Test (2011) 20:524–548
DOI 10.1007/s11749-010-0216-0

ORIGINAL PAPER

Asymptotic behavior of robust estimators in partially linear models with missing responses: the effect of estimating the missing probability on the simplified marginal estimators

Ana Bianco · Graciela Boente · Wenceslao González-Manteiga · Ana Pérez-González

Received: 21 August 2009 / Accepted: 8 October 2010 / Published online: 4 November 2010

Abstract In this paper, we consider a semiparametric partially linear regression model where missing data occur in the response. We derive the asymptotic behavior of the robust estimators for the regression parameter and of the weighted simplified location estimator introduced in Bianco et al. (Comput. Stat. Data Anal. 54:546–564, 2010a). For the latter, consistency results and the asymptotic distribution are derived when the missing probability is known and also when it is estimated.

Keywords Fisher-consistency · Kernel weights · M-location functionals · Missing at random · Nonparametric regression · Robust estimation

Mathematics Subject Classification (2000) 62G35 · 62F12

Communicated by Domingo Morales.

A. Bianco · G. Boente
Departamento de Matemáticas and Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires 1428, Argentina

A. Bianco
e-mail: abianco@dm.uba.ar

G. Boente
e-mail: gboente@dm.uba.ar

W. González-Manteiga
Departamento de Estatística e Investigación Operativa, Facultad de Matemáticas, Universidad de Santiago de Compostela, Campus Sur., 15706 Santiago de Compostela, Spain
e-mail: wenceslao.gonzalez@usc.es

A. Pérez-González (✉)
Departamento de Estadística e Investigación Operativa, Universidad de Vigo, Campus Orense. Campus Universitario As Lagoas s/n, 32004 Ourense, Spain
e-mail: anapg@uvigo.es

1 Introduction

Consider the partially linear regression model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(t_i) + \epsilon_i$, $1 \le i \le n$, where the response $y_i \in \mathbb{R}$ and the covariates $(\mathbf{x}_i^{\mathrm{T}}, t_i)$ are such that $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$, while the errors ϵ_i are i.i.d., independent of $(\mathbf{x}_i^{\mathrm{T}}, t_i)$ satisfying $E(\epsilon_i) = 0$ and $\mathrm{VAR}(\epsilon_i) < \infty$. Partly linear models are more flexible than standard linear models since they have a parametric and a nonparametric component. They can be a suitable choice when one suspects that the response y linearly depends on $\mathbf{x}$ but is nonlinearly related to t. This model has gained attention in recent years. An extensive description of the different results obtained in partly linear regression models can be found in Härdle et al. (2000). He et al. (2002) considered M-type estimates for repeated measurements using B-splines, while Bianco and Boente (2004) considered a kernel-based three-step procedure to define robust estimates under the partly linear model.

In practice, some response variables may be missing, by design (as in two-stage studies) or by happenstance. As is well known, the methods described above are designed for complete data sets and problems arise when missing observations are present. Even if 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. Actually, missing responses are very common in opinion polls, market research surveys, mail enquiries, socioeconomic investigations, medical studies and other scientific experiments. Wang et al. (2004) considered inference on the mean of y under regression imputation of missing responses based on the semiparametric regression model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + g_0(t_i) + \epsilon_i$. The estimator of the regression parameter $\boldsymbol{\beta}_0$, introduced by Wang et al. (2004), is a least-squares regression estimator defined by considering preliminary kernel estimators, of the quantities $E(\delta_1 \mathbf{x}_1 | t_1 = t)/E(\delta_1 | t_1 = t)$ and $E(\delta_1 y_1 | t_1 = t)/E(\delta_1 | t_1 = t)$, where $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if y_i is missing. Based on this estimator, estimators of the marginal mean of the responses y are defined using an imputation estimator and a number of propensity-score weighting estimators. On the other hand, Wang and Sun (2007) considered estimators of the regression coefficients and the nonparametric function using either imputation, semiparametric regression surrogate or an inverse marginal-probability weighted approach. These estimators are based on weighted means of the response variables, and so they are highly sensitive to anomalous data. This fact motivated the need of considering procedures resistant to outliers as those given in Bianco et al. (2010a), who introduced robust estimators based on bounded score functions together with algorithms to compute them. Moreover, consistency of the marginal estimators was derived under certain regularity conditions therein.

In this paper, we go further and we focus our attention on the asymptotic behavior of the robust estimators of the regression parameter and the marginal location y, say θ, 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 reviews the definition of the robust semiparametric estimators. The consistency and the asymptotic distribution of the regression parameter are derived in Sect. 3, while the asymptotic distribution of the marginal location estimator is studied in Sect. 4 where weak consistency results are also discussed. For the marginal simplified location estimator, the asymptotic distribution is derived in the situation in which the missing probability is known and also when it is estimated under two different frameworks. In many situations, a parametric model can be assumed for the missing probability and the influence of estimating the parameters of the model on the distribution of the marginal location estimators needs to be quantified. In particular, if a logistic model is assumed and the parameters are estimated using the maximum likelihood estimator, a reduction in the variance is obtained with respect to the estimator computed with the true missing probability, denoted $p(\mathbf{x}, t)$. On the other hand, if the parameters are estimated robustly, we argue that a larger variance may be obtained. Besides, if a kernel estimator is used to estimate $p(\mathbf{x}, t)$, then a reduction of variance is always achieved and so, as recommended in Bianco et al. (2010a), this estimator should be used whenever it is possible. Finally, Sect. 5 presents some concluding remarks while technical proofs are relegated to the Appendix.

2 The robust estimators

Suppose we obtain a random sample of incomplete data $(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i, \delta_i)$, $1 \le i \le n$, where $\delta_i = 1$ if y_i is observed, $\delta_i = 0$ if y_i is missing. Furthermore, assume that the responses y_i satisfy a partially linear model, i.e.,

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

where the errors ϵ_i are independent, independent of $(\mathbf{x}_i^{\mathrm{T}}, t_i)$ and identically distributed with symmetric distribution $F_0(\cdot)$, that is, we assume that the error scale equals 1 to identify the parameter σ_0.

Let $(y, \mathbf{x}^{\mathrm{T}}, t, \delta)$ be a random vector with the same distribution as $(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i, \delta_i)$. As mentioned in the Introduction, our aim is to study the asymptotic behavior of the robust estimators of the regression parameter and the marginal location. For that purpose, an ignorable missing mechanism will be imposed by assuming that y is missing at random (MAR), that is, δ and y are conditionally independent given $(\mathbf{x}, t)$, i.e., $P(\delta = 1 | (y, \mathbf{x}, t)) = P(\delta = 1 | (\mathbf{x}, t)) = p(\mathbf{x}, t)$.

For the sake of completeness, we will briefly remind the definition of the estimators.

2.1 Estimators of the regression parameter and regression function

As mentioned in Bianco et al. (2010a), the estimation of the robust location conditional functional related to each component of $\mathbf{x}_i$ causes no problem since the data set is complete, while that of the response y_i is problematic since there are missing responses. As noted therein, if one proceeds 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 was considered by combining the M-smoothers defined in Boente et al. (2009) with robust regression estimators. Let ψ_1 be an odd and bounded score function and ρ be a *rho*-function as defined in Maronna et al. (2006, Chap. 2), i.e., a function ρ such that $\rho(x)$ is a nondecreasing function of $|x|$, $\rho(0)=0$, and $\rho(x)$ is increasing for $x>0$ when $\rho(x)<\|\rho\|_\infty$. If ρ is bounded, it is also assumed that $\|\rho\|_\infty=1$. We will consider kernel smoothers weights for the nonparametric component adapted to the missing setting, which are given by $w_i(\tau)=K((t_i-\tau)/h_n)\delta_i\{\sum_{j=1}^n K((t_j-\tau)/h_n)\delta_j\}^{-1}$, with K a kernel function, i.e., a non-negative integrable function on $\mathbb{R}$ and h_n the bandwidth parameter.

Let F be the distribution of $(y,\mathbf{x}^{\mathrm{T}},t,\delta)$. To define a robust estimator for the regression parameter, Bianco et al. (2010a) proceed as follows.

