$,
- $\rho(x)$ is increasing for $x > 0$ when $\rho(x) < \|\rho\|_\infty$,
- if ρ is bounded, it is also assumed that $\|\rho\|_\infty = 1$.

To define a robust estimator, we can proceed as follows:

- **Step 1**. For each t and $\boldsymbol{\beta}$, define $g_{\boldsymbol{\beta}}(t)$ and its related estimate $\widehat{g}_{\boldsymbol{\beta}}(t)$ using the simplified local M-smoothers defined in Boente et al. (2009). That is, $g_{\boldsymbol{\beta}}(t)$ and $\widehat{g}_{\boldsymbol{\beta}}(t)$ are, respectively, the solutions of

$$E\left[\delta\psi_1\left(\frac{Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - g_{\boldsymbol{\beta}}(t)}{\sigma}\right)\middle| T = t\right] = 0, \tag{5}$$

$$\sum_{i=1}^{n} w_i(t)\psi_1\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \widehat{g}_{\boldsymbol{\beta}}(t)}{\widehat{s}(t)}\right) = 0, \tag{6}$$

  where $\widehat{s}(t)$ is a preliminary robust consistent scale estimator.
- **Step 2**. The functional $\boldsymbol{\beta}(F)$, where F is the distribution of $(\delta, Y, \mathbf{X}^{\mathrm{T}}, T)$, is defined as

$$\boldsymbol{\beta}(F) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} E\left[\delta\rho\left(\frac{Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - g_{\boldsymbol{\beta}}(T)}{\sigma}\right)\upsilon(\mathbf{X})\right]$$

  and its related estimate as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i\rho\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \widehat{g}_{\boldsymbol{\beta}}(t_i)}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i),$$

  with $\widehat{\sigma}$ a preliminary estimate of the scale σ_0, i.e., a robust M-scale computed using an initial (possible inefficient) estimate of $\boldsymbol{\beta}$ with high breakdown point. Therefore, if $\psi = \rho'$ denotes the derivative of ρ, the estimator is the solution of

$$\sum_{i=1}^{n} \delta_i\psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \widehat{g}_{\widehat{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i)\left(\mathbf{x}_i + \frac{\partial}{\partial\boldsymbol{\beta}}\widehat{g}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right) = 0. \tag{7}$$

- **Step 3**. Then, the functional $g(t, F)$ is defined as $g(t, F) = g_{\boldsymbol{\beta}(F)}(t)$, while the estimate of the nonparametric component is $\widehat{g}_n(t) = \widehat{g}_{\widehat{\boldsymbol{\beta}}}(t)$.

An estimator of the regression function m is thus given by $\widehat{m}(\mathbf{x}, t) = \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t)$.

In the Appendix, it is shown that, under mild conditions, these functionals are Fisher-consistent since the errors ϵ are independent of (δ, T). Moreover, it will be shown that Fisher-consistency is preserved under the heteroscedastic model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(t_i) + \sigma_0(\mathbf{x}_i, t_i)\epsilon_i$, $1 \le i \le n$, where the errors ϵ_i are i.i.d. and independent of the covariates.

Remark 2.1.1. As in nonparametric regression without missing observations, the aim of a robust smoother, as the local M-estimator, is to provide reliable estimations when outlier observations are present in the responses y_i. Indeed, the researcher is seeking for consistent estimators of the regression functions $g_{\boldsymbol{\beta}}(t)$ and $m(\mathbf{x}, t)$ without requiring moment conditions on the errors ϵ_i. This includes the well-known α-contaminated neighborhood for the error distribution. More precisely, in a robust framework, one seeks procedures that remain valid when $\epsilon_i \sim F_0 \in \mathcal{F}_\alpha = \{G : G(y) = (1-\alpha)G_0(y) + \alpha H(y)\}$, with H any symmetric distribution and G_0 a central model with possible first or second moments. In fact, the same framework can be considered in this paper. In these neighborhoods, no moment conditions are required to the errors and outliers correspond to deviations in the error distribution. Moreover, the condition that ψ_1 is an odd function and the errors have a symmetric distribution can be replaced by $E(\psi_1(\epsilon/\sigma)) = 0$ and $E(\psi(\epsilon/\sigma)) = 0$, for any $\sigma > 0$, which are standard conditions in robustness in order to guarantee Fisher-consistency of the location or regression parameters. Further discussion can be found in He et al. (2002), Bianco et al. (2006) and Boente et al. (2009). Otherwise, if this assumption is not fulfilled the regression and location estimators are asymptotically biased; see for instance, Maronna et al. (2006, Chapter 4).

On the other hand, as in any regression model, leverage points in the explanatory variables $\mathbf{x}$ can cause breakdown. To overcome this problem, *GM*- and *S*-estimators have been introduced; see for instance, Maronna et al. (2006). In **Step 2**, we have considered a score function ρ combined with a weight υ to include both families of estimators. Our proposal is thus resistant against outliers in the residuals and in the carriers $\mathbf{x}$ as well.

2.2. Estimation of the marginal location

Let us denote by θ the marginal location of Y; for instance, we are interested in the M-location parameter of the response Y solution of $\lambda(a,\sigma) = E\psi_2((Y-a)/\sigma) = 0$ for all σ, where ψ_2 is an odd and bounded score function. When $\psi_2(u) = \mathrm{sg}(u) = I_{(0,\infty)}(u) - I_{(-\infty,0)}(u)$, θ is the median of Y. The same score functions ψ_1 and ψ_2 can be considered both in **Step 1** and when computing the marginal parameter estimators defined below.

Denote by $\widehat{\sigma}$, $\widehat{\sigma}_{\mathrm{WI}}$ and $\widehat{\sigma}_{\mathrm{A}}$ robust consistent estimators of the marginal scale of the variables involved, such as the MAD. Since we only have the responses at hand, the unknown values can be predicted using $\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i)$, where $\widehat{g}_n(t)$ and $\widehat{\boldsymbol{\beta}}$ are defined in Section 2.1. Besides, to correct the bias caused in the estimation by the missing data mechanism, an estimator of the missingness probability needs to be considered. Denote by $p_n(\mathbf{x}, t)$ any estimator of $p(\mathbf{x}, t)$, such as the nonparametric kernel estimator

$$p_n(\mathbf{x},t) = \frac{\sum_{i=1}^{n} K_1\left(\frac{\mathbf{x}_i-\mathbf{x}}{\lambda_n}\right) K_2\left(\frac{t_i-t}{b_n}\right)\delta_i}{\sum_{j=1}^{n} K_1\left(\frac{\mathbf{x}_j-\mathbf{x}}{\lambda_n}\right) K_2\left(\frac{t_j-t}{b_n}\right)}, \tag{8}$$

where $K_1 : \mathbb{R}^p \to \mathbb{R}$ and $K_2 : \mathbb{R} \to \mathbb{R}$ are kernel functions and λ_n and h_n denote the smoothing parameters. If a parametric model is assumed, other choices for estimating $p(\mathbf{x}, t)$ can be considered.

- **Weighted simplified M-estimate**. This estimate uses the complete sample and is the solution, $\widehat{\theta}_{\mathrm{WS}}$, of

$$\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} \psi_2\left(\frac{y_i - \widehat{\theta}_{\mathrm{WS}}}{\widehat{\sigma}}\right) = 0.$$

- **Averaged M-estimate**. This estimator uses the predicted values to compute the marginal parameter estimator. If the error distribution is symmetric, as assumed, and $Z = m(\mathbf{X}, T) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(T) = \theta + u$ with u having symmetric distribution, we get that the median of the distribution of Y equals the median of Z. The same will happen when considering an M-functional, that is, Y and Z will have the same M-location, and so we get the estimator, $\widehat{\theta}_{\mathrm{A}}$ as the solution of

$$\sum_{i=1}^{n} \psi_2\left(\frac{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i) - \widehat{\theta}_{\mathrm{A}}}{\widehat{\sigma}_{\mathrm{A}}}\right) = 0.$$

- **Weighted imputed M-estimate**. This estimator combines the ideas of the previous ones by imputing the missing responses. The estimate $\widehat{\theta}_{\mathrm{WI}}$ is the solution of

$$\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} \psi_2\left(\frac{y_i - \widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right) + \sum_{i=1}^{n} \left(1 - \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)}\right) \psi_2\left(\frac{\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i) - \widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right) = 0. \tag{9}$$

The Fisher-consistency of the related functionals is derived in the Appendix.

2.2.1. On the strong robustness

In the classical setting, the target parameter is the mean $\theta = E(Y)$. When considering $\psi_2(t) = \mathrm{sg}(t)$ the target is now the median of the response Y. For general score functions ψ_2, the target is the robust M-location functional related to ψ_2, as introduced in Huber (1981).

It is worth noticing that the assumption of symmetry required to the error's distribution is needed if we want to guarantee that we are estimating the same quantity when using all robust location functionals. As discussed in Remark 2.1.1, it can be replaced by $E(\psi_2(\epsilon/\sigma)) = 0$, for any $\sigma > 0$. Furthermore, the weak continuity of these robust location functionals for bounded score functions can be seen in Huber (1981). Therefore, by applying this functional to weak consistent estimators of the distribution of Y, we obtain consistent and asymptotically strongly robust estimators of θ. These results can, thus, be applied in our missing data setting by defining suitable empirical distributions.

Note that all the estimators introduced in Section 2.2 can be written as M-functionals applied to some modified empirical distribution. In fact, we have:

- $\widehat{\theta}_{\mathrm{WS}}$ is the solution of $\widehat{\lambda}_n(a, \widehat{\sigma}) = 0$ with

$$\widehat{\lambda}_n(a,\sigma) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} \psi_2\left(\frac{y_i - a}{\sigma}\right) = \int \psi_2\left(\frac{y-a}{\sigma}\right) \mathrm{d}\widehat{F}_n(y),$$

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} I_{(-\infty,y]}(y_i).$$

From the above discussion and since $\widehat{F}_n(y)$ provides weak consistent estimators of $F(y)$ under mild conditions, the weighted simplified M-estimator provides asymptotically strongly robust consistent estimators.

- Denote by $\widehat{m}_i$ the predicted values using the partially linear model (1), $\widehat{m}_i = \widehat{m}(\mathbf{x}_i, t_i) = \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i)$. Then, $\widehat{\theta}_{\mathrm{A}}$ is the solution of $\widetilde{\lambda}_n(a, \widehat{\sigma}_{\mathrm{A}}) = 0$,

$$\widetilde{\lambda}_n(a,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\psi_2\left(\frac{\widehat{m}_i - a}{\sigma}\right) = \int \psi_2\left(\frac{z-a}{\sigma}\right)\mathrm{d}\widetilde{F}_n(z),$$

$$\widetilde{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,z]}(\widehat{m}_i).$$

  In this case, if $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}$ are robust consistent estimators of $\boldsymbol{\beta}_0$ and g_0, $\widetilde{F}_n$ will be a weak consistent estimator of the distribution, F_Z, of $Z = m(\mathbf{X}, T)$. Thus, the average M-estimators are a sequence of asymptotically strongly robust consistent estimators.
- $\widehat{\theta}_{\mathrm{WI}}$ is the solution of $\widehat{\widetilde{\lambda}}_n(a, \widehat{\sigma}_{\mathrm{WI}}) = 0$ with

$$\widehat{\widetilde{\lambda}}_n(a,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{p_n(\mathbf{x}_i,t_i)}\psi_2\left(\frac{y_i-a}{\sigma}\right) + \left(1-\frac{\delta_i}{p_n(\mathbf{x}_i,t_i)}\right)\psi_2\left(\frac{\widehat{m}_i-a}{\sigma}\right) = \int \psi_2\left(\frac{y-a}{\sigma}\right)\mathrm{d}\widehat{\widetilde{F}}_n(y),$$

$$\widehat{\widetilde{F}}_n(y) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i}{p_n(\mathbf{x}_i,t_i)}I_{(-\infty,y]}(y_i) + \left(1-\frac{\delta_i}{p_n(\mathbf{x}_i,t_i)}\right)I_{(-\infty,y]}(\widehat{m}_i)\right].$$

  As will be shown, $\widehat{\widetilde{F}}_n$ is a weak consistent estimator of F if $p_n(\mathbf{x}, t)$ is a consistent estimator of $p(\mathbf{x}, t) = P(\delta = 1|(\mathbf{X}, T) = (\mathbf{x}, t))$. Thus, the weighted imputed M-estimators provide a sequence of asymptotically strongly robust consistent estimators.

2.2.2. Selection of the smoothing parameter

Although the bias and variance of the marginal location estimators are less sensitive to the bandwidth than in other semiparametric settings (see for instance Cheng (1994) and Wang and Sun (2007)), a least squares cross-validation scheme for choosing the smoothing parameter is usually considered. Besides, the sensitivity of L^2 cross-validation methods in nonparametric regression was pointed out by, among others, Wang and Scott (1994) and Cantoni and Ronchetti (2001) who also proposed robust alternatives, while Boente et al. (2009) considered the case when missing responses are present.

The ideas of robust cross-validation can be adapted to the present situation. Let $\widehat{\theta}$ be the robust estimator to be considered, i.e., the average or the weighted imputed one. Denote by $\widehat{\theta}^{(i)}(h)$ the estimator computed with bandwidth h using all the data except $(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i)$. Taking into account that the L^2 cross-validation criterion tries to measure both bias and variance, it would be sensible to introduce, as in Bianco and Boente (2007), a new measure that establishes a trade-off between robust measures of bias and variance. Let μ_n and σ_n denote robust estimators of location and scale, respectively. A robust cross-validation criterion can be defined by minimizing on h

$$RCV_{\mathrm{R}}(h) = \mu_n^2\left(\widehat{r}_i(h), w(t_i)\right) + \sigma_n^2\left(\widehat{r}_i(h), w(t_i)\right), \tag{10}$$

where the weight function w may be chosen so as to protect against boundary effects, $\widehat{r}_i(h) = y_i - \widehat{\theta}^{(i)}(h)$ are the residuals and $\mu_n(u_i, w_i)$ and $\sigma_n(u_i, w_i)$ indicate that to compute the robust location and scale, respectively, each observation u_i receives a weight w_i. As location estimator, μ_n, one can consider the median, while σ_n can be taken as the bisquare a-scale estimator or the Huber τ-scale estimator. For the situation that we are dealing with, it is enough to compute RCV_{R} with the observations at hand, i.e., to compute RCV_{R} we use only the observed residuals $\{\widehat{r}_i\}_{i:\delta_i=1}$ and discard the incomplete vectors.

3. Algorithm

3.1. Computation of the parametric and nonparametric components

We will consider kernel smoother weights for the nonparametric component which are given by (4). In this section, we describe an algorithm which is a slight modification of the procedure described in Maronna et al. (2006, Chapter 5). Let ρ_0 and ρ be two bounded *rho*-functions such that $\rho_0 \geq \rho$.

When $\boldsymbol{\beta} \in \mathbb{R}^p$, $p = 1, 2$, the algorithm for computing the estimator $\widehat{\boldsymbol{\beta}}$ defined in **Step 2** may be based on a search over a grid of points as follows.

- **Step A0**. Take a net $\boldsymbol{\beta}_j$ of possible values for $\boldsymbol{\beta}, j = 1, \ldots, J$.
- **Step A1**. Fix $1 \leq j \leq J$. We first compute the regression function estimate $\widehat{g}_{\boldsymbol{\beta}}(t)$ for each $\boldsymbol{\beta} = \boldsymbol{\beta}_j$ of the net and each t_i and also, an estimator for the scale σ_0.
  - ⋆ For any $1 \leq i \leq n$, evaluate $\widehat{g}_{j,i} = \widehat{g}_{\boldsymbol{\beta}_j}(t_i)$ using the simplified M-estimator introduced by Boente et al. (2009) applied to $\{(y_k - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}_j, t_k, \delta_k)\}_{1\leq k\leq n}$ i.e., as the solution of

$$\sum_{k=1}^{n} w_k(t_i)\psi_1\left(\frac{y_k - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}_j - \widehat{g}_{j,i}}{\widehat{s}(t_i)}\right) = 0, \tag{11}$$

where $\widehat{s}(t_i)$ is a preliminary robust scale estimator, such as the local MAD, i.e., $\widehat{s}(t) = \mathrm{mad}_{k\in\mathcal{I}(h_n)}|r_{k,j} - \mathrm{median}_{\ell\in\mathcal{I}(h_n)}(r_{\ell,j})|$ with $r_{k,j} = y_k - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}_j$ and $\mathcal{I}(h_n) = \{\ell : 1 \le \ell \le n, \delta_\ell = 1 \text{ and } |t_\ell - t| \le h_n\}$.

⋆ Compute

$$L_0(\boldsymbol{\beta}_j) = \underset{1\le i\le n:\delta_i=1}{\mathrm{median}}\left(\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j - \widehat{g}_{j,i}\right)^2\right).$$

- **Step A2**. In order to define the residuals scale estimator, let $\widehat{\boldsymbol{\beta}}_{\mathrm{INI}} = \beta_{j_0}$ be the preliminary estimator of $\boldsymbol{\beta}$ such that

$$\widehat{\boldsymbol{\beta}}_{\mathrm{INI}} = \underset{j}{\mathrm{argmin}}\, L_0(\boldsymbol{\beta}_j) = L_0(\boldsymbol{\beta}_{j_0}),$$

and, for each $1 \le i \le n$, let $\widehat{g}_{\mathrm{INI},i}$ be the solution of (11) when using $\widehat{\boldsymbol{\beta}}_{\mathrm{INI}}$. Note that there is no need to evaluate again the solutions $\widehat{g}_{\mathrm{INI},i}$, since they were already computed in **Step A1** for all the values of $\boldsymbol{\beta}$ in the grid.

Let $m = \sum_{i=1}^{n}\delta_i$. The estimator of the scale σ, $\widehat{\sigma}$, is then defined as the solution of

$$\frac{1}{m}\sum_{i=1}^{n}\delta_i\rho_0\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{INI}} - \widehat{g}_{\mathrm{INI},i}}{\widehat{\sigma}}\right) = \frac{1}{2}. \tag{12}$$

- **Step A3**. To compute the final estimator of $\boldsymbol{\beta}$, let

$$L(\boldsymbol{\beta}_j) = \sum_{i=1}^{n}\delta_i\rho\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j - \widehat{g}_{j,i}}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i),$$

where the estimators $\widehat{g}_{j,i}$ are obtained in **Step A1** as the solutions of (11). Note that, as in **Step A2**, the estimators $\widehat{g}_{j,i}$ do not need to be computed again since we have already calculated them in **Step A1**. A common choice is $\upsilon \equiv 1$, since ρ is bounded.

Let $\widehat{\boldsymbol{\beta}}$ be the value minimizing L over the grid, i.e., $\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{1\le j\le J}L(\boldsymbol{\beta}_j)$.

- **Step A4**. The estimator of the nonparametric component is the solution $\widehat{g}_n(t) = \widehat{g}_{\widehat{\boldsymbol{\beta}}}(t)$ of

$$\sum_{i=1}^{n} w_i(t)\psi_1\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \widehat{g}_n(t)}{\widehat{s}(t)}\right) = 0.$$

When $p > 2$, a subsampling scheme may be considered as is usual in robust linear regression when computing, for instance, S-estimators. The algorithmic complexity involved in this setting is mainly due to the fact that we cannot ensure that a reweighted procedure will decrease the objective function, as it does in linear regression, since we follow now a robust profile-likelihood approach. To be more precise, in order to find an approximate solution to (7), we can compute the objective function over a "large" finite set of candidate solutions, and replace the minimization over $\boldsymbol{\beta} \in \mathbb{R}^p$ in **Step 2** by minimizing $L(\boldsymbol{\beta})$ over that finite set. The main problem is how to obtain a set of candidate solutions and a possible way is to adapt the well-known subsampling methods used in linear regression models. Let us assume that initial estimates of the regression function g_0 are available, denoted by $\widetilde{g}_{\mathrm{INI}}$. Among the complete data, i.e., those with $\delta_i = 1$, one may take subsamples of size p, $\{(y_{i_j}, \mathbf{x}_{i_j}, t_{i_j}) : 1 \le j \le p\}$, where $i_1 < \cdots < i_p$ and $\mathcal{I} = \{i_1, \ldots, i_p\} \in \{i : \delta_i = 1\}$. Clearly, more than p observed responses are needed. For each $\mathcal{I}$ find $\boldsymbol{\beta}_{\mathcal{I}}$ that satisfies the exact fit on the adjusted responses $y_{i_j} - \widetilde{g}_{\mathrm{INI}}(t_{i_j}) = x_{i_j}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{I}}$ for $i \in \mathcal{I}$. If a subsample is collinear, it is replaced by another. Then, the problem of minimizing $L(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathbb{R}^p$ is replaced by the finite problem of minimizing $L(\boldsymbol{\beta}_{\mathcal{I}})$ over $\mathcal{I}$. Since choosing all $\binom{m}{p}$ subsamples where $m = \#\{i : \delta_i = 1\}$ would be prohibitive unless both m and p are rather small, we choose N of them at random.

To improve the computational time, the initial estimator $\widetilde{g}_{\mathrm{INI}}$ of the regression function g_0 can be computed as $(\widetilde{\boldsymbol{\beta}}(t), \widetilde{g}_{\mathrm{INI}}(t)) = \mathrm{argmin}_{(\boldsymbol{\beta},a)}\sum_{i=1}^{n}\delta_i K((t - t_i)/h)\,\rho\left((y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - a)/\widehat{\sigma}\right)\upsilon(\mathbf{x}_i)$. Note that these estimators correspond to defining weighted regression estimators over the complete data set, where the weights are higher if the observation $(y_i, \mathbf{x}_i, t_i)$ is such that t_i is close to t. It is worth noticing that in order to compute these preliminary estimators the uniform kernel $K(u) = I_{[-1,1]}(u)/2$ can be considered so that any routine for computing a high breakdown point estimator can be used locally, i.e., for the observations lying in the neighborhood of t. Moreover, in order to guarantee the computation of the initial estimators, one needs to ensure that a large enough number of observations are available in each neighborhood of t_i. Besides, arguments similar to those considered in Appendix A.2 allow one to show that the functionals related to $(\widetilde{\boldsymbol{\beta}}(t), \widetilde{g}_{\mathrm{INI}}(t))$ are Fisher-consistent if condition (c) stated therein holds. Note that this preliminary estimator of $\boldsymbol{\beta}$ depends on t, since it is computed locally and so, it may have a nonparametric rate of convergence.

3.2. Computation of the robust marginal location estimators

To compute $\widehat{\theta}_{\mathrm{WS}}$ and $\widehat{\theta}_{\mathrm{A}}$, any standard algorithm for computing M-estimators can be used. For instance, they can be computed iteratively using reweighting, as described in the location setting in Chapter 2 of Maronna et al. (2006). On

the other hand, the following algorithm can be used to compute $\widehat{\theta}_{\mathrm{WI}}$. Using that $\widehat{\theta}_{\mathrm{WI}}$ is the solution of (9) and defining $W_2(u) = \psi_2(u)/u$, $p_i = p_n(\mathbf{x}_i, t_i)$ and $\widehat{m}_i = \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i)$, we get that

$$\sum_{i=1}^{n} \frac{\delta_i}{p_i}\psi_2\left(\frac{y_i - \widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right) + \sum_{i=1}^{n}\left(1 - \frac{\delta_i}{p_i}\right)\psi_2\left(\frac{\widehat{m}_i - \widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right) = 0$$

and so

$$\widehat{\theta}_{\mathrm{WI}} = \frac{\sum_{i=1}^{n}\left[\frac{\delta_i}{p_i}W_2\left(\frac{y_i-\widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right)y_i + \left(1-\frac{\delta_i}{p_i}\right)W_2\left(\frac{\widehat{m}_i-\widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right)\widehat{m}_i\right]}{\sum_{i=1}^{n}\left[\frac{\delta_i}{p_i}W_2\left(\frac{y_i-\widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right) + \left(1-\frac{\delta_i}{p_i}\right)W_2\left(\frac{\widehat{m}_i-\widehat{\theta}_{\mathrm{WI}}}{\widehat{\sigma}_{\mathrm{WI}}}\right)\right]}.$$

Let $\theta^{(0)} = \widehat{\theta}_{\mathrm{A}}$ and $\widehat{\sigma}_{\mathrm{A}} = \mathrm{mad}_{1\le i\le n}(\widehat{m}_i)$ The algorithm can be defined as follows:

- For $k = 0, 1, \ldots,$ given $\theta^{(k)}$ define

$$\theta^{(k+1)} = \frac{\sum_{i=1}^{n}\left[\frac{\delta_i}{p_i}W_2\left(\frac{y_i-\theta^{(k)}}{\widehat{\sigma}_{\mathrm{WI}}}\right)y_i + \left(1-\frac{\delta_i}{p_i}\right)W_2\left(\frac{\widehat{m}_i-\theta^{(k)}}{\widehat{\sigma}_{\mathrm{WI}}}\right)\widehat{m}_i\right]}{\sum_{i=1}^{n}\left[\frac{\delta_i}{p_i}W_2\left(\frac{y_i-\theta^{(k)}}{\widehat{\sigma}_{\mathrm{WI}}}\right) + \left(1-\frac{\delta_i}{p_i}\right)W_2\left(\frac{\widehat{m}_i-\theta^{(k)}}{\widehat{\sigma}_{\mathrm{WI}}}\right)\right]}.$$

- Iterate until convergence occurs or for a fixed number of steps $k_{\max}$.