Step 1. For each τ and $\boldsymbol{\beta}$, define $g_{\boldsymbol{\beta}}(\tau)$ and its related estimate $\widehat{g}_{\boldsymbol{\beta}}(\tau)$ as the solution of $S^{(1)}(g_{\boldsymbol{\beta}}(\tau),\boldsymbol{\beta},\tau)=0$ and $S_n^{(1)}(\widehat{g}_{\boldsymbol{\beta}}(\tau),\boldsymbol{\beta},\tau)=0$, respectively, where

$$S^{(1)}(a,\boldsymbol{\beta},\tau)=E\left[\delta\psi_1\left(\frac{y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-a}{\sigma_{\boldsymbol{\beta}}}\right)\upsilon(\mathbf{x})|t=\tau\right], \tag{2}$$

$$S_n^{(1)}(a,\boldsymbol{\beta},\tau)=\sum_{i=1}^n w_i(\tau)\psi_1\left(\frac{y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-a}{\widehat{s}_{\boldsymbol{\beta}}}\right)\upsilon(\mathbf{x}_i), \tag{3}$$

with $\widehat{s}_{\boldsymbol{\beta}}$ a preliminary robust consistent scale estimator of $\sigma_{\boldsymbol{\beta}}$, the scale of $y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-g_{\boldsymbol{\beta}}(\tau)$, and υ a weight function.

Step 2. Let $H(\boldsymbol{\beta})=E[\delta\rho((y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-g_{\boldsymbol{\beta}}(t))/\sigma_0)\upsilon(\mathbf{x})]$ and

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

with $\widehat{\sigma}$ a preliminary estimate of the scale σ_0, i.e., a robust M-scale computed using an initial (possibly inefficient) estimate of $\boldsymbol{\beta}_0$ with high breakdown point.
The functional $\boldsymbol{\beta}(F)$ and its related estimate $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}_n$ are defined as $\boldsymbol{\beta}(F)=\operatorname{argmin}_{\boldsymbol{\beta}} H(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\beta}}=\operatorname{argmin}_{\boldsymbol{\beta}} H_n(\boldsymbol{\beta})$.

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

As in any regression model, leverage points in the explanatory variables $\mathbf{x}$ can cause breakdown. To overcome this problem, *GM*-, *S*- and *MM*-estimators have been introduced; see for instance, Maronna et al. (2006). In Step 2, a loss function ρ combined with a weight υ is introduced to include both families of estimators. This proposal is thus resistant against outliers in the residuals and in the carriers $\mathbf{x}$ as well. In most situations, when considering *MM*-estimators, one chooses $\upsilon(\mathbf{x})\equiv 1$ since *MM*-estimators already control high-leverage points. An algorithm to compute these estimators is described in Bianco et al. (2010a). Therein, the authors considered initial *LMS*-estimators combined with *S*-estimators adapted to the partly linear setting, to obtain *MM*-estimators of the regression parameter.

Finally, let $\psi=\rho'$ be the derivative of the loss function ρ. Thus, the regression estimator defined in Step 2 is the solution of

$$H_n^{(1)}(\widehat{\boldsymbol{\beta}})=\sum_{i=1}^n\delta_i\psi\left(\frac{y_i-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}-\widehat{g}_{\widehat{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i)\left(\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\widehat{g}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right)=0. \tag{4}$$

2.2 Estimators of the marginal location

Let θ be the marginal location of y. For instance, if we are interested in the M-location parameter of y related to an increasing and bounded score function ψ_2, θ is the solution of $\lambda(a,\varsigma) = E\psi_2((y-a)/\varsigma) = 0$ for all $\varsigma > 0$. When the distribution of y is symmetric around θ and ψ_2 is odd, the M-location parameter coincides with the point of symmetry of y.

On the other hand, when considering redescending score functions, it is better to define the functional through a minimization problem to guarantee its uniqueness. In this case, let ρ_2 be a bounded *rho*-function, as described above, such that $\rho_2' = \psi_2$. Denote $\zeta(a,\varsigma) = E\rho_2((y-a)/\varsigma)$; then, θ is defined as $\theta = \operatorname{argmin}_a \zeta(a,\varsigma_0)$ with ς_0 the marginal scale. Note that, from Theorem 10.2 in Maronna et al. (2006), if y has a density f which is a decreasing function of $|y-\theta|$ and ρ_2 is any *rho*-function, then $\zeta(a,\varsigma)$ has a unique minimum at $a=\theta$ for any $\varsigma > 0$. Besides, as above, ψ_2 is an odd and bounded score function and θ is a solution of $\lambda(a,\varsigma) = 0$ for all ς. When $\rho_2(u) = |u|$, $\psi_2(u) = \mathrm{sg}(u) = I_{(0,\infty)}(u) - I_{(-\infty,0)}(u)$ and θ is the median of y.

Denote by $\widehat{\varsigma}$ any robust consistent estimator of the marginal scale ς_0 of the responses y, such as the MAD. To correct the bias caused in the estimation by the missing mechanism, an estimator of the missing probability needs to be considered. Denote by $p_n(\mathbf{x},t)$ any estimator of $p(\mathbf{x},t)$. The *weighted simplified M-estimate* was defined in Bianco et al. (2010a) as the solution, $\widehat{\theta}$, of $U_n(p_n,\widehat{\varsigma},\theta) = 0$ with

$$U_n(q,\varsigma,\theta) = \sum_{i=1}^{n} \frac{\delta_i}{q(\mathbf{x}_i,t_i)} \psi_2\left(\frac{y_i-\theta}{\varsigma}\right). \tag{5}$$

For redescending ψ_2 functions, it is better to define $\widehat{\theta}$ as the value $\widehat{\theta} = \operatorname{argmin}_\theta D_n(p_n,\widehat{\varsigma},\theta)$ where

$$D_n(q,\varsigma,\theta) = \sum_{i=1}^{n} \frac{\delta_i}{q(\mathbf{x}_i,t_i)} \rho_2\left(\frac{y_i-\theta}{\varsigma}\right). \tag{6}$$

3 Asymptotic behavior of the regression parameter estimators

In this section, we will derive the strong consistency and the asymptotic normality of the regression parameter.

3.1 Consistency of $\widehat{\boldsymbol{\beta}}$

We will assume that $t \in \mathcal{T} \subset \mathbb{R}$, and let $\mathcal{T}_0 \subset \mathcal{T}$ be a compact set. For any continuous function $v : \mathcal{T} \to \mathbb{R}$, we will denote $\|v\|_\infty = \sup_{t\in\mathcal{T}} |v(t)|$ and $\|v\|_{0,\infty} = \sup_{t\in\mathcal{T}_0} |v(t)|$. We will need the following set of assumptions:

C1. The function ρ and ψ_1 are continuous and bounded. Moreover, the function ρ is Lipschitz and v is bounded.

C2. The kernel $K : \mathbb{R} \to \mathbb{R}$ is an even, non-negative, continuous and bounded function, with bounded variation, satisfying $\int K(u)\,du = 1$, $\int u^2 K(u)\,du < \infty$ $|u|K(u) \to 0$ as $|u| \to \infty$.

C3. The bandwidth sequence h_n is such that $h_n \to 0$, $nh_n/\log(n) \to \infty$.

C4. The marginal density f_T of t is a bounded function. Moreover, given any compact set $\mathcal{T}_0 \subset \mathcal{T}$ there exists a positive constant $A_1(\mathcal{T}_0)$ such that $A_1(\mathcal{T}_0) < f_T(\tau)$ for all $\tau \in \mathcal{T}_0$.

C5. The function $S^{(1)}(a, \boldsymbol{\beta}, \tau)$ satisfies the following equicontinuity condition: for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $\tau_1, \tau_2 \in \mathcal{T}_0$ and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{K}$, a compact set in $\mathbb{R}^d$,

$$\begin{aligned} &|\tau_1 - \tau_2| < \delta \quad \text{and} \quad \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| < \delta \\ &\Rightarrow \quad \sup_{a \in \mathbb{R}} \left| S^{(1)}(a, \boldsymbol{\beta}_1, \tau_1) - S^{(1)}(a, \boldsymbol{\beta}_2, \tau_2) \right| < \epsilon. \end{aligned}$$

C6. The function $S^{(1)}(a, \boldsymbol{\beta}, \tau)$ is continuous, and $g_{\boldsymbol{\beta}}(\tau)$ is a continuous function of $(\boldsymbol{\beta}, \tau)$.