4. Main results

In this section, we will derive the strong consistency of the marginal location M-estimators under the following conditions:

A1 $\psi_2 : \mathbb{R} \to \mathbb{R}$ is a bounded, differentiable function with bounded derivative ψ_2', such that $\int |\psi_2'(u)|\mathrm{d}u < \infty$.
A2 $\inf_{(\mathbf{x},t)} p(\mathbf{x}, t) = A > 0$.
A3 $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x}, t) - p(\mathbf{x}, t)| \xrightarrow{a.s.} 0$.

Assumption **A1** is a standard condition on the score function ψ_2, while **A2** states that response variables are observed, which is a common assumption in the literature.

Theorem 4.1. *Assume that* **A2** *and* **A3** *hold. Then:*
(a) $\|\widehat{F}_n - F\|_\infty = \sup_y |\widehat{F}_n(y) - F(y)| \xrightarrow{a.s.} 0$.
(b) *If in addition* $\widehat{\sigma} \xrightarrow{a.s.} \sigma_0$, **A1** *holds and in a neighborhood of* θ, *the function* $\lambda(a, \sigma_0)$ *has a unique change of sign, there exists a solution* $\widehat{\theta}_{\mathrm{ws}}$ *of* $\widehat{\lambda}_n(a, \widehat{\sigma}) = 0$, *such that* $\widehat{\theta}_{\mathrm{ws}} \xrightarrow{a.s.} \theta$.

Theorem 4.2. *Denote by* $\Pi(Q, P)$ *the Prohorov distance between the probability measures* Q *and* P. *Let* $\widetilde{m}(\mathbf{x}, t)$ *be an estimator of* $m(\mathbf{x}, t)$ *such that for any compact sets* $\mathcal{K}_1 \in \mathbb{R}^p$ *and* $\mathcal{K}_2 \in \mathbb{R}$,

$$\sup_{\substack{\mathbf{x}\in\mathcal{K}_1\\ t\in\mathcal{K}_2}} |\widetilde{m}(\mathbf{x}, t) - m(\mathbf{x}, t)| \xrightarrow{a.s.} 0.$$

Then:
(a) $\Pi(\widetilde{P}_n, P_Z)| \xrightarrow{a.s.} 0$ *where* P_Z *is the probability measure induced by* $Z = m(\mathbf{X}, T)$ *and*

$$\widetilde{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(\widehat{m}(\mathbf{x}_i, t_i)) = \frac{1}{n}\sum_{i=1}^{n} I_A(\widehat{m}_i).$$

(b) *If in addition* $\widehat{\sigma}$ *is an estimator of the scale* σ_Z *of* Z *such that* $\widehat{\sigma} \xrightarrow{a.s.} \sigma_Z$, ***A1*** *holds and in a neighborhood of* θ, *the function* $\lambda(a, \sigma_Z)$ *has a unique change of sign, there exists a solution* $\widetilde{\theta}$ *of* $\widetilde{\widetilde{\lambda}}_n(a, \widehat{\sigma}) = 0$, *such that* $\widetilde{\theta} \xrightarrow{a.s.} \theta$ *where*

$$\widetilde{\widetilde{\lambda}}_n(a, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\psi_2\left(\frac{\widetilde{m}_i - a}{\sigma}\right) = \int \psi_2\left(\frac{z-a}{\sigma}\right)\mathrm{d}\widetilde{\widetilde{F}}_n(z)$$

$$\widetilde{m}_i = \widetilde{m}(\mathbf{x}_i, t_i)$$

$$\widetilde{\widetilde{F}}_n(z) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,z]}(\widetilde{m}_i).$$

Note that Theorem 4.2 entails the following result.

Corollary 4.1. *Assume that:*

(i) $\widehat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}_0$.

(ii) *For any compact set* $\mathcal{K}$, $\sup_{t\in\mathcal{K}} |\widehat{g}_n(t) - g_0(t)| \xrightarrow{a.s.} 0$.

If in addition $\widehat{\sigma}_{\text{A}} \xrightarrow{a.s.} \sigma_Z$, ***A1*** *holds and in a neighborhood of* θ, *the function* $\lambda(a, \sigma_Z)$ *has a unique change of sign, there exists a solution* $\widehat{\theta}_{\text{A}}$ *of* $\widetilde{\lambda}_n(a, \widehat{\sigma}_{\text{A}}) = 0$, *such that* $\widehat{\theta}_{\text{A}} \xrightarrow{a.s.} \theta$.

Theorem 4.3. *Assume that* ***A2*** *and* ***A3*** *hold. Then:*

(a) $\|\widehat{\widehat{F}}_n - F\|_\infty = \sup_y |\widehat{\widehat{F}}_n(y) - F(y)| \xrightarrow{a.s.} 0$.

(b) *If in addition* $\widehat{\sigma}_{\text{WI}} \xrightarrow{a.s.} \sigma_0$, ***A1*** *holds and in a neighborhood of* θ, *the function* $\lambda(a, \sigma_0)$ *has a unique change of sign, there exists a solution* $\widehat{\theta}_{\text{WI}}$ *of* $\widehat{\widehat{\lambda}}_n(a, \widehat{\sigma}_{\text{WI}}) = 0$, *such that* $\widehat{\theta}_{\text{WI}} \xrightarrow{a.s.} \theta$.

4.1. Some comments

It is worth noticing that Theorem 4.3 entails that $\widehat{\theta}_{\text{WI}} \xrightarrow{a.s.} \theta$ even when the estimators of the regression function will not be consistent when we estimate consistently the probability of missing data. Obviously, the same happens with $\widehat{\theta}_{\text{WS}}$ that uses the observations at hand. Besides, $\widehat{\theta}_{\text{A}}$ is consistent if the regression model is correct without any need of estimating the probability of missing data.

One could try to combine both proposals in order to get the double-protected property in the sense of Scharfstein et al. (1999). Let

$$\widehat{\theta} = \widehat{\theta}_{\text{A}} + \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} \left(\widehat{\theta}_{\text{WS}} - \widehat{\theta}_{\text{WA}}\right)$$

with $\widehat{\theta}_{\text{WA}}$ the solution of

$$\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} \psi_2\left(\frac{\mathbf{x}_i^{\text{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i) - \widehat{\theta}_{\text{WA}}}{\widehat{\sigma}_{\text{WA}}}\right) = 0,$$

where $\widehat{\sigma}_{\text{WA}}$ is a preliminary robust scale estimator. It is clear that, if **A3** holds, $\widehat{\theta} \xrightarrow{a.s.} \theta$ since:

- $\widehat{\theta}_{\text{WS}} \xrightarrow{a.s.} \theta$,
- $\widehat{\theta}_{\text{WA}} \xrightarrow{a.s.} \theta_{\text{WA}}(F)$,
- $\widehat{\theta}_{\text{A}} \xrightarrow{a.s.} \theta_{\text{A}}(F)$,
- $\theta_{\text{A}}(F) = \theta_{\text{WA}}(F)$,

where $\theta_{\text{A}}(F)$ and $\theta_{\text{WA}}(F)$ stand for the functionals related to each proposal (see Appendix A.3).

However, if $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x}, t) - p^\star(\mathbf{x}, t)| \xrightarrow{a.s.} 0$ with $p^\star(\mathbf{x}, t) \neq p(\mathbf{x}, t)$ but $m(\mathbf{x}, t) = \mathbf{x}^{\text{T}}\boldsymbol{\beta}_0 + g_0(t)$, from the consistency of $\widehat{\theta}_{\text{A}}$, we get that $\widehat{\theta} \xrightarrow{a.s.} \theta + E\left(p(\mathbf{X}, T)/p^\star(\mathbf{X}, T)\right)\left(\theta_{\text{WS}}(F) - \theta_{\text{WA}}(F)\right)$ and we cannot ensure that the two location functionals, $\theta_{\text{WS}}(F)$ and $\theta_{\text{WA}}(F)$, will be equal. However, note that, when ψ_2 is the identity function, this equality holds due to the linearity of the expectation.

Moreover, assume that there exists an M-functional $\theta(F)$ such that if $\widehat{\theta}$ is the corresponding estimator then $\widehat{\theta}$ satisfies the double-protected property. Hence, assuming that scale σ is known, we must have:

(a) If $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x}, t) - p(\mathbf{x}, t)| \xrightarrow{a.s.} 0$, then $\theta(F)$ should be equal to $\theta_{\text{WS}}(F)$, i.e., it should satisfy

$$E\frac{\delta}{p(\mathbf{X}, T)} \psi_2\left(\frac{Y - \theta(F)}{\sigma}\right) = 0.$$

(b) If $\sup_{(\mathbf{x},t)} |\widehat{m}_n(\mathbf{x}, t) - m(\mathbf{x}, t)| \xrightarrow{a.s.} 0$, then $\theta(F)$ should be equal to $\theta_{\text{A}}(F)$, i.e., it should satisfy

$$E\psi_2\left(\frac{m(\mathbf{X}, T) - \theta(F)}{\sigma}\right) = 0.$$

Thus, if one wants to obtain a robust and double-protected M-estimator, both equations should be satisfied. Clearly, when $\psi_2 \equiv \text{id}$, this is fulfilled when the errors ϵ are independent of $(\mathbf{x}, t)$ and have symmetric distribution. So, if $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x}, t) - p^\star(\mathbf{x}, t)| \xrightarrow{a.s.} 0$ and $\sup_{(\mathbf{x},t)} |\widehat{m}(\mathbf{x}, t) - m^\star(\mathbf{x}, t)| \xrightarrow{a.s.} 0$, we need that:

(a) if $p^\star(\mathbf{x}, t) = p(\mathbf{x}, t)$, then $\theta(F)$ will satisfy

$$E\frac{\delta}{p^\star(\mathbf{X}, T)}\psi_2\left(\frac{Y-\theta(F)}{\sigma}\right) = E\frac{p^\star(\mathbf{X}, T)}{p^\star(\mathbf{X}, T)}\psi_2\left(\frac{m(\mathbf{X}, T)+\sigma_0\epsilon-\theta(F)}{\sigma}\right) = 0,$$

or equivalently in this situation

$$E\psi_2\left(\frac{m(\mathbf{X}, T)+\sigma_0\epsilon-\theta(F)}{\sigma}\right) = 0.$$

b) On the other hand, if $m^\star(\mathbf{x}, t) = m(\mathbf{x}, t)$, $\theta(F)$ should satisfy

$$E\psi_2\left(\frac{m(\mathbf{X}, T)-\theta(F)}{\sigma}\right) = 0.$$

For the regular score functions used in robustness this seems difficult to attain due to the non-linearity of ψ_2. Another possibility could be to consider the solution $\widehat{\theta}$ of $\lambda_n^\star(\widehat{\theta}, \widehat{\sigma}_{\text{WA}}) = 0$ with

$$\lambda_n^\star(a, \sigma) = \widetilde{\lambda}_n(a, \sigma) + \left(\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)}\right)\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)}\left[\psi_2\left(\frac{y_i-a}{\sigma}\right) - \psi_2\left(\frac{\widehat{m}_i-a}{\sigma}\right)\right] = 0,$$

where the $\widehat{m}_i$ are the values predicted using the partially linear model (1), $\widehat{m}_i = \widehat{m}(\mathbf{x}_i, t_i) = \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i)$ and $\widetilde{\lambda}_n(a, \sigma)$ was defined in Section 2.2.1 as

$$\widetilde{\lambda}_n(a, \sigma) = \frac{1}{n}\sum_{i=1}^n \psi_2\left(\frac{\widehat{m}_i-a}{\sigma}\right).$$