Remark 3.1 If the conditional distribution of $\mathbf{x}|t = \tau$ is continuous with respect to τ, the continuity and boundedness of ψ_1 stated in C1 entail that $S^{(1)}(a, \boldsymbol{\beta}, \tau)$ is continuous. Assumption C3 ensures that for each fixed a and $\boldsymbol{\beta}$ we have convergence of the kernel estimates to their mean, while C5 guarantees that the bias term converges to 0. Assumption C4 is a standard condition in semiparametric models. On the other hand, assumption C5 is fulfilled under C1 if the following equicontinuity condition holds: for any $\epsilon > 0$ there exist compact sets $\mathcal{K}_1 \subset \mathbb{R}$ and $\mathcal{K}_d \subset \mathbb{R}^d$ such that, for any $\tau \in \mathcal{T}_0$, $P((y, \mathbf{x}) \in \mathcal{K}_1 \times \mathcal{K}_d | t = \tau) > 1 - \epsilon$, which holds, for instance, if $x_{ij} = \phi_j(t_i) + u_{ij}$, $1 \le i \le n$, $1 \le j \le p$, where ϕ_j are continuous functions and u_{ij} are i.i.d. and independent of t_i.

It is worth noticing that C1 to C4 were also required in Bianco and Boente (2004) who introduced and studied robust estimators of $\boldsymbol{\beta}$ when there are no missing observations. On the other hand, as mentioned in Sect. 2, when missing responses arise, a profile-likelihood approach is needed. Hence, instead of requiring symmetry and equicontinuity to the conditional distributions of $y|t = \tau$ and $x_j|t = \tau$, $1 \le j \le p$, as in Bianco and Boente (2004), C5 and C6 need to be considered.

Theorem 3.1 *Let $\mathcal{K} \subset \mathbb{R}^d$ and $\mathcal{T}_0 \subset \mathcal{T}$ be compact sets such that $\mathcal{T}_\delta \subset \mathcal{T}$ where $\mathcal{T}_\delta$ is the closure of a δ neighborhood of $\mathcal{T}_0$. Assume that* C1 *to* C6 *and the following conditions hold*:

(i) *ψ_1 is of bounded variation*,

(ii) $\inf_{\boldsymbol{\beta} \in \mathcal{K}} \sigma_{\boldsymbol{\beta}} > 0$ *and* $\sup_{\boldsymbol{\beta} \in \mathcal{K}} |\widehat{s}_{\boldsymbol{\beta}} - \sigma_{\boldsymbol{\beta}}| \xrightarrow{\text{a.s.}} 0$, *where $\sigma_{\boldsymbol{\beta}}$ as defined in Step* 1.

Then, *we have*:

(a) $\sup_{\boldsymbol{\beta} \in \mathcal{K}, a \in \mathbb{R}} \|S_n^{(1)}(a, \boldsymbol{\beta}, \cdot) - S^{(1)}(a, \boldsymbol{\beta}, \cdot)\|_{0,\infty} \xrightarrow{\text{a.s.}} 0$.

(b) *If, in addition*, $S^{(1)}(a, \boldsymbol{\beta}, \tau) = 0$ *has a unique root* $g_{\boldsymbol{\beta}}(\tau)$, *then* $\sup_{\boldsymbol{\beta} \in \mathcal{K}} \|\widehat{g}_{\boldsymbol{\beta}} - g_{\boldsymbol{\beta}}\|_{0,\infty} \xrightarrow{\text{a.s.}} 0$.

The proof of Theorem 3.1 follows the same arguments as those in Theorem 3.1 of Boente et al. (2006), using the fact that assumption (ii) implies that the family of functions $\mathcal{F} = \{f(y, \mathbf{x}) = \psi_1((y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + a)/\sigma)\upsilon(\mathbf{x}),\ \boldsymbol{\beta} \in \mathcal{K}, a \in \mathbb{R}, \sigma > 0\}$ has covering number $N(\epsilon, \mathcal{F}, L^1(\mathbb{Q})) \leq A\epsilon^{-W}$, for any probability $\mathbb{Q}$ and $0 < \epsilon < 1$. Besides, the condition that $S^{(1)}(a, \boldsymbol{\beta}, \tau) = 0$ has a unique root is fulfilled if ψ_1 is a nondecreasing function and strictly increasing in a neighborhood of 0.

Theorem 3.2 *Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of $H_n(\boldsymbol{\beta})$, where $H_n(\boldsymbol{\beta})$ is defined in Step* 2 *with $\widehat{g}_{\boldsymbol{\beta}}$ satisfying* $\sup_{\boldsymbol{\beta}\in\mathcal{K}} \|\widehat{g}_{\boldsymbol{\beta}} - g_{\boldsymbol{\beta}}\|_{0,\infty} \xrightarrow{\text{a.s.}} 0$ *for any compact sets $\mathcal{K} \subset \mathbb{R}^d$ and $\mathcal{T}_0 \subset \mathcal{T}$. If* C1 *holds and $\widehat{\sigma} \xrightarrow{\text{a.s.}} \sigma_0$, then*

(a) $\sup_{\boldsymbol{\beta}\in\mathcal{K}} |H_n(\boldsymbol{\beta}) - H(\boldsymbol{\beta})| \xrightarrow{\text{a.s.}} 0$.

(b) *If, in addition, there exists a compact set $\mathcal{K}_1$ such that* $\lim_{m\to\infty} P(\bigcap_{n\geq m} \widehat{\boldsymbol{\beta}} \in \mathcal{K}_1) = 1$ *and $H(\boldsymbol{\beta})$ has a unique minimum at $\boldsymbol{\beta}_0$, then $\widehat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_0$.*

We also omit the proof of Theorem 3.2, since it follows as Theorem 3.2 of Boente et al. (2006).

Remark 3.2 Theorems 3.1 and 3.2 entail that $\|\widehat{g}_{\widehat{\boldsymbol{\beta}}} - g_0\|_{0,\infty} \xrightarrow{\text{a.s.}} 0$, for any compact set $\mathcal{T}_0 \subset \mathcal{T}$, since $g_{\boldsymbol{\beta}}(t)$ is continuous.

3.2 Asymptotic normality of $\widehat{\boldsymbol{\beta}}$

From now on, $\mathcal{T}$ is assumed to be a compact set. Assumptions N1–N6 under which the resulting estimates are asymptotically normally distributed are detailed in the Appendix.

Theorem 3.3 *Assume that t_1 is a random variable with distribution on a compact set $\mathcal{T}$. Assume that* N1–N6 *in the Appendix hold and that $\widehat{\sigma} \xrightarrow{p} \sigma_0$. Then, for any consistent solution $\widehat{\boldsymbol{\beta}}$ of* (4), *we have that*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N\big(\mathbf{0}, \sigma_0^2\mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-1}\big),$$

where the symmetric matrix $\mathbf{A}$ *is defined in* N3 *and* $\boldsymbol{\Sigma}$ *is defined in* N4.

It is worth noticing that, when $\upsilon(\mathbf{x}) \equiv 1$, the efficiency of the robust estimator $\widehat{\boldsymbol{\beta}}$ with respect to its linear relative, i.e., the least square estimator, equals $[E\psi'(\epsilon)]^{-2}E\psi^2(\epsilon)$, which corresponds to the very well known efficiency of any robust location M-estimator. This situation includes, in particular, *MM*-estimators and so, the same asymptotic efficiency as in the regression model is obtained in this case.

4 Asymptotic behavior of $\widehat{\theta}$

In this section, we will derive the asymptotic distribution of the weighted simplified M-estimate, $\widehat{\theta}$, under different situations, i.e., when the missing probability is assumed to be known or when it is estimated either parametrically or using a kernel approach. Different asymptotic variances are obtained in each situation. The goal of this section is to validate theoretically the results observed in the simulation study performed in Bianco et al. (2010a), i.e., to prove that estimating nonparametrically the missing probability reduces the variance of the estimator. The asymptotic results for the marginal location estimators to be derived will help to the comprehension of some non-intuitive numerical results observed in Bianco et al. (2010a). In particular, when a logistic model for the missing probability is assumed, the order among the asymptotic variances is discussed in detail in Sect. 4.3.

It is worth noticing that our results require weak consistency of the proposed estimators, i.e., that $\widehat{\theta} \xrightarrow{p} \theta$. Theorem 4.1 of Bianco et al. (2010a) states that there exists a solution $\widehat{\theta}$ of $U_n(p_n, a, \widehat{\varsigma}) = 0$, such that $\widehat{\theta} \xrightarrow{\text{a.s.}} \theta$ under strong uniform consistency of the missing probability, i.e., $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x}, t) - p(\mathbf{x}, t)| \xrightarrow{\text{a.s.}} 0$, smoothness conditions to the score function ψ_2 and if $\inf_{(\mathbf{x},t)} p(\mathbf{x}, t) = A > 0$, which states that some response variables are observed at each neighborhood of $(\mathbf{x}, t)$. It is worth noticing that the results obtained in Bianco et al. (2010a) do not give a full answer when the score function ψ_2 is not an increasing function. For that reason, we begin by establishing in Sect. 4.1 weak consistency results for both increasing and redescending score functions which are the most common choices in robustness. The results will be derived under the different scenarios for the missing probability to be considered later, i.e., when $p(\mathbf{x}, t)$ is assumed to be known and when it is consistently estimated.

4.1 Weak consistency of $\widehat{\theta}$

We begin by deriving consistency of the marginal location estimators when the missing probability is assumed to be known.

Proposition 4.1 *Let U_n and D_n be defined in* (5) *and* (6), *respectively. Assume that* $\widehat{\varsigma} \xrightarrow{p} \varsigma_0$.

(a) *Let $\widehat{\theta}^{(1)}_{\psi_2}$ be the solution of $U_n(p, \widehat{\varsigma}, a) = 0$, where the score function $\psi_2 : \mathbb{R} \to \mathbb{R}$ is a bounded and increasing function. Then, if ψ_2 is continuously differentiable with first derivative ψ_2' such that $u\psi_2'(u)$ is bounded, we have that $\widehat{\theta}^{(1)}_{\psi_2} \xrightarrow{p} \theta$, where θ is such that $E\psi_2((y - \theta)/\varsigma) = 0$ for all $\varsigma > 0$.*

(b) *Denote $\widehat{\theta}^{(1)}_{\rho_2} = \operatorname{argmin}_a D_n(p, \widehat{\varsigma}, a)$, where the loss function $\rho_2 : \mathbb{R} \to \mathbb{R}$ is a bounded rho-function. Then, if ρ_2 is continuously differentiable with first derivative ψ_2 such that $u\psi_2(u)$ is bounded, we have that $\widehat{\theta}^{(1)}_{\rho_2} \xrightarrow{p} \theta$, where θ is the unique solution of $\theta = \operatorname{argmin}_a E\rho_2((y - a)/\varsigma)$ for all $\varsigma > 0$.*

It is worth noticing that the differentiability assumption required to the score function ψ_2 or to the loss function ρ_2 is needed to deal with the scale parameter estimator. Using standard empirical process techniques, this condition can be replaced by the weaker condition $\lim_{\varsigma \to \varsigma_0} E \sup_{a \in \mathbb{R}} |\psi_2((y - a)/\varsigma) - \psi_2((y - a)/\varsigma_0)| = 0$ or $\lim_{\varsigma \to \varsigma_0} E \sup_{a \in \mathbb{R}} |\rho_2((y - a)/\varsigma) - \rho_2((y - a)/\varsigma_0)| = 0$, respectively.

In most real data applications, the missing probability is unknown and so $\widehat{\theta}^{(1)}_{\psi_2}$ or $\widehat{\theta}^{(1)}_{\rho_2}$ cannot be computed. In this situation, consistent estimators of $p(\mathbf{x}, t)$ need to be considered. The following result states the consistency of the marginal location estimators described in Sect. 2.2, under mild consistency assumptions on the missing probability estimators.

Proposition 4.2 *Let U_n and D_n be defined in* (5) *and* (6), *respectively. Assume that $\widehat{\varsigma} \xrightarrow{p} \varsigma_0$. Moreover, assume that the following conditions hold*:

(i) $\inf_{(\mathbf{x},t)} p(\mathbf{x},t) = A > 0$,

(ii) $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x},t) - p(\mathbf{x},t)| \xrightarrow{p} 0$.

Then,

(a) *If the score function $\psi_2 : \mathbb{R} \to \mathbb{R}$ is a bounded, increasing and continuously differentiable function with first derivative ψ_2' such that $u\psi_2'(u)$ is bounded, we have that the solution $\widehat{\theta}_{\psi_2}$ of $U_n(p_n, \widehat{\varsigma}, a) = 0$ satisfies that $\widehat{\theta}_{\psi_2} \xrightarrow{p} \theta$, where θ is such that $E\psi_2((y-\theta)/\varsigma) = 0$ for all $\varsigma > 0$.*

(b) *If the loss function $\rho_2 : \mathbb{R} \to \mathbb{R}$ is a bounded rho-function, continuously differentiable with first derivative ψ_2 such that $u\psi_2(u)$ is bounded, and $\widehat{\theta}_{\rho_2} = \operatorname{argmin}_a D_n(p_n, \widehat{\varsigma}, a)$, $\widehat{\theta}_{\rho_2} \xrightarrow{p} \theta$, where θ is the unique minimizer of $\zeta(a, \varsigma) = E\rho_2((y-a)/\varsigma)$ for all $\varsigma > 0$.*

Remark 4.1 Two situations arise when estimating the missing probability. The practitioner may choose a kernel estimator or a parametric approach based on previous information.

When considering a nonparametric estimator, the estimator is defined as $p_n(\mathbf{x},t) = p_{n,b_n}(\mathbf{x},t)$ where

$$p_{n,b_n}(\mathbf{x},t) = \sum_{i=1}^{n} K_1\left(\frac{\mathbf{w}_i - \mathbf{w}}{b_n}\right)\delta_i \left\{\sum_{j=1}^{n} K_1\left(\frac{\mathbf{w}_j - \mathbf{w}}{b_n}\right)\right\}^{-1}, \tag{7}$$

with $K_1 : \mathbb{R}^{d+1} \to \mathbb{R}$ a kernel function to be selected by the researcher, $\mathbf{w} = (\mathbf{x}^{\mathrm{T}}, t)^{\mathrm{T}}$ and b_n denoting the smoothing parameter. In this case, if $b_n \to 0$ and $nb_n^{d+1}/\log(n) \to +\infty$ and $K_1(\mathbf{w}) = \kappa(\|\mathbf{w}\|)$, where $\kappa : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ is a bounded variation function, analogous arguments to those considered in Pollard (1984, p. 35, Example 38) allow to show that $\sup_{(\mathbf{x},t)} |p_n(\mathbf{x},t) - p(\mathbf{x},t)| \xrightarrow{\text{a.s.}} 0$. The weaker assumption $nb_n^{d+1}/\log(b_n^{d+1}) \to +\infty$ can be required to obtain just convergence in probability modifying the proof of Theorem 37 in Pollard (1984).