Note that if $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x}, t) - p^\star(\mathbf{x}, t)| \xrightarrow{a.s.} 0$ and $\sup_{(\mathbf{x},t)} |\widehat{m}(\mathbf{x}, t) - m^\star(\mathbf{x}, t)| \xrightarrow{a.s.} 0$, $\widehat{\theta}$ will be consistent to the solution $\theta^\star(F)$ of $\lambda^\star(a, \sigma) = 0$ with

$$\lambda^\star(a, \sigma) = \lambda_Z(a, \sigma) + \left(E\frac{p(\mathbf{X}, T)}{p^\star(\mathbf{X}, T)}\right)E\left\{\frac{p(\mathbf{X}, T)}{p^\star(\mathbf{X}, T)}\left[\psi_2\left(\frac{Y-a}{\sigma}\right) - \psi_2\left(\frac{Z-a}{\sigma}\right)\right]\right\},$$

where $\lambda_Z(a, \sigma) = E\psi_2((Z-a)/\sigma)$, $Z = m^\star(\mathbf{X}, T)$. Again:

(a) If $p^\star(\mathbf{x}, t) = p(\mathbf{x}, t)$, then $\lambda^\star(a, \sigma) = E\psi_2((Y-a)/\sigma)$ and so $\theta^\star(F) = \theta_{\text{ws}}(F)$, attaining the desired Fisher-consistency.
(b) If $m^\star(\mathbf{x}, t) = m(\mathbf{x}, t)$, then, if $R(\mathbf{X}, T) = p(\mathbf{X}, T)/p^\star(\mathbf{X}, T)$, we have

$$\lambda^\star(a, \sigma) = E\psi_2\left(\frac{m(\mathbf{X}, T)-a}{\sigma}\right) + (ER(\mathbf{X}, T))\,E\left(R(\mathbf{X}, T)\left[\psi_2\left(\frac{m(\mathbf{X}, T)+\sigma_0\epsilon-a}{\sigma}\right) - \psi_2\left(\frac{m(\mathbf{X}, T)-a}{\sigma}\right)\right]\right).$$

Thus, using that $m(\mathbf{X}, T)$ has a symmetric distribution around θ, we obtain

$$\lambda^\star(\theta, \sigma) = (ER(\mathbf{X}, T))\,E\left(R(\mathbf{X}, T)\left[\psi_2\left(\frac{m(\mathbf{X}, T)+\sigma_0\epsilon-\theta}{\sigma}\right) - \psi_2\left(\frac{m(\mathbf{X}, T)-\theta}{\sigma}\right)\right]\right).$$

So, if we want that $\theta^\star(F) = \theta_{\text{A}}(F) = \theta$ we need that

$$E\left(R(\mathbf{X}, T)\left[\psi_2\left(\frac{m(\mathbf{X}, T)+\sigma_0\epsilon-\theta}{\sigma}\right) - \psi_2\left(\frac{m(\mathbf{X}, T)-\theta}{\sigma}\right)\right]\right) = 0. \tag{13}$$

Eq. (13) will be fulfilled for instance if the ratio $R(\mathbf{X}, T)$ is an even function of $m(\mathbf{X}, T) - \theta$. This assumption seems unnatural and that is why these estimators were not considered in this paper. Again, if ψ_2 is the identity function, (13) is automatically fulfilled.

Analogous conclusions can be obtained if we define, as in Wang and Sun (2007), the M-location estimator of the weighted responses $(\delta_i/p_n(\mathbf{x}_i, t_i))\,y_i + (1 - \delta_i/p_n(\mathbf{x}_i, t_i))\left(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i)\right)$.

5. Monte Carlo study

A simulation study was carried out for when the regression parameter has dimension 1. The S-code is available upon request to the authors.

In all tables, WSE_{LS}, AE_{LS} and WIE_{LS} denote the classical estimates obtained using the weighted simplified estimate, the averaged estimate and the weighted imputed estimate, respectively. On the other hand, the corresponding robust estimates are denoted as WSE_{R}, AE_{R} and WIE_{R}.

The aims of this study are to compare:

- the behavior of the classical and robust estimators under contamination and for normal samples, for different missing data probabilities;

- the performances of the robust proposals, the weighted simplified, the averaged and the weighted imputed estimators, among themselves and also with that of the robust estimator that would be computed if the complete data set were available (WSE_R); note that this estimator, which corresponds to $p(\mathbf{x}, t) \equiv 1$, cannot be computed in practice, and the goal is to detect which of the proposals would give mean square errors closer to those that would be obtained if there were no missing responses.

In both the classical and robust smoothing procedures, we have used the gaussian kernel with standard deviation 0.37 such that the interquartile range is 0.5. The robust smoothing procedure used local M-estimates with score function ψ_1, the bisquare function with tuning constant 4.685, using local medians as initial estimates. The chosen tuning constant for the local M-estimator gives a 95% efficiency with respect to its linear relative. The same score function was used to compute the marginal estimators, that is, we choose $\psi_2 = \psi_1$.

The robust estimator of the regression parameter $\boldsymbol{\beta}$ was computed as described in Section 3 using as *rho*-functions the bisquare function, that is,

$$\rho_0(x) = \rho_1\left(\frac{x}{c_0}\right) \quad \text{and} \quad \rho(x) = \rho_1\left(\frac{x}{c_1}\right) \tag{14}$$

with $c_0 = 1.56$, $c_1 \geq c_0$ and $\rho_1(x) = \min(1, 1-(1-x^2)^3)$. The value selected for c_0 ensures Fisher-consistency of the scale when the errors are gaussian, while $c_1 = 4.68$ guarantees that under a regression model the *MM*-estimates will achieve 95% efficiency.

Two different models have been considered for studying the behavior of the proposed methods. For both of them, due to the expensive computing time when evaluating the robust estimators, we only performed 500 replications generating independent samples of size $n = 100$.

5.1. Model I

Under this model, we first generate observations as

$$z_i = \boldsymbol{\beta}_0 \mathbf{x}_i + 2\sin(4\pi(t_i - 0.5)) + \sigma_0 \epsilon_i \quad 1 \leq i \leq n, \tag{15}$$

where $\boldsymbol{\beta}_0 = 2$, $\left(\mathbf{x}_i^{\mathrm{T}}, t_i\right)^{\mathrm{T}} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (0, \frac{1}{2})^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1/\left(6\sqrt{3}\right) \\ 1/\left(6\sqrt{3}\right) & 1/36 \end{pmatrix}$, and $\epsilon_i \sim N(0, 1)$ with $\sigma_0^2 = 0.25$ in the non-contaminated case.

The results for normal data sets will be indicated by C_0 in Tables 1–7, while C_1 to C_4 will denote the following contaminations:

- C_1: $\epsilon_1, \ldots, \epsilon_n$, are i.i.d., $0.9N(0, 1) + 0.1N(0, 25)$. This contamination corresponds to inflating the errors and thus, will only affect the variance of the location estimates.
- C_2: $\epsilon_1, \ldots, \epsilon_n$, are i.i.d., $0.9N(0, 1) + 0.1N(0, 25)$ and artificially ten observations of the response z_i but not of the carriers $\mathbf{x}_i$ were modified to be equal to 20 at equally spaced values of t. This case corresponds to introducing outliers with high residuals. The aim of this contamination is to study changes in bias in the estimation of the location parameter.
- C_3: $\epsilon_1, \ldots, \epsilon_n$, are i.i.d., $0.9N(0, 1) + 0.1N(0, 25)$ and artificially ten observations of the carriers $\mathbf{x}_i$ but not of the response z_i were modified to be equal to 20 at equally spaced values of t. This case corresponds to introducing high leverage points. The aim of this contamination is to study changes in bias in the estimation of the location parameter when using the averaged and the weighted imputed estimates, since this contamination affects mainly the estimation of the regression parameter.
- C_4: $\epsilon_1, \ldots, \epsilon_n$, are i.i.d., $0.9N(0, 1) + 0.1N(0, 25)$ and artificially five observations of the carriers $\mathbf{x}_i$ and five of the response z_i were modified to be equal to 20 and -20, respectively, at equally spaced values of t. The modified observations of the response were not allocated at the same t as those of the carriers. This case corresponds to introducing both high leverage points and high residuals. The aim of this contamination is to study changes in bias in the different estimators of the location parameter since this contamination affects the regression parameter and also the marginal one.

In a first step, when computing the marginal estimators, the missing data probabilities are not estimated but assumed to be known, i.e., we have taken $p_n(x, t) = p(x, t)$. This decision was taken in order to avoid increasing biases due to the estimation of the missing data probability (see, for instance, Chen et al. (2006)). Moreover, as discussed in Wang et al. (1998), it seems natural to argue that the weighted estimators using the estimated probabilities are at least as efficient as those using the true model and this phenomenon will also be studied below.

The missing data models considered can be described as follows. Let $\pi(u) = 0.4 + 0.5(\cos(2u + 0.4))^2$. We define $y_i = z_i$ if $\delta_i = 1$, and as missing otherwise, where the δ_i are generated according to the following missingness mechanism (see (3)):

(a) $p(x, t) \equiv 1$ that corresponds to the complete data situation. As in nonparametric regression with missing responses, in this case, the WSE_{LS} and WIE_{LS} give the same results. However, although WSE_R and WIE_R also should be identical, they provide slightly different results since the algorithms used to compute them were not identical. To compute WSE_R we have used the S-plus routine *location.m* with 20 iterations, while to compute the WIE_R, we used the reweighted method described in Section 3.2 with $k_{\max} = 10$, i.e., only ten iterations were performed.

Table 1
Minimum, mean value and maximum of the least squares cross-validation error, under C_0 and C_2, for **Model I**.

$p(\mathbf{x}, t)$	0.8		$\pi(t)$		$\pi(x)$		$\pi(xt)$		
	AE_{LS}	WIE_{LS}	AE_{LS}	WIE_{LS}	AE_{LS}	WIE_{LS}	AE_{LS}	WIE_{LS}	
Minimum	5.6703	5.6702	6.7913	6.7911	6.0852	6.0846	5.7455	5.7442	
Mean	5.6713	5.6709	6.7972	6.8010	6.0880	6.0863	5.7496	5.7491	C_0
Maximum	5.6717	5.6713	6.7992	6.8056	6.0891	6.0871	5.7516	5.7512	
Minimum	34.7509	34.7516	40.9682	40.9641	35.0603	35.1356	36.3458	36.3825	
Mean	34.7542	34.7540	41.0222	41.0219	35.0668	35.1434	36.3476	36.3922	C_2
Maximum	34.7554	34.7550	41.5475	41.5549	35.0831	35.1623	36.3497	36.4062	

Table 2
Biases, standard deviations and mean square errors of the classical and robust procedures, under C_0 to C_4, when $h = 0.2$ and $p(\mathbf{x}, t) = 1$, for **Model I**.