On the other hand, in the parametric setting, assume that $p(\mathbf{x},t) = G(\mathbf{x},t,\boldsymbol{\lambda}_0)$, where $\boldsymbol{\lambda}_0 \in \mathbb{R}^q$ is an unknown parameter to be estimated and let $\widehat{\boldsymbol{\lambda}}$ be any consistent estimator of $\boldsymbol{\lambda}$, i.e., such that $\widehat{\boldsymbol{\lambda}} \xrightarrow{p} \boldsymbol{\lambda}_0$. Hence, the estimator of the missing probability is defined as $p_n(\mathbf{x},t) = p_{n,\widehat{\boldsymbol{\lambda}}}(\mathbf{x},t) = G(\mathbf{x},t,\widehat{\boldsymbol{\lambda}})$. Then, assumption (ii) in Proposition 4.2 is fulfilled if $G(\mathbf{x},t,\boldsymbol{\lambda})$ is equicontinuous in $\boldsymbol{\lambda}$ at $\boldsymbol{\lambda}_0$, i.e., for any $\epsilon > 0$ there exists $\delta > 0$ such that $|\boldsymbol{\lambda} - \boldsymbol{\lambda}_0| < \delta$ implies that $|G(\mathbf{x},t,\boldsymbol{\lambda}) - G(\mathbf{x},t,\boldsymbol{\lambda}_0)| < \epsilon$ for any $(\mathbf{x}^{\mathrm{T}}, t)$. For instance, when $G(\mathbf{x},t,\boldsymbol{\lambda})$ is a continuous function of all its arguments and $(\mathbf{x}^{\mathrm{T}}, t)$ lies in a compact set, this condition is fulfilled.

4.2 Asymptotic normality of $\widehat{\theta}$

Assumptions NM1–NM8 under which the estimators are asymptotically normally distributed are stated in the Appendix. From now on, we will denote $u = (y - \theta)/\varsigma_0$.

Theorem 4.1 *Let U_n be defined in (5). Assume that* NM1–NM3 *in the Appendix hold and that $\widehat{\varsigma} \xrightarrow{p} \varsigma_0$. Denote by $\widehat{\theta}^{(1)}$ the solution of $U_n(p, \widehat{\varsigma}, \theta) = 0$, i.e., the weighted simplified estimator assuming that the missing probability is known. If $\widehat{\theta}^{(1)} \xrightarrow{p} \theta$, where $E\psi_2((y-\theta)/\varsigma) = 0$ for all $\varsigma > 0$, we have that $\sqrt{n}(\widehat{\theta}^{(1)} - \theta) \xrightarrow{\mathcal{D}} N(0, \upsilon^{(1)})$, where $\upsilon^{(1)} = E(\psi_2^2(u)/p(\mathbf{x}, t))(E\psi_2'(u))^{-2} = \gamma^{(1)}(E\psi_2'(u))^{-2}$.*

Note that in this situation, the efficiency with respect to the classical simplified estimator, i.e., when $\psi_2(u) = u$, is not the efficiency of the location estimator when no missing data are present, since a factor $1/p(\mathbf{x}, t)$ depending on the missing probability appears in the numerator's expectation. Therefore, the efficiency of the estimators depends on the proportion of missing data appearing in the sample.

Theorem 4.2 *Let U_n be defined in (5). Assume that* NM1–NM5 *in the Appendix hold and that $\widehat{\varsigma} \xrightarrow{p} \varsigma_0$. Moreover, assume that $p(\mathbf{x}_i, t_i) = G(\mathbf{x}_i, t_i, \boldsymbol{\lambda}_0)$, where $\boldsymbol{\lambda}_0 \in \mathbb{R}^q$, and let $p_{n,\widehat{\boldsymbol{\lambda}}}(\mathbf{x}_i, t_i) = G(\mathbf{x}_i, t_i, \widehat{\boldsymbol{\lambda}})$, where $\widehat{\boldsymbol{\lambda}}$ is an estimator of $\boldsymbol{\lambda}$ such that $\widehat{\boldsymbol{\lambda}} \xrightarrow{p} \boldsymbol{\lambda}_0$. Denote by $\widehat{\theta}^{(2)}$ the solution of $U_n(p_{n,\widehat{\boldsymbol{\lambda}}}, \widehat{\varsigma}, \theta) = 0$, i.e., assuming a parametric model for the missing probability. If $\widehat{\theta}^{(2)} \xrightarrow{p} \theta$, where $E\psi_2((y-\theta)/\varsigma) = 0$ for all $\varsigma > 0$, we have that $\sqrt{n}(\widehat{\theta}^{(2)} - \theta) \xrightarrow{\mathcal{D}} N(0, \upsilon^{(2)})$, where $\upsilon^{(2)} = \gamma^{(2)}(E\psi_2'(u))^{-2}$ with*

$$\begin{aligned}
\gamma^{(2)} &= E\left[\frac{\delta}{G(\mathbf{x}, t, \boldsymbol{\lambda}_0)}\psi_2(u) - \boldsymbol{\eta}(\delta, \mathbf{x}, t)^{\mathrm{T}} E\left(\frac{G'(\mathbf{x}, t, \boldsymbol{\lambda}_0)}{G(\mathbf{x}, t, \boldsymbol{\lambda}_0)}\psi_2(u)\right)\right]^2 \\
&= E\frac{\psi_2^2(u)}{p(\mathbf{x}, t)} + E\left(\psi_2(u)\frac{G'(\mathbf{x}, t, \boldsymbol{\lambda}_0)}{G(\mathbf{x}, t, \boldsymbol{\lambda}_0)}\right)^{\mathrm{T}} \left\{\boldsymbol{\Sigma}\, E\left(\psi_2(u)\frac{G'(\mathbf{x}, t, \boldsymbol{\lambda}_0)}{G(\mathbf{x}, t, \boldsymbol{\lambda}_0)}\right)\right. \\
&\quad \left. - 2E\left[\frac{\delta\psi_2(u)\boldsymbol{\eta}(\delta, \mathbf{x}, t)}{G(\mathbf{x}, t, \boldsymbol{\lambda}_0)}\right]\right\}
\end{aligned}$$

and $\boldsymbol{\eta}$ and $\boldsymbol{\Sigma}$ given in NM5.

We will now study the asymptotic distribution of the weighted simplified estimator when the missing probability is estimated using a kernel estimator as in (7) where $K_1 : \mathbb{R}^{d+1} \to \mathbb{R}$ is a kernel function, $\mathbf{w} = (\mathbf{x}^{\mathrm{T}}, t)^{\mathrm{T}}$ and b_n denotes the smoothing parameter.

Theorem 4.3 *Let U_n be defined in (5). Assume that* NM1–NM3 *and* NM6–NM8 *in the Appendix hold and that $\widehat{\varsigma} \xrightarrow{p} \varsigma_0$. Let $p_{n,b_n}(\mathbf{x}_i, t_i)$ be the kernel estimator defined in (7). Denote by $\widehat{\theta}^{(3)}$ the solution of $U_n(p_{n,b_n}, \widehat{\varsigma}, \theta) = 0$, i.e., using the nonparametric estimator of the missing probability, $p_{n,b_n}(\mathbf{x}, t)$, defined in (7). If $\widehat{\theta}^{(3)} \xrightarrow{p} \theta$,* *where* $E\psi_2((y-\theta)/\varsigma)=0$ *for all* $\varsigma>0$, *we have that* $\sqrt{n}(\widehat{\theta}^{(3)}-\theta)\xrightarrow{\mathcal{D}} N(0,\upsilon^{(3)})$ *where* $\upsilon^{(3)}=\gamma^{(3)}(E\psi_2'(u))^{-2}$ *and*

$$\gamma^{(3)}=E\left(\frac{\delta}{p(\mathbf{x},t)}\psi_2\left(\frac{y-\theta}{\varsigma_0}\right)-\frac{(\delta-p(\mathbf{x},t))}{p(\mathbf{x},t)}r(\mathbf{x},t)\right)^2,$$

with $r(\mathbf{x},t)=E(\psi_2(u)|\mathbf{x},t)$.

Remark 4.2 Using that

$$E\left(\frac{\delta}{p(\mathbf{x},t)}\big(\delta-p(\mathbf{x},t)\big)|(y,\mathbf{x},t)\right)=E\left(\frac{\delta}{p(\mathbf{x},t)}\big(\delta-p(\mathbf{x},t)\big)|(\mathbf{x},t)\right)=1-p(\mathbf{x},t),\tag{8}$$

and after some algebra, we get that

$$\gamma^{(3)}=E\frac{\psi_2^2(u)}{p(\mathbf{x},t)}-E\left(\frac{(1-p(\mathbf{x},t))}{p(\mathbf{x},t)}r^2(\mathbf{x},t)\right).$$

Hence, for any missing model we have that $\upsilon^{(3)}\le\upsilon^{(1)}$ and so, the marginal location estimator $\widehat{\theta}^{(3)}$ computed estimating the missing probability through a kernel estimator is more efficient than $\widehat{\theta}^{(1)}$. Note that both estimators have equal variance if and only if $E((1-p(\mathbf{x},t))r^2(\mathbf{x},t)/p(\mathbf{x},t))=\mathbf{0}$, i.e., if and only if there are no missing observations, since $E(\psi_2(u)|\mathbf{x},t)=0$ a.e. holds only if $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+g(t)$ is constant, which is a situation to be discarded in practice. This gain of efficiency was discussed by several authors when covariates are missing and is related to the sample adjustment obtained through the kernel estimator of the missing probability (see Sect. 5 for further comments).

4.3 Some comments under a logistic model for the missing probability

Denote by $F_{\mathrm{L}}(s)=(1+e^{-s})^{-1}$ the logistic distribution function and let us assume that the missing probability is given by the logistic model, i.e., that $p(\mathbf{x},t)=F_{\mathrm{L}}(\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda}_0)$ and $G(\mathbf{x},t,\boldsymbol{\lambda})=F_{\mathrm{L}}(\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda})$ where $\mathbf{v}=(1,\mathbf{x}^{\mathrm{T}},t)^{\mathrm{T}}$. Hence, $G'(\mathbf{x},t,\boldsymbol{\lambda})=F_{\mathrm{L}}(\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda})[1-F_{\mathrm{L}}(\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda})]\mathbf{v}$ and so,

- $G'(\mathbf{x},t,\boldsymbol{\lambda}_0)=p(\mathbf{x},t)[1-p(\mathbf{x},t)]\mathbf{v}$,
- $E(G'(\mathbf{x},t,\boldsymbol{\lambda}_0)\psi_2(u)/p(\mathbf{x},t))=E((1-p(\mathbf{x},t)\psi_2(u)\mathbf{v})$.

It is worth noticing that in this situation, NM4(b) and NM6 hold. Besides, using that $\{\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda},\boldsymbol{\lambda}\in\mathbb{R}^{d+2}\}$ is a finite-dimensional class of functions and so, it has polynomial discrimination, and the permanence properties stated in van der Vaart and Wellner (1996), we get that $\mathcal{G}=\{G(\mathbf{x},t,\boldsymbol{\lambda}),\boldsymbol{\lambda}\in\mathbb{R}^{d+2}\}$ has finite-entropy entailing that NM4(a) is fulfilled. Moreover, (c) and (d) in NM4 are satisfied if

$$E\big(\big(1-F_{\mathrm{L}}\big(\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda}\big)\big)|v_j|\big)<\infty\quad E\big(\big(1+e^{\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda}_0}\big)|v_jv_\ell|\big)<\infty\quad\text{for } 1\le j,\ell\le q.\tag{9}$$

Note that, when the missing probability only depends on t, i.e., when $\mathbf{v}=(1,t)^{\mathrm{T}}$, (9) is fulfilled since we are assuming that t belongs to a compact set. In this case, NM3 also holds.

In order to analyze the relation between the asymptotic variances of $\widehat{\theta}^{(1)}$, $\widehat{\theta}^{(2)}$ and $\widehat{\theta}^{(3)}$, we shall consider two situations for estimating the parameter $\boldsymbol{\lambda}$. In the first scenario, we will assume that the parameters are estimated using the maximum likelihood estimators, while in the second one, we will assume that they are estimated using a robust procedure. Note that under mild conditions, both estimators admit a Bahadur expansion as required in NM5.

The calculations to be done include, in particular, the classical estimators, i.e., those corresponding to $\psi_2(t) = t$, for which, up to our knowledge, there are no results regarding the theoretical comparison of the asymptotic variances of the marginal location estimator when the missing probability is known and when it is estimated parametrically, using the maximum likelihood estimator, or even, as discussed in Remark 4.2, nonparametrically using a kernel approach. Note also that, when $\psi_2(t) = t$, NM1 and NM2 are clearly satisfied.

(i) Let us thus assume that $\widehat{\boldsymbol{\lambda}}$ is the maximum likelihood estimator. This estimator can be considered instead of a robust one, such as that defined in Croux and Haesbroeck (2003), if we suspect that no outliers are present in the covariates $\mathbf{x}$ or if we know that $p(\mathbf{x}, t)$ only depends on t where no outliers appear, i.e., if in the above model, $\mathbf{v} = (1, t)^{\mathrm{T}}$. This last situation is also included in the sequel just by taking into account the new expression for $\mathbf{v}$. As mentioned above, the maximum likelihood estimator admits a Bahadur expansion with $\boldsymbol{\eta}(\delta, \mathbf{x}, t) = \mathbf{A}_1^{-1}(\delta - p(\mathbf{x}, t))\mathbf{v}$, where $\mathbf{A}_1 = Ep(\mathbf{x}, t)(1 - p(\mathbf{x}, t))\mathbf{v}\mathbf{v}^{\mathrm{T}}$ which implies that the matrix $\boldsymbol{\Sigma}$ defined in NM5 equals $\mathbf{A}_1^{-1}$. We want to show that $\gamma^{(3)} \le \gamma^{(2)} \le \gamma^{(1)}$, i.e., even if a logistic model is assumed, the best performance of the location estimators is attained when estimating the missing probability nonparametrically.

For the parametric situation, the asymptotic variance is given by $\gamma^{(2)} = E(\psi_2^2(u)/p(\mathbf{x}, t)) + \nu^{(2)}$ with

$$\begin{aligned}\nu^{(2)} &= E\big(\big(1 - p(\mathbf{x}, t)\big)\psi_2(u)\mathbf{v}^{\mathrm{T}}\big)\mathbf{A}_1^{-1}\Bigg\{E\big(\big(1 - p(\mathbf{x}, t)\big)\psi_2(u)\mathbf{v}\big) \\ &\qquad - 2E\left[\psi_2(u)\frac{\delta(\delta - p(\mathbf{x}, t))}{p(\mathbf{x}, t)}\mathbf{v}\right]\Bigg\} \\ &= -E\big(\big(1 - p(\mathbf{x}, t)\big)\psi_2(u)\mathbf{v}^{\mathrm{T}}\big)\mathbf{A}_1^{-1}E\big(\big(1 - p(\mathbf{x}, t)\big)\psi_2(u)\mathbf{v}\big), \qquad (10)\end{aligned}$$

where we have used (8). Hence, $\nu^{(2)} \le 0$ which entails that $\gamma^{(2)} \le \gamma^{(1)}$ and equality holds if and only if $E((1 - p(\mathbf{x}, t))\psi_2(u)\mathbf{v}) = \mathbf{0}$ that happens obviously if there are no missing observations.

In this case, we can also compare the asymptotic variances of the marginal location estimator when the missing probability is estimated parametrically or using a kernel estimator, since from Remark 4.2 we already know that a nonparametric approach is better than assuming the probability to be known.