	WSE_{LS}	AE_{LS}	WIE_{LS}	WSE_R	AE_R	WIE_R	
bias	−0.0072	−0.0077	−0.0072	−0.0135	−0.0137	−0.0132	
sd	0.2486	0.2515	0.2486	0.2612	0.2653	0.2608	C_0
MSE	0.0618	0.0633	0.0618	0.0684	0.0706	0.0682	
bias	−0.0081	−0.0091	−0.0081	−0.0171	−0.0161	−0.0166	
sd	0.2636	0.2666	0.2636	0.2767	0.2702	0.2762	C_1
MSE	0.0696	0.0712	0.0696	0.0769	0.0733	0.0766	
bias	1.9922	2.0318	1.9922	−0.01287	−0.0078	−0.0126	
sd	0.2517	0.2574	0.2517	0.2888	0.2703	0.2884	C_2
MSE	4.0322	4.1946	4.0322	0.0836	0.0731	0.0833	
bias	0.0017	0.0010	0.0017	−0.0058	−0.0119	−0.0056	
sd	0.2601	0.2621	0.2601	0.2730	0.2831	0.2724	C_3
MSE	0.0676	0.0687	0.0676	0.0746	0.0803	0.0742	
bias	−1.0070	−1.0502	−1.0070	−0.0145	−0.0081	−0.0141	
sd	0.2563	0.2551	0.2563	0.2803	0.2763	0.2800	C_4
MSE	1.0796	1.1681	1.0796	0.0788	0.0764	0.0786	

(b) $p(x, t) \equiv 0.8$, corresponding to data missing completely at random.
(c) $p(x, t) = \pi(t)$.
(d) $p(x, t) = \pi(x)$.
(e) $p(x, t) = \pi(xt)$.

The classical least squares cross-validation method constructs an asymptotically optimal data-driven bandwidth and thus, adaptive data-driven estimators, by minimizing

$$CV_{LS}(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left(y_i - \widehat{\theta}^{(i)}(h)\right)^2 w^2(t_i).$$

In our study, the function w was chosen equal to 1. In order to study the sensitivity of the resulting estimator to the bandwidth, Table 1 shows the minimum, mean and maximum values of CV_{LS} as a function of h for the missing data models (b) to (e), for the average and weighted average estimators and for one of the samples generated as above. We have generated the sample according to C_0 and C_2 to show the sensitivity of the least squares procedure and we have computed the cross-validation errors for a grid of 20 equally spaced values of h between 0.05 and 1. As could be seen, when considering both the classical or the robust procedures, for non-contaminated samples, the cross-validation error of all the estimators is almost constant on its domain, showing the lack of sensitivity of the marginal estimators to the smoothing parameter. It is also clear that the least squares cross-validation error is highly sensitive to anomalous data, since its values are almost 7 times those obtained with the non-contaminated samples. The robust procedure has a behavior similar to that described for the least squares method, under C_0. For instance, when considering $p(x, t) = 0.4 + 0.5(\cos(2t + 0.4))^2$, the minimum ($m$) and maximum values ($M$) of the robust cross-validation function related to AE_R and WIE_R are $m_{AE_R} = 7.3511$, $M_{AE_R} = 7.3953$ and $m_{WIE_R} = 7.3601$ and $M_{WIE_R} = 7.3943$, respectively. The robust cross-validation $RCV_R(h)$, defined in (10), is much more stable under contamination. Effectively, under C_2, the minimum (m) and maximum values (M) of the robust cross-validation function are $m_{AE_R} = 12.0666$, $M_{AE_R} = 12.1572$ and $m_{WIE_R} = 11.9976$ and $M_{WIE_R} = 12.0888$ and the shape of the function is almost the same as in the non-contaminated situation.

Considering the above discussion, a robust cross-validation procedure was not performed in this preliminary study, taking into account that it is very expensive computationally when it is combined with the robust profile procedure and since it was clear from the results obtained that, in all situations, the bandwidth choice did not seem crucial for the estimation of

Table 3
Biases, standard deviations and mean square errors of the classical and robust procedures, under C_0 to C_4, when $h = 0.2$ and $p(\mathbf{x}, t) = 0.80$, for **Model I**.

	WSE_{LS}	AE_{LS}	WIE_{LS}	WSE_R	AE_R	WIE_R	
bias	−0.0032	−0.0062	−0.0054	−0.0103	−0.0126	−0.0124	
sd	0.2786	0.2511	0.2481	0.2957	0.2646	0.2605	C_0
MSE	0.0776	0.0631	0.0616	0.0875	0.0701	0.0680	
bias	−0.0036	−0.0071	−0.0059	−0.0136	−0.0166	−0.0161	
sd	0.2981	0.2704	0.2673	0.3152	0.2702	0.2793	C_1
MSE	0.0889	0.0731	0.0715	0.0996	0.0733	0.0783	
bias	1.9818	2.0232	1.9846	−0.0102	−0.00756	−0.0145	
sd	0.4017	0.4016	0.3954	0.3280	0.2712	0.2933	C_2
MSE	4.0889	4.2547	4.0952	0.1077	0.0736	0.0862	
bias	0.0066	0.0101	0.0101	−0.0021	−0.0118	−0.0019	
sd	0.2945	0.2840	0.2819	0.3110	0.2844	0.2781	C_3
MSE	0.0868	0.0807	0.0796	0.0967	0.0811	0.0773	
bias	−0.9914	−1.0299	−0.9888	−0.0112	−0.0079	−0.0121	
sd	0.3764	0.3776	0.3716	0.3200	0.2773	0.2891	C_4
MSE	1.1246	1.2033	1.1158	0.1025	0.0769	0.0837	

Table 4
Biases, standard deviations and mean square errors of the classical and robust procedures, under C_0 to C_4, when $h = 0.2$ and $p(\mathbf{x}, t) = 0.4 + 0.5\cos^2(2t + 0.4)$, for **Model I**.

	WSE_{LS}	AE_{LS}	WIE_{LS}	WSE_R	AE_R	WIE_R	
bias	−0.0016	−0.0227	−0.0102	−0.0104	−0.0313	−0.0183	
sd	0.3750	0.2618	0.2579	0.4009	0.2748	0.2703	C_0
MSE	0.1406	0.0691	0.0666	0.1608	0.0765	0.0734	
bias	−0.0054	−0.0285	−0.0159	−0.0145	−0.0369	−0.0250	
sd	0.4018	0.2933	0.2899	0.4269	0.2850	0.3006	C_1
MSE	0.1614	0.0868	0.0843	0.1821	0.0826	0.0910	
bias	1.9820	1.9777	1.9753	−0.0096	−0.0288	−0.0260	
sd	0.8037	0.8111	0.8079	0.4435	0.2889	0.3223	C_2
MSE	4.5747	4.5691	4.5547	0.1968	0.0843	0.1046	
bias	0.0046	0.0022	0.0271	−0.0035	−0.0325	−0.0092	
sd	0.4008	0.4080	0.4041	0.4254	0.3018	0.3130	C_3
MSE	0.1607	0.1665	0.1640	0.1810	0.0921	0.0981	
bias	−0.9985	−0.9497	−0.9112	−0.0162	−0.0291	−0.0258	
sd	0.6778	0.8168	0.8146	0.4340	0.2952	0.3247	C_4
MSE	1.4564	1.5692	1.4937	0.1887	0.0880	0.1061	

the marginal location parameter. Even when we have performed the simulation with bandwidths $h = 0.1$, 0.2 and 0.4, we only present in this paper the results for $h = 0.2$. In fact, all the bandwidths considered lead to the same conclusions.

The performance of the location estimators was measured using the bias, standard deviation and mean square error (bias, sd and MSE, respectively) in Tables 2–6. Boxplots can be found in Bianco et al. (2009).

The results reported in Tables 2–6 show that when there is no contamination, the linear estimators perform better than the robust ones (both in bias and mean square error) showing the loss of efficiency related to the score function used to compute the location M-estimators. On the other hand, the robust estimators show their advantage over the classical ones when outliers with high residuals are present having a similar performance when contaminating the errors. In fact, the MSE of the classical estimators is more than 20 times larger than the one observed under no contamination and also than the MSE of the robust estimators which are much more stable under C_2 or C_4. Note that this is mainly due to the increased bias of the classical estimators using any of the methods (weighted simplified, averaged or weighted imputed one). This explains the better efficiency of the robust estimators under C_2 and C_4. It is worth noticing that the classical estimators seem stable with respect to contaminating only the carriers with high leverage points C_3. This is natural when using the weighted simplified estimator since responses with large residuals were not included in this contamination, but it could seem unnatural when using the average or weighted imputed estimators, since it could be expected that large values of the covariates $\mathbf{x}$ would lead the classical estimate to explode. However, the good performance observed is mainly due to the fact that the least squares regression parameter estimates β almost as 0, in all the missingness schemes. Effectively, under C_3 the median of $\widehat{\boldsymbol{\beta}}_{LS}$ is close to 0 while $\widehat{\boldsymbol{\beta}}_R$ is close to 2 as it should be. The same behavior is observed also under C_4, where the classical estimates of $\boldsymbol{\beta}$ exhibit a large bias which decreases the influence of the leverage points when using the average or

Table 5
Biases, standard deviations and mean square errors of the classical and robust procedures, under C_0 to C_4, when $h = 0.2$ and $p(\mathbf{x}, t) = 0.4 + 0.5\cos^2(2x + 0.4)$, for **Model I**.

	WSE_{LS}	AE_{LS}	WIE_{LS}	WSE_R	AE_R	WIE_R	
bias	−0.0040	−0.0090	−0.0068	−0.0139	−0.0155	−0.0137	
sd	0.3336	0.2547	0.2539	0.3551	0.2684	0.2668	C_0
MSE	0.1113	0.0650	0.0645	0.1263	0.0723	0.0713	
bias	−0.0042	−0.0095	−0.0085	−0.0173	−0.0202	−0.0170	
sd	0.3545	0.2784	0.2789	0.3754	0.2769	0.2904	C_1
MSE	0.1257	0.0776	0.0778	0.1412	0.0771	0.0846	
bias	1.9946	2.0265	1.9953	−0.0093	−0.0114	−0.0125	
sd	0.6059	0.5389	0.5776	0.3835	0.2784	0.3055	C_2
MSE	4.3456	4.3970	4.3148	0.1471	0.0777	0.0935	
bias	0.0082	0.0011	0.0088	−0.0023	−0.0152	−0.0021	
sd	0.3423	0.3073	0.3096	0.3623	0.2921	0.2880	C_3
MSE	0.1172	0.0944	0.0959	0.1313	0.0856	0.0829	
bias	−0.9857	−1.0313	−0.9883	−0.0112	−0.0115	−0.0132	
sd	0.5026	0.4605	0.4896	0.3731	0.2846	0.2998	C_4
MSE	1.2241	1.2757	1.2165	0.1393	0.0811	0.0900	

Table 6
Biases, standard deviations and mean square errors of the classical and robust procedures, under C_0 to C_4, when $h = 0.2$ and $p(\mathbf{x}, t) = 0.4 + 0.5\cos^2(2xt + 0.4)$, for **Model I**.

	WSE_{LS}	AE_{LS}	WIE_{LS}	WSE_R	AE_R	WIE_R	
bias	−0.0124	−0.0203	−0.0086	−0.0308	−0.0283	−0.0158	
sd	0.3238	0.2535	0.2511	0.3481	0.2666	0.2637	C_0
MSE	0.1050	0.0647	0.0631	0.1221	0.0719	0.0698	
bias	−0.0131	−0.0229	−0.0098	−0.0347	−0.0323	−0.0189	
sd	0.3421	0.2725	0.2719	0.3647	0.2732	0.2834	C_1
MSE	0.1172	0.0748	0.0739	0.1342	0.0757	0.0807	
bias	1.9982	1.9659	1.9995	−0.0262	−0.0235	−0.0202	
sd	0.5411	0.4862	0.5206	0.3757	0.2744	0.2983	C_2
MSE	4.2855	4.1010	4.2690	0.1419	0.0759	0.0894	
bias	−0.0036	−0.1115	0.0021	−0.0215	−0.0269	−0.0062	
sd	0.3364	0.2917	0.3122	0.3583	0.2871	0.2857	C_3
MSE	0.1132	0.0975	0.0975	0.1288	0.0831	0.0817	
bias	−1.0132	−1.1413	−1.0053	−0.0309	−0.0233	−0.0179	
sd	0.4784	0.4513	0.4862	0.3697	0.2799	0.2946	C_4
MSE	1.2555	1.5063	1.2470	0.1376	0.0789	0.0871	

the weighted imputed estimators. In both cases, the classical regression parameter estimators are useless and in this sense, the least squares procedures seem not reliable for estimating the marginal parameter.