We want thus to show that $\gamma^{(3)} \le \gamma^{(2)}$ and hence $\upsilon^{(3)} \le \upsilon^{(2)}$, which means that the nonparametric estimator of the missing probability gives whenever it is possible to compute the smallest asymptotic variance. Let us recall that

$\gamma^{(3)} = E(\psi_2^2(u)/p(\mathbf{x},t)) + \nu^{(3)}$ with $\nu^{(3)} = -E((1-p(\mathbf{x},t))\, r^2(\mathbf{x},t)/p(\mathbf{x},t))$. On the other hand, $\gamma^{(2)} = E(\psi_2^2(u)/p(\mathbf{x},t)) + \nu^{(2)}$ with $\nu^{(2)}$ given in (10), where $\mathbf{A}_1 = Ep(\mathbf{x},t)(1-p(\mathbf{x},t))\mathbf{v}\mathbf{v}^{\mathrm{T}}$. Note that $E((1-p(\mathbf{x},t))\psi_2(u)\mathbf{v}^{\mathrm{T}}) = E((1-p(\mathbf{x},t))r(\mathbf{x},t)\mathbf{v}^{\mathrm{T}})$. Hence, in order to compare the asymptotic variances and using the expressions given above, we only need to compare the quantities $a^{(2)}$ and $a^{(3)}$, where

$$a^{(2)} = -\nu^{(2)} = E\big(\big(1-p(\mathbf{x},t)\big)r(\mathbf{x},t)\mathbf{v}^{\mathrm{T}}\big)\mathbf{A}_1^{-1}E\big(\big(1-p(\mathbf{x},t)\big)r(\mathbf{x},t)\mathbf{v}\big),$$

$$a^{(3)} = E\left(\frac{(1-p(\mathbf{x},t))}{p(\mathbf{x},t)}\, r^2(\mathbf{x},t)\right).$$

Clearly, if $a^{(3)} = 0$ then $a^{(2)} = 0$, so we can assume that $a^{(3)} > 0$. Let $\mathbf{A}_1 = \mathbf{C}_1\mathbf{C}_1^{\mathrm{T}}$ and denote $\mathbf{z} = \mathbf{C}_1^{-1}(\delta - p(\mathbf{x},t))\mathbf{v}$ and $\xi = (\delta - p(\mathbf{x},t))p(\mathbf{x},t)^{-1}r(\mathbf{x},t)$. Then, $E(\mathbf{z}) = 0$, $E(\xi) = 0$, $E(\mathbf{z}\mathbf{z}^{\mathrm{T}}) = \mathbf{I}$, $a^{(2)} = \|E(\xi\mathbf{z})\|^2$, while $a^{(3)} = E(\xi^2) = Var(\xi)$. If we denote $\boldsymbol{\rho} = E(\xi\mathbf{z})$ and $\boldsymbol{\Sigma}^{\star} = E(\mathbf{s}\mathbf{s}^{\mathrm{T}})$ with $\mathbf{s} = (\xi, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$, we have that $\boldsymbol{\Sigma}^{\star} = \begin{pmatrix} a^{(3)} & \boldsymbol{\rho}^{\mathrm{T}} \\ \boldsymbol{\rho} & \mathbf{I} \end{pmatrix}$ is a non-negative definite matrix. Note that since $\det(\boldsymbol{\Sigma}^{\star}) = a^{(3)} \det(\mathbf{I} - (1/a^{(3)})\,\boldsymbol{\rho}\boldsymbol{\rho}^{\mathrm{T}}) \geq 0$, the eigenvalue $1 - (1/a^{(3)})\,\boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\rho}$ of the matrix $\mathbf{I} - (1/a^{(3)})\,\boldsymbol{\rho}\boldsymbol{\rho}^{\mathrm{T}}$ is non-negative and so, $a^{(2)} = \|\boldsymbol{\rho}\|^2 \leq a^{(3)}$, as desired.

(ii) In some situations, the parameters of the logistic model need to be estimated robustly, for instance, if we suspect that high leverage points in the carriers $\mathbf{x}$ are present. We can carry on the robust estimation using, for instance, a weighted maximum likelihood estimator or the estimator defined in Croux and Haesbroeck (2003), i.e., $\widehat{\boldsymbol{\lambda}} = \operatorname{argmin}_{\boldsymbol{\lambda}} \sum_{i=1}^{n} w(\mathbf{x}_i)\varphi(\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\lambda};\delta_i)$ where $\varphi(s;0) = \varphi(-s;1)$, $\varphi(s;0) = \rho(-\ln(1-F_{\mathrm{L}}(s))) + C(F_{\mathrm{L}}(s)) + C(1-F_{\mathrm{L}}(s)) - C(1)$ with ρ a score function and $C(s) = \int_0^s \rho'(-\ln u)du$ the correction term ensuring Fisher-consistency. The weighted maximum likelihood estimator corresponds to the choice $\rho(s) = s$, while the estimators considered in Bianco and Yohai (1996) use a bounded ρ function. Then, using the results in Bianco and Martínez (2009), we have that $\boldsymbol{\eta}(\delta,\mathbf{x},t) = -\mathbf{A}_{1,\mathrm{R}}^{-1}w(\mathbf{x})\Psi(\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda}_0;\delta)\mathbf{v}$, where $\Psi(s;0) = \partial\varphi(s;0)/\partial s$, $\Psi(s;1) = -\Psi(-s;0)$ and

$$\mathbf{A}_{1,\mathrm{R}} = E\left\{ w(\mathbf{x})\frac{\partial^2}{\partial s^2}\varphi(s;\delta)\bigg|_{s=\mathbf{v}^{\mathrm{T}}\boldsymbol{\lambda}_0} \mathbf{v}\mathbf{v}^{\mathrm{T}} \right\}.$$

Straightforward calculations lead to

$$E\left(\frac{\delta\boldsymbol{\eta}(\delta,\mathbf{x},t)}{p(\mathbf{x},t)}\psi_2(u)\right) = \mathbf{A}_{1,\mathrm{R}}^{-1}E\big(\psi_2(u)w(\mathbf{x})\big(1-p(\mathbf{x},t)\big)D(\mathbf{x},t)\mathbf{v}\big),$$

$$\boldsymbol{\Sigma} = \mathbf{A}_{1,\mathrm{R}}^{-1}E\big(w^2(\mathbf{x})\big(1-p(\mathbf{x},t)\big)p(\mathbf{x},t)D^2(\mathbf{x},t)\mathbf{v}\mathbf{v}^{\mathrm{T}}\big)\mathbf{A}_{1,\mathrm{R}}^{-1},$$

where $D(\mathbf{x},t) = (1-p(\mathbf{x},t))C'(p(\mathbf{x},t)) + p(\mathbf{x},t)C'(1-p(\mathbf{x},t))$. Therefore, $\gamma^{(2)} = E(\psi_2^2(u)/p(\mathbf{x},t)) + \nu^{(2)}$ with

$$\nu^{(2)} = E\big(\big(1-p(\mathbf{x},t)\big)\psi_2(u)\mathbf{v}^{\mathrm{T}}\big)\big\{\boldsymbol{\Sigma}E\big(\big(1-p(\mathbf{x},t)\big)\psi_2(u)\mathbf{v}\big) - 2\mathbf{A}_{1,\mathrm{R}}^{-1}E\big(\psi_2(u)w(\mathbf{x})\big(1-p(\mathbf{x},t)\big)D(\mathbf{x},t)\mathbf{v}\big)\big\}.$$

In particular, if $w(\mathbf{x}) = w^2(\mathbf{x})$, which corresponds to a $0-1$ weight function and $\rho(s) = s$, i.e., when considering the weighted maximum likelihood, we have that $\mathbf{A}_{1,\mathrm{R}} = E\{w(\mathbf{x})p(\mathbf{x},t)(1-p(\mathbf{x},t))\mathbf{v}\mathbf{v}^{\mathrm{T}}\}$, $D(\mathbf{x},t) \equiv 1$ and so, $E(w^2(\mathbf{x})(1-p(\mathbf{x},t))p(\mathbf{x},t)D^2(\mathbf{x},t)\mathbf{v}\mathbf{v}^{\mathrm{T}}) = \mathbf{A}_{1,\mathrm{R}}$ implying that $\nu^{(2)} = \mathbf{b}^{\mathrm{T}}\mathbf{A}_{1,\mathrm{R}}^{-1}\mathbf{b} - 2\mathbf{b}^{\mathrm{T}}\mathbf{A}_{1,\mathrm{R}}^{-1}\mathbf{b}_w$, where $\mathbf{b} = E((1-p(\mathbf{x},t))\psi_2(u)\mathbf{v})$ and $\mathbf{b}_w = E(\psi_2(u)w(\mathbf{x})(1-p(\mathbf{x},t))\mathbf{v})$. Depending on the choice of the weight function w, i.e., on the tuning constant selected to cut-off outliers, the inner product $\mathbf{b}^{\mathrm{T}}\mathbf{A}_{1,\mathrm{R}}^{-1}\mathbf{b}_w$ can be much smaller than the squared norm $\mathbf{b}^{\mathrm{T}}\mathbf{A}_{1,\mathrm{R}}^{-1}\mathbf{b}$, leading to a positive value of $\nu^{(2)}$. In this situation, the variance of the robust marginal location estimator $\widehat{\theta}^{(2)}$ will be larger than that of the estimator $\widehat{\theta}^{(1)}$ computed with the true missing probability. This fact is consistent with the simulation results obtained in Bianco et al. (2010a) and opposite to the conclusions obtained when the parameters of the missing probability are estimated using the classical maximum likelihood estimator, leading to a larger loss of efficiency when robust estimators are used.