It is worth noticing that, when there is no contamination, except for the complete data estimators, the weighted imputed estimators (linear or robust) perform better than the two other competitors leading to smaller mean square errors. Their advantage over the weighted simplified estimator is especially reflected when the missing data probability depends on x, t or both variables. In this situation (see Tables 4–6), the weighted simplified estimators have almost twice the mean square errors of the weighted imputed ones. The worst situation for the weighted simplified procedure is when $p(x, t)$ only depends on t. This fact can be explained since it gives the larger proportion of missing data in each sample, near 70%, while in the two other situations the proportion of missing data is about 65%. The same conclusion holds under C_1. However, under C_2 to C_4 a different behavior is observed for the classical and robust estimators. As expected, the weighted imputed M-estimators perform much better than the weighted simplified method leading to almost the same ratios between the mean square error of the weighted simplified M-estimator and the weighted imputed M-estimator as in the non-contaminated situation. In contrast, when using the linear estimators, the mean square errors of the three estimators are almost the same due to the bias of all procedures. Note also that, under C_2 and C_4, when using the robust estimators the smallest mean square errors are attained by the averaged M-estimators, although they are almost of the same order as the weighted imputed M-estimator.

In order to evaluate the impact of the estimation of the missing data probabilities on the final marginal estimator, we have performed a moderate simulation study for some of the situations discussed above. We have considered the setting in which the missing data probability is given by $p(x, t) = \pi(t)$ since the results for the other missing data schemes lead to

Table 7
Biases, standard deviations and mean square errors of the classical and robust procedures, under C_0 to C_4, when $h = 0.2$ and $p(\mathbf{x}, t) = 0.4 + 0.5\cos^2(2t + 0.4)$ is estimated using a kernel estimator with bandwidth $\gamma = 0.2$, for **Model I**.

	WSE_{LS}	WIE_{LS}	WSE_R	WIE_R	
bias	−0.0206	−0.0109	−0.0309	−0.0189	
sd	0.3314	0.2569	0.3554	0.2694	C_0
MSE	0.1102	0.0661	0.1273	0.0729	
bias	−0.0259	−0.0173	−0.0364	−0.0261	
sd	0.3564	0.2895	0.3797	0.3013	C_1
MSE	0.1277	0.0841	0.1455	0.0914	
bias	1.9529	1.9682	−0.0306	−0.0266	
sd	0.8035	0.8171	0.4008	0.3225	C_2
MSE	4.4593	4.5414	0.1616	0.1047	
bias	−0.0138	0.0246	−0.0232	−0.0111	
sd	0.3600	0.4032	0.3840	0.3128	C_3
MSE	0.1298	0.1632	0.1480	0.0980	
bias	−1.0052	−0.9092	−0.0363	−0.0280	
sd	0.6623	0.8173	0.3908	0.3253	C_4
MSE	1.4491	1.4946	0.1541	0.1066	

similar conclusions. We have estimated $\pi(t)$ using a kernel density estimator with bandwidth $\gamma = 0.2$ and gaussian kernel K as described above. Table 7 reports summary measures for the marginal location estimators. It is worth noticing that the results corresponding to AE_{LS} and AE_R are not given since these estimators do not depend on estimators of the missing data probabilities. The results obtained for other choices for the smoothing parameter γ are similar to those reported here. The comparison between the behaviors of the robust and linear estimators in this case is similar to that observed when the missing data probability is assumed to be known. On the other hand, as expected, when estimating $p(x, t)$ the bias of the estimators increases; in particular, in most cases WSE_{LS} (WSE_R) has a larger bias than AE_{LS} (AE_R), in spite of what happens when $p(x, t)$ is known. However, mean square errors reported in Table 7 are smaller to those given in Table 4 due to a reduction in the standard deviation. This surprising fact has been observed by, for instance, Wang et al. (1997) in the classical setting.

5.2. Model II

As in Robins et al. (1994) in some situations a parametric model, $p_{\boldsymbol{\alpha}}(\mathbf{x}, t)$, $\boldsymbol{\alpha} \in \mathbb{R}^q$, for the missing data probabilities can be assumed. The goal of the following simulation study is to analyze the impact of estimating the unknown parameters $\boldsymbol{\alpha}$ on the final marginal estimators. Again the observations z_i satisfy model (15) where $\boldsymbol{\beta}_0 = 2$, but the covariates were generated as $\mathbf{x}_i \sim N(0, 1)$ and $t_i \sim \mathcal{U}(0, 1)$, while the errors are $\epsilon_i \sim N(0, 1)$ with $\sigma_0^2 = 0.25$ in the non-contaminated case. The results for normal data sets will be indicated by $C_{0,L}$ in the tables, while $C_{1,L}$ refers to the following contamination:

- $C_{1,L}$: $\epsilon_1, \ldots, \epsilon_n$, are i.i.d., $0.9N(0, 1) + 0.1N(0, 25)$ and artificially five observations of the carriers $\mathbf{x}_i$ and five of the response z_i were modified to be equal to 5 and −20, respectively, at equally spaced values of t. The modified observations of the response were not allocated at the same t as those of the carriers.

Note that $C_{1,L}$ is similar to C_4. The subscript L indicates the fact that the missing data scheme considered is a logistic model, where $p(x, t) = 1/(1 + \exp(-2x - 12(t - 0.5)))$. As above, we then define $y_i = z_i$ if $\delta_i = 1$, and as missing otherwise, to obtain the missing responses. This missing data model is analogous to that considered by Croux and Haesbroeck (2003) who studied the behavior of robust estimators under a logistic regression model. In Tables 8 and 9 we summarize the results obtained under Model II for non-contaminated samples and under $C_{1,L}$, when the true probability or the estimated one is used to compute the marginal estimators, respectively. As in Wang et al. (1997), the missing data probability was estimated using the parametric model and using a smoothed estimator. The nonparametric estimator was computed, using a product gaussian kernel, as defined in (8) with $b_n = 0.2$ and $\lambda_n = 0.4$. On the other hand, when computing the classical marginal location estimators, $\boldsymbol{\alpha}$ was estimated using the maximum likelihood method, while for the robust marginal procedures the robust estimators implemented by Croux and Haesbroeck (2003) were used. As in Section 5.1, the classical and robust estimators of θ perform similarly under $C_{0,L}$ while under $C_{1,L}$ the robust procedures show their advantage, either when the missing data probability is assumed to be known or when it is estimated. Moreover, as in Section 5.1, the standard deviations of the classical estimators are reduced when estimating parametrically the missing data probabilities or when using the Nadaraya–Watson estimator. However, this phenomenon is not observed for the robust estimators when estimating parametrically the missing data probabilities. It is worth noticing that when $\boldsymbol{\alpha}$ is estimated, biases are also reduced under $C_{0,L}$ for both the classical and robust methods. This fact was already pointed out by Wang et al. (1997) when the missing data probabilities depend on both covariates, for the estimators defined therein. Besides, when the missing data probabilities are estimated nonparametrically, biases are enlarged. In all cases, except when $\boldsymbol{\alpha}$ is estimated

Table 8
Biases, standard deviations and mean square errors of the classical and robust procedures, under $C_{0,\mathrm{L}}$ and $C_{1,\mathrm{L}}$, when $h = 0.2$ and the missing data probability is assumed to be known, for **Model II**.

	WSE_{LS}	AE_{LS}	WIE_{LS}	WSE_{R}	AE_{R}	WIE_{R}	
bias	0.2896	−0.0920	−0.0735	0.2717	−0.0843	−0.0702	
sd	0.5710	0.3089	0.3387	0.6350	0.2972	0.3300	$C_{0,\mathrm{L}}$
MSE	0.4099	0.1039	0.1201	0.4770	0.0954	0.1139	
bias	−0.8423	−0.7951	−0.9040	0.2669	−0.1896	−0.2088	
sd	1.1216	0.8471	1.2317	0.6709	0.4937	0.7005	$C_{1,\mathrm{L}}$
MSE	1.9674	1.3497	2.3344	0.5214	0.2797	0.5344	

Table 9
Biases, standard deviations and mean square errors of the classical and robust procedures, under $C_{0,\mathrm{L}}$ and $C_{1,\mathrm{L}}$, when $h = 0.2$ and the missing data probability is estimated parametrically or nonparametrically, for **Model II**.

	Parametric estimator of $p(x, t)$				Nadaraya–Watson estimator of $p(x, t)$				
	WSE_{LS}	WIE_{LS}	WSE_{R}	WIE_{R}	WSE_{LS}	WIE_{LS}	WSE_{R}	WIE_{R}	
bias	0.2708	−0.0838	0.2166	−0.0856	0.4283	−0.0906	0.4239	−0.0898	
sd	0.5291	0.3377	0.7649	0.4476	0.2911	0.3089	0.3054	0.2973	$C_{0,\mathrm{L}}$
MSE	0.3533	0.1211	0.6319	0.2077	0.2682	0.1036	0.2729	0.0964	
bias	−0.8799	−0.8581	0.2152	−0.2001	−0.7877	−0.8146	0.3864	−0.2897	
sd	1.1104	1.0871	0.7853	0.8359	0.6049	0.8689	0.3251	0.5558	$C_{1,\mathrm{L}}$
MSE	2.0073	1.9182	0.6630	0.7388	0.9865	1.4186	0.2550	0.3928	

using a robust procedure, estimating p reduces mean square errors of the marginal location estimators under $C_{0,\mathrm{L}}$; the lower values correspond to estimating the missing data probability with the kernel estimator.

6. An example

Daniel and Wood (1980) studied 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 $\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}}$ represent the feed compositions and the covariate t is the logarithm of a combination of process conditions scaled to [0, 1]. We computed the estimators with bandwidth $h = 0.06$. In order to avoid boundary effects and to improve the performance of the regression function estimator, we used Gasser and Müller's weights with boundary kernels, as described in González-Manteiga and Aneiros-Pérez (2003).

We first compute the estimates of the marginal location θ for this data set. The robust estimators were computed as in Section 5. We have used the gaussian kernel with standard deviation 0.37 and the robust smoothing procedure used local M-estimates with score function ψ_1, the bisquare function with tuning constant 4.685, and the local medians as initial estimates. The same score function was used to compute the marginal estimators, that is, we choose $\psi_2 = \psi_1$. On the other hand, to compute the robust estimator of the regression parameter $\boldsymbol{\beta}$ we used as *rho*-functions the bisquare function given in (14) with $c_0 = 1.56$ and $c_1 = 3.44$ that guarantees, under a regression model, an 85% efficiency. In this case, due to the dimension of the covariates, we have selected a smaller tuning constant to make a trade-off between bias and efficiency. We will denote with a superscript (1) the resulting estimators computed with the whole data set, i.e., with $p(\mathbf{x}, t) \equiv 1$. The values obtained are $\widehat{\theta}^{(1)}_{\mathrm{WS},\mathrm{LS}} = 91.855$, $\widehat{\theta}^{(1)}_{\mathrm{A},\mathrm{LS}} = 91.901$, $\widehat{\theta}^{(1)}_{\mathrm{WI},\mathrm{LS}} = 91.855$, $\widehat{\theta}^{(1)}_{\mathrm{WS},\mathrm{R}} = 91.667$, $\widehat{\theta}^{(1)}_{\mathrm{A},\mathrm{R}} = 91.670$ and $\widehat{\theta}^{(1)}_{\mathrm{WI},\mathrm{R}} = 91.668$. It is worth noticing that, in all cases, the robust procedures lead to smaller values than the classical ones. To study the significance of their differences, we have computed, through jackniffe, the standard deviations of the difference $\Delta^{(1)}_{\alpha} = \widehat{\theta}^{(1)}_{\alpha,\mathrm{R}} - \widehat{\theta}^{(1)}_{\alpha,\mathrm{LS}}$ between the least squares and robust estimates, with $\alpha = \mathrm{WS}, \mathrm{A}$ or WI. These values were used to calculate the ratios $R^{(1)}_{\alpha} = \Delta^{(1)}_{\alpha}/\widehat{\sigma}_{\Delta^{(1)}_{\alpha}}$ leading to $R^{(1)}_{\mathrm{WS}} = 1.885$, $R^{(1)}_{\mathrm{A}} = 2.000$ and $R^{(1)}_{\mathrm{WI}} = 1.735$. These results show that the classical and robust estimators are significantly different from each other which make us suspect that some atypical data influencing the marginal estimators are present.

Indeed, Daniel and Wood (1980) discussed the presence of three anomalous observations (labeled 75 to 77) which correspond to high values of octanes associated with high leverage points. We repeat the analysis for the classical procedure excluding these three observations and we also compute standard deviations of the least squares estimates, reported between brackets, through jackniffe. The subscript NO is added to identify the estimators computed without these three outlying data. The values obtained are $\widehat{\theta}^{(1)}_{\mathrm{WS},\mathrm{LS},\mathrm{NO}} = 91.674\ (0.118)$, $\widehat{\theta}^{(1)}_{\mathrm{A},\mathrm{LS},\mathrm{NO}} = 91.716\ (0.130)$, $\widehat{\theta}^{(1)}_{\mathrm{WI},\mathrm{LS},\mathrm{NO}} = 91.674\ (0.118)$ which are quite similar to the values obtained with the related robust method. Moreover, we compute the standardized absolute differences, *SAD*, between the classical estimates computed with all data and without the outliers. To compute the standard deviation of the differences $D^{(1)}_{\alpha,\mathrm{LS}} = \widehat{\theta}^{(1)}_{\alpha,\mathrm{LS}} - \widehat{\theta}^{(1)}_{\alpha,\mathrm{LS},\mathrm{NO}}$ and $D^{(1)}_{\alpha,\mathrm{R}} = \widehat{\theta}^{(1)}_{\alpha,\mathrm{R}} - \widehat{\theta}^{(1)}_{\alpha,\mathrm{LS},\mathrm{NO}}$, we have proceeded using an adaptive jackniffe as follows. For each $1 \le i \le n$, denote as $\widehat{\theta}^{(-i)}$ the estimators computed without the i-th observation. Clearly, when computing the least squares estimator with the sample without outliers only those indexes i with $i \neq 75, 76, 77$ are used since these observations are not in the new sample. We have then compared the differences

computed when removing the observation $(y_i, \mathbf{x}_i, t_i)$ where $i \neq 75, 76, 77$ to estimate the standard deviations of the differences $D^{(1)}_{\alpha,\text{LS}}$ and $D^{(1)}_{\alpha,\text{R}}$. Let us denote the standardized differences as $SAD^{(1)}_{\alpha,\text{LS}} = |D^{(1)}_{\alpha,\text{LS}}|/\widehat{\sigma}_{D^{(1)}_{\alpha,\text{LS}}}$ and $SAD^{(1)}_{\alpha,\text{R}} = |D^{(1)}_{\alpha,\text{R}}|/\widehat{\sigma}_{D^{(1)}_{\alpha,\text{R}}}$, respectively. The observed values are $SAD^{(1)}_{\text{WS},\text{LS}} = 2.016$, $SAD^{(1)}_{\text{A},\text{LS}} = 1.979$, $SAD^{(1)}_{\text{WI},\text{LS}} = 2.016$, $SAD^{(1)}_{\text{WS},\text{R}} = 0.100$, $SAD^{(1)}_{\text{A},\text{R}} = 0.583$ and $SAD^{(1)}_{\text{WI},\text{R}} = 0.098$ showing the high sensitivity of the classical procedure to anomalous data.

Missing responses were introduced completely at random with probability $p(\mathbf{x}, t) \equiv 0.8$ and the analysis described was repeated. The values of the estimators with outliers are $\widehat{\theta}_{\text{WS},\text{LS}} = 92.013$, $\widehat{\theta}_{\text{A},\text{LS}} = 91.905$, $\widehat{\theta}_{\text{WI},\text{LS}} = 91.888$, $\widehat{\theta}_{\text{WS},\text{R}} = 91.782$, $\widehat{\theta}_{\text{A},\text{R}} = 91.678$ and $\widehat{\theta}_{\text{WI},\text{R}} = 91.690$. As with the complete data set, the robust procedures lead to smaller values than the classical ones. When excluding the three outliers identified by Daniel and Wood (1980), the results are $\widehat{\theta}_{\text{WS},\text{LS},\text{NO}} = 91.795\,(0.130)$, $\widehat{\theta}_{\text{A},\text{LS},\text{NO}} = 91.711\,(0.122)$, $\widehat{\theta}_{\text{WI},\text{LS},\text{NO}} = 91.697\,(0.110)$ which are quite similar to the values obtained with the related robust method. Besides, we computed the standardized absolute differences defined as above with the data set with missing observations, that is, as above, for $\alpha = \text{WS}, \text{A}$ and WI, we define $D_{\alpha,\text{LS}} = \widehat{\theta}_{\alpha,\text{LS}} - \widehat{\theta}_{\alpha,\text{LS},\text{NO}}$, $D_{\alpha,\text{R}} = \widehat{\theta}_{\alpha,\text{R}} - \widehat{\theta}_{\alpha,\text{LS},\text{NO}}$ and denote by $SAD_{\alpha,\text{LS}}$ and $SAD_{\alpha,\text{R}}$ the ratios $SAD_{\alpha,\text{LS}} = \left|D_{\alpha,\text{LS}}\right|/\widehat{\sigma}_{D_{\alpha,\text{LS}}}$ and $SAD_{\alpha,\text{R}} = \left|D_{\alpha,\text{R}}\right|/\widehat{\sigma}_{D_{\alpha,\text{R}}}$. The results obtained are $SAD_{\text{WS},\text{LS}} = 2.062$, $SAD_{\text{A},\text{LS}} = 1.983$, $SAD_{\text{WI},\text{LS}} = 2.116$, $SAD_{\text{WS},\text{R}} = 0.166$, $SAD_{\text{A},\text{R}} = 0.421$, $SAD_{\text{WI},\text{R}} = 0.112$ and so, the resistance to anomalous data of the robust proposals is preserved under this missing data scheme. We have also computed, using jackniffe, the standard deviations of the difference $\Delta_\alpha = \widehat{\theta}_{\alpha,\text{R}} - \widehat{\theta}_{\alpha,\text{LS}}$ between the least squares and robust estimates. These values were used to calculate the ratios $R_\alpha = \Delta_\alpha/\widehat{\sigma}_{\Delta_\alpha}$ leading to $R_{\text{WS}} = 1.934$, $R_{\text{A}} = 2.175$ and $R_{\text{WI}} = 2.092$. These results show that the classical and robust estimators are significantly different from each other and confirm the sensitivity of all the procedures based on linear estimators, in particular of the average linear estimator when high leverage points are present.

7. Final comments

We have introduced three robust procedures for estimating the marginal location parameter under a partially linear model when there are missing observations in the response variable and it can be suspected that anomalous observations are present in the sample. All procedures are Fisher-consistent and thus they lead to strongly consistent estimators.

Under the contaminations considered, they show their advantage over the classical estimators. Moreover, the average and weighted imputed M-estimators, although they are computationally more expensive, should be used since they perform better than the weighted simplified M-estimator in all situations. Both the classical and robust procedures do not seem to be very sensitive to the choice of the smoothing parameter and so an exhaustive bandwidth search can be avoided. As mentioned by Wang and Sun (2007) the selection of bandwidths is not so critical if one is only interested in estimation of parametric components.

The results of our simulation study suggest that smaller mean square errors can be attained using a smooth estimator of the missing data probabilities instead of a parametric one if the dimension of the covariates and the number of observations allow one to compute the kernel estimator.

Acknowledgments

The authors would like to thank two anonymous referees for their valuable comments and suggestions that led to improvement of the presentation of the paper. This work began while Graciela Boente and Ana Bianco were visiting the Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela. This research was partially supported by Grants PIP 112-200801-00216 from CONICET, PICT 21407 and 821 from ANPCYT and X-018 from the Universidad de Buenos Aires at Buenos Aires, Argentina, and also by the DGICYT Spanish Grants MTM2005-00820 and MTM2008-03010 (European FEDER support included).

Appendix

In Appendix A.1 we give the proofs of the theorems stated in Section 4, while in Appendices A.2 and A.3, we will study the Fisher-consistency of the given proposals.

A.1. Proofs

Proof of Theorem 4.1. (a) It is enough to show that for any borelian set B, $\widehat{\phi}(B) \xrightarrow{a.s.} P(Y \in B)$ where

$$\widehat{\phi}(B) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} I_B(y_i).$$

Note that $\widehat{\phi}(B) = S_{1n} + S_{2n}$ where

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

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

Using **A2** and **A3**, we have that $|S_{1n}| \xrightarrow{a.s.} 0$. On the other hand, using the strong law of large numbers and the MAR assumption, we have that $S_{2n} \xrightarrow{a.s.} P(Y \in B)$, concluding the proof.

The proof of (b) follows easily using (a), the following bound:

$$\sup_{\substack{a\in\mathbb{R} \\ \sigma\in[\sigma_0/2, 2\sigma_0]}} |\widehat{\lambda}(a,\sigma) - \lambda(a,\sigma)| \le \int |\psi_2'(u)| \mathrm{d}u \, \|\widehat{F}_n - F\|_\infty,$$

the continuity of $\lambda(a,\sigma)$ as a function of σ and the fact that in a neighborhood of θ, the function $\lambda(a,\sigma_0)$ has a unique change of sign. □

Proof of Theorem 4.2. (a) It is enough to show that $\Pi(\widetilde{P}_n, \widetilde{\widetilde{P}}_n)| \xrightarrow{a.s.} 0$ with

$$\widetilde{\widetilde{P}}_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_A(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(t_i)).$$

This result follows if we show that for any bounded and continuous function $f : \mathbb{R} \longrightarrow \mathbb{R}$ we have that

$$\left| E_{\widetilde{\widetilde{P}}_n}(f) - E_{\widetilde{P}_n}(f) \right| \xrightarrow{a.s.} 0$$

which follows using arguments analogous to those considered in Lemma 1 of Bianco and Boente (2004).
The proof of (b) is derived as in Theorem 4.1 using the following bound:

$$\sup_{\substack{a\in\mathbb{R} \\ \sigma\in[\sigma_0/2, 2\sigma_0]}} |\widetilde{\widetilde{\lambda}}_n(a,\sigma) - \lambda_Z(a,\sigma)| \le 2\|\psi_2'\|_\infty \Pi(\widetilde{P}_n, P_Z)$$

with $\lambda_Z(a,\sigma) = E\psi_2((Z-a)/\sigma)$. □

Proof of Theorem 4.3. (a) It is enough to show that for any borelian set B, $\widehat{\widehat{\phi}}(B) \xrightarrow{a.s.} P(Y \in B)$ where

$$\widehat{\widehat{\phi}}(B) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)} I_B(y_i) + \frac{1}{n}\sum_{i=1}^{n} \left[\left(1 - \frac{\delta_i}{p_n(\mathbf{x}_i, t_i)}\right) I_B(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i)) \right].$$

Note that $\widehat{\widehat{\phi}}(B) = \widehat{\phi}(B) + S_{1n} + S_{2n}$ where $\widehat{\phi}(B)$ is defined in the proof of Theorem 4.1 and

$$S_{1n} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{p_n(\mathbf{x}_i, t_i)} - \frac{1}{p(\mathbf{x}_i, t_i)} \right] \delta_i \, I_B(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i))$$

$$S_{2n} = \frac{1}{n}\sum_{i=1}^{n} \left[1 - \frac{\delta_i}{p(\mathbf{x}_i, t_i)} \right] I_B(\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{g}_n(t_i)).$$

In the proof of Theorem 4.1 it was shown that $\widehat{\phi}(B) \xrightarrow{a.s.} P(Y \in B)$. Using **A2**, we have that

$$|S_{1n}| \le \frac{1}{A \inf_{(\mathbf{x},t)} p_n(\mathbf{x}, t)} \sup_{(\mathbf{x},t)} |p_n(\mathbf{x}, t) - p(\mathbf{x}, t)|$$

which together with **A3** entails that $S_{1n} \xrightarrow{a.s.} 0$. Besides,

$$|S_{2n}| \le \frac{1}{n}\sum_{i=1}^{n} \left[1 - \frac{\delta_i}{p(\mathbf{x}_i, t_i)} \right]$$

and so, using the strong law of large numbers, we have that $S_{2n} \xrightarrow{a.s.} 0$, concluding the proof.
The proof of (b) follows as in Theorem 4.1. □

A.2. Fisher-consistency of the parametric and nonparametric components

Fisher-consistency will be derived under the more general heteroscedastic model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(t_i) + \sigma_0(\mathbf{x}_i, t_i)\,\epsilon_i$, $1 \le i \le n$, with the errors ϵ_i i.i.d. and independent of the covariates. Moreover, we will assume that for any $\sigma > 0$:

(a) $E(\psi_1(\epsilon/\sigma)) = 0$,
(b) $E(\psi_1((\epsilon - a)/\sigma)) = 0 \Rightarrow a = 0$,
(c) $E(\rho((\epsilon - a)/\sigma)) \ge E(\rho(\epsilon/\sigma))$.

For a discussion on condition (a) see Remark 2.1.1. Note that condition (b) is fulfilled if, for instance, ψ_1 is strictly monotone. On the other hand, using that ρ is a *rho*-function, from well-known results on robust location estimation, we have that the symmetry of the error distribution implies (c).

We first consider the functionals defined in Section 2.1. Note that

$$E\left[\delta\psi_1\left(\frac{Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - g_{\boldsymbol{\beta}}(T)}{\sigma}\right)\middle| T\right] = E\left[p(\mathbf{X}, T)\psi_1\left(\frac{Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - g_{\boldsymbol{\beta}}(T)}{\sigma}\right)\middle| T\right]$$
$$= E\left\{p(\mathbf{X}, T)E\left[\psi_1\left(\frac{\mathbf{X}^{\mathrm{T}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + (g_0(T) - g_{\boldsymbol{\beta}}(T)) + \sigma_0(\mathbf{X}, T)\epsilon}{\sigma}\right)\middle| (\mathbf{X}, T)\ \right]\middle| T\right\}.$$

Thus, it is easy to see that condition (b) entails that $g_{\boldsymbol{\beta}_0}(t) \equiv g_0(t)$. On the other hand, we get easily that

$$E\left[\delta\rho\left(\frac{Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - g_{\boldsymbol{\beta}}(T)}{\sigma}\right)\upsilon(\mathbf{X})\right] = E\left[p(\mathbf{X}, T)\rho\left(\frac{Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - g_{\boldsymbol{\beta}}(T)}{\sigma}\right)\upsilon(\mathbf{X})\right]$$
$$= E\left[p(\mathbf{X}, T)\rho\left(\frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(T) + \sigma_0(\mathbf{X}, T)\epsilon - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - g_{\boldsymbol{\beta}}(T)}{\sigma}\right)\upsilon(\mathbf{X})\right]$$
$$= E\left[p(\mathbf{X}, T)\upsilon(\mathbf{X})E\left\{\rho\left(\frac{\sigma_0(\mathbf{X}, T)\epsilon + \mathbf{X}^{\mathrm{T}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + g_0(T) - g_{\boldsymbol{\beta}}(T)}{\sigma}\right)\middle| (\mathbf{X}, T)\right\}\right].$$

From the independence between the errors and the covariates and condition (c), we get

$$E\left\{\rho\left(\frac{\sigma_0(\mathbf{x}, t)\epsilon + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + g_0(t) - g_{\boldsymbol{\beta}}(t)}{\sigma}\right)\middle| (\mathbf{X}, T) = (\mathbf{x}, t)\right\}$$
$$= E\left\{\rho\left(\frac{\sigma_0(\mathbf{x}, t)\epsilon + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + g_0(t) - g_{\boldsymbol{\beta}}(t)}{\sigma}\right)\right\}$$
$$\ge E\left\{\rho\left(\frac{\sigma_0(\mathbf{x}, t)\epsilon}{\sigma}\right)\right\}.$$

and so $\boldsymbol{\beta}(F) = \boldsymbol{\beta}_0$, which concludes the proof.

A.3. Fisher-consistency of the marginal location functionals

For the sake of simplicity, throughout this section we assume that the error distribution is symmetric and $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(t_i) = \theta + u_i$ where u_i has a symmetric distribution too. The functionals related to the proposed estimators are given by:

- **Weighted simplified functional**. This functional is the solution, $\theta_{\mathrm{ws}}(F)$, of

$$E\frac{\delta}{p(\mathbf{X}, T)}\psi_2\left(\frac{Y - \theta_{\mathrm{ws}}(F)}{\sigma}\right) = 0.$$

Note that, by taking conditional expectation and using that we have a MAR missingness scheme, we have

$$E\frac{\delta}{p(\mathbf{X}, T)}\psi_2\left(\frac{Y - \theta_{\mathrm{s}}(F)}{\sigma}\right) = E\frac{p(\mathbf{X}, T)}{p(\mathbf{X}, T)}\psi_2\left(\frac{Y - \theta_{\mathrm{s}}(F)}{\sigma}\right)$$
$$= E\psi_2\left(\frac{Y - \theta_{\mathrm{s}}(F)}{\sigma}\right),$$

and so $\theta_{\mathrm{ws}}(F) = \theta$ if $u + \sigma_0\epsilon$ has a symmetric distribution.

- **Averaged M-functional**. The functional $\theta_{\mathrm{A}}(F)$ is the solution of

$$E\psi_2\left(\frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(F) + g_{\boldsymbol{\beta}(F)}(T) - \theta_{\mathrm{A}}(F)}{\sigma}\right) = 0.$$

Using that ϵ has a symmetric distribution, from Appendix A.2 we get $\boldsymbol{\beta}(F) = \boldsymbol{\beta}_0$ and $g_{\boldsymbol{\beta}(F)} = g_0$. Thus,

$$E\psi_2\left(\frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(F) + g_{\boldsymbol{\beta}(F)}(T) - \theta_{\mathrm{MA}}(F)}{\sigma}\right) = E\psi_2\left(\frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(T) - \theta_{\mathrm{MA}}(F)}{\sigma}\right)$$
$$= E\psi_2\left(\frac{u + \theta - \theta_{\mathrm{MA}}(F)}{\sigma}\right).$$

Since we have assumed that u has a symmetric distribution, we obtain that $\theta_{\mathrm{A}} = \theta$.

- **Weighted imputed functional**. The functional $\theta_{\mathrm{WI}}(F)$ solves

$$E\left[\frac{\delta}{p(\mathbf{X}, T)}\psi_2\left(\frac{Y - \theta_{\mathrm{w}}(F)}{\sigma}\right) + \left(1 - \frac{\delta}{p(\mathbf{X}, T)}\right)\psi_2\left(\frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(F) + g_{\boldsymbol{\beta}(F)}(T) - \theta_{\mathrm{w}}(F)}{\sigma}\right)\right] = 0.$$

As above, we get

$$E\left[\frac{\delta}{p(\mathbf{X}, T)}\psi_2\left(\frac{Y - \theta_{\mathrm{w}}(F)}{\sigma}\right) + \left(1 - \frac{\delta}{p(\mathbf{X}, T)}\right)\psi_2\left(\frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(F) + g_{\boldsymbol{\beta}(F)}(T) - \theta_{\mathrm{w}}(F)}{\sigma}\right)\right]$$
$$= E\left[\frac{p(\mathbf{X}, T)}{p(\mathbf{X}, T)}\psi_2\left(\frac{Y - \theta_{\mathrm{w}}(F)}{\sigma}\right) + \left(1 - \frac{p(\mathbf{X}, T)}{p(\mathbf{X}, T)}\right)\psi_2\left(\frac{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(F) + g_{\boldsymbol{\beta}(F)}(T) - \theta_{\mathrm{w}}(F)}{\sigma}\right)\right]$$
$$= E\psi_2\left(\frac{Y - \theta_{\mathrm{w}}(F)}{\sigma}\right)$$

and so $\theta_{\mathrm{WI}}(F) = \theta$.

References

Aerts, M., Claeskens, G., Hens, N., Molenberghs, G., 2002. Local multiple imputation. Biometrika 89 (2), 375–388.
Bianco, A., Boente, G., 2004. Robust estimators in semiparametric partly linear regression models. J. Statist. Plann. Inference 122, 229–252.
Bianco, A., Boente, G., 2007. Robust estimators under a semiparametric partly linear autoregression model: Asymptotic behavior and bandwidth selection. J. Time Ser. Anal. 28, 274–306.
Bianco, A., Boente, G., González-Manteiga, W., Pérez-González, A., 2009. Estimation of the marginal location under a partially linear model with missing responses. Available at: http://www.ic.fcen.uba.ar/preprints/biaboegonper.pdf.
Bianco, A., Boente, G., Martínez, E., 2006. Robust tests in semiparametric partly linear models. Scandinavian J. Statist. 33, 435–450.
Boente, G., González-Manteiga, W., Pérez-González, A., 2009. Robust nonparametric estimation with missing data. J. Statist. Plann. Inference 139, 571–592.
Cantoni, E., Ronchetti, E., 2001. Resistant selection of the smoothing parameter for smoothing splines. Statist. Comput. 11, 141–146.
Chen, J., Fan, J., Li, K., Zhou, H., 2006. Local quasi-likelihood estimation with data missing at random. Statist. Sinica 16, 1071–1100.
Chen, J., Shao, J., 2000. Nearest neighbor imputation for survey data. J. Official Statist. 16, 113–131.
Cheng, P.E., 1994. Nonparametric estimation of mean functionals with data missing at random. J. Amer. Statist. Assoc. 89, 81–87.
Chu, C.K., Cheng, P.E., 1995. Nonparametric regression estimation with missing data. J. Statist. Plann. Inference 48, 85–99.
Croux, C., Haesbroeck, G., 2003. Implementing the Bianco and Yohai estimator for logistic regression. Comput. Statist. Data Anal. 44, 273–295.
Daniel, C., Wood, F., 1980. Fitting Equations to Data: Computer Analysis of Multifactor Data. Wiley, New York.
González-Manteiga, W., Aneiros-Pérez, G., 2003. Testing in partial linear regression models with dependent errors. J. Nonparametr. Stat. 15, 93–111.
González-Manteiga, W., Pérez-González, A., 2004. Nonparametric mean estimation with missing data. Comm. Statist. Theory Methods 33, 277–303.
He, X., Zhu, Z., Fung, W., 2002. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika 89, 579–590.
Huber, P., 1981. Robust Statistics. Wiley, New York.
Maronna, R., Martin, D., Yohai, V., 2006. Robust Statistics: Theory and Methods. Wiley, New York.
Meng, X.-L., 2000. Missing data: Dial M for ??? J. Amer. Statist. Assoc. 95 (452), 1325–1330.
Neyman, J., 1938. Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc. 33, 101–116.
Robins, J., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846–866.
Scharfstein, D., Rotnitzky, A., Robins, J., 1999. Adjusting for nonignorable drop out in semiparametric nonresponse models (with discussion). J. Amer. Statist. Assoc. 94, 1096–1146.
Wang, C., Wang, S., Gutierrez, R., Carroll, R., 1998. Local linear regression for generalized linear models with missing data. Ann. Statist. 26, 1028–1050.
Wang, C., Wang, S., Zhao, L.P., Ou, S.T., 1997. Weighted semiparametric estimation in regression analysis regression with missing covariates data. J. Amer. Statist. Assoc. 92, 512–525.
Wang, F., Scott, D., 1994. The L_1 method for robust nonparametric regression. J. Amer. Statist. Assoc. 89, 65–76.
Wang, Q., Linton, O., Härdle, W., 2004. Semiparametric regression analysis with missing response at random. J. Amer. Statist. Assoc. 99 (466), 334–345.
Wang, Q., Sun, Z., 2007. Estimation in partially linear models with missing responses at random. J. Multivariate Anal. 98, 1470–1493.
Wang, W., Rao, J., 2002. Empirical likelihood-based inference under imputation for missing response data. Ann. Statist. 30, 896–924.
Yates, F., 1933. The analysis of replicated experiments when the field results are incomplete. Emporium J. Exp. Agriculture 1, 129–142.