5 Concluding remarks

Under a partially linear model when there are missing observations in the response variable but the covariates $(\mathbf{x}^{\mathrm{T}}, t)$ are totally observed, the classical procedures fail to give reliable estimations when it can be suspected that anomalous observations are present in the sample. Robust procedures to estimate the regression parameter and the marginal location y were introduced in Bianco et al. (2010a). In this paper, the consistency and asymptotic distribution of the regression parameter estimators are obtained. Moreover, we show that the weighted simplified M-estimators, $\widehat{\theta}$, related to increasing or redescending score functions ψ_2, lead to weakly consistent estimators. Besides, we derive their asymptotic distribution, when the missing probability is assumed to be known or when it is estimated either parametrically or using a kernel approach. Different asymptotic variances are obtained in each situation.

The obtained theoretical results validate the numerical outcomes observed in the simulation study performed in Bianco et al. (2010a), since they allow to show that estimating nonparametrically the missing probability reduces the variance of the marginal estimator either when the probability is known or, under a logistic missing model, when it is estimated parametrically using the maximum likelihood estimator. This counter-intuitive phenomenon was also observed by several authors, such as Pierce (1982), Rosenbaum (1987), Robins et al. (1994, 1995), Wang et al. (1998) and the references given therein. When the covariates are missing, Wang et al. (1997) discussed the gain of efficiency of the estimators of θ via adjustment of the missing probability. A heuristic argument justifying this behavior for general parameter estimation with missing covariates was given in Robins et al. (1994). The same arguments can be applied for missing responses. When the missing probability is modeled parametrically and the unknown quantities are estimated using maximum likelihood estimators, the gain of efficiency is related to the linear expansion given in the Appendix together with the joint asymptotic distribution of $\sum_{i=1}^{n}\delta_i p^{-1}(\mathbf{w}_i)\psi_2(u_i)/\sqrt{n}$ and of $\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)$, and so the optimality arguments used in Pierce (1982) can be considered to explain the effect of replacing estimators for the true parameters.

On the other hand, when the parameters are estimated robustly using a weighted maximum likelihood method with weight function w, the robust estimators of the marginal location $\widehat{\theta}^{(2)}$ may have a higher loss of efficiency. To be more precise, depending on the tuning constant selected to cut-off outliers, the variance of the robust marginal location estimator $\widehat{\theta}^{(2)}$ may be larger than that of the estimator $\widehat{\theta}^{(1)}$ computed with the true missing probability and so, larger than that of $\widehat{\theta}^{(3)}$, the estimator based on a kernel approach. In this sense, we recommend using a smooth estimator of the missing probabilities instead of a parametric one, if the dimension of the covariates and the number of observations allow to compute the kernel estimator.

Acknowledgements This work began while Ana Bianco and Graciela Boente 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 0821 from ANPCYT and X-018 from the Universidad de Buenos Aires at Buenos Aires, Argentina and also by the DGICYT Spanish Grant MTM2008-03010 (European FEDER support included). We also wish to thank two anonymous referees and the Associate Editor for valuable comments which led to an improved version of the original paper.

Appendix

6.1 Proof of the asymptotic normality of the regression estimates

For the sake of simplicity, we denote by ψ' and ψ'' the first and second derivatives of ψ. Moreover, let $\mathbf{z} = \mathbf{z}(\boldsymbol{\beta}_0) = \mathbf{x} + (\partial g_{\boldsymbol{\beta}}(t)/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$, $\mathbf{z}_i = \mathbf{z}_i(\boldsymbol{\beta}_0) = \mathbf{x}_i + (\partial g_{\boldsymbol{\beta}}(t_i)/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ and

$$\widehat{\gamma}(\boldsymbol{\beta}, \tau) = \widehat{g}_{\boldsymbol{\beta}}(\tau) - g_{\boldsymbol{\beta}}(\tau), \qquad \widehat{\gamma}_0(\tau) = \widehat{\gamma}(\boldsymbol{\beta}_0, \tau), \tag{11}$$

$$\widehat{\mathrm{v}}_j(\boldsymbol{\beta}, \tau) = \frac{\partial \widehat{\gamma}(\boldsymbol{\beta}, \tau)}{\partial \beta_j}, \qquad \widehat{\mathrm{v}}_{j,0}(\tau) = \widehat{\mathrm{v}}_j(\boldsymbol{\beta}_0, \tau). \tag{12}$$

We list the conditions needed for the asymptotic normality of the regression parameter estimators, followed by general comments on those conditions. The first condition is on the preliminary estimate of $g_{\boldsymbol{\beta}}(\tau)$, while the others ones concern the score functions and the underlying model distributions.

N1. (a) The functions $\widehat{g}_{\boldsymbol{\beta}}(\tau)$ and $g_{\boldsymbol{\beta}}(\tau)$ are continuously differentiable with respect to $(\boldsymbol{\beta}, \tau)$ and twice continuously differentiable with respect to $\boldsymbol{\beta}$ such that $(\partial^2 g_{\boldsymbol{\beta}}(\tau)/\partial \beta_j \partial \beta_\ell)|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is bounded. Furthermore, for any $1 \le j, \ell \le p$, $\partial^2 g_{\boldsymbol{\beta}}(\tau)/\partial \beta_j \partial \beta_\ell$ satisfies the following equicontinuity condition:

$$\forall \epsilon > 0, \exists \delta > 0\colon \quad |\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0| < \delta$$
$$\Rightarrow \quad \left\| \frac{\partial^2}{\partial \beta_j \partial \beta_\ell} g_{\boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1} - \frac{\partial^2}{\partial \beta_j \partial \beta_\ell} g_{\boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\|_\infty < \epsilon.$$

(b) $\|\widehat{g}_{\widehat{\boldsymbol{\beta}}} - g_0\|_\infty \xrightarrow{p} 0$, for any consistent estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$.

(c) For each $\tau \in \mathcal{T}$ and $\boldsymbol{\beta}$, $\widehat{\gamma}(\boldsymbol{\beta}, \tau) \xrightarrow{p} 0$. Moreover, $n^{1/4}\|\widehat{\gamma}_0\|_\infty \xrightarrow{p} 0$ and $n^{1/4}\|\widehat{\mathrm{v}}_{j,0}\|_\infty \xrightarrow{p} 0$ for all $1 \leq j \leq p$.

(d) There exists a neighborhood of $\boldsymbol{\beta}_0$ with closure $\mathcal{K}$ such that for any $1 \leq j, \ell \leq p$, $\sup_{\boldsymbol{\beta}\in\mathcal{K}}(\|\widehat{\mathrm{v}}_j(\boldsymbol{\beta}, \cdot)\|_\infty + \|\partial\widehat{\mathrm{v}}_j(\boldsymbol{\beta}, \cdot)/\partial\beta_\ell\|_\infty) \xrightarrow{p} 0$.

(e) $\|\partial\widehat{\gamma}_0/\partial\tau\|_\infty + \|\partial\widehat{\mathrm{v}}_{j,0}/\partial\tau\|_\infty \xrightarrow{p} 0$ for any $1 \leq j \leq p$.

N2. The functions υ and $\Upsilon(\mathbf{x}) = \mathbf{x}\upsilon(\mathbf{x})$ are bounded and continuous. The function $\psi = \rho'$ is an odd, bounded and twice continuously differentiable function with bounded derivatives ψ' and ψ'', such that $\varphi_1(s) = s\psi'(s)$ and $\varphi_2(s) = s\psi''(s)$ are bounded. Moreover, the function ψ_1 is a bounded and continuously differentiable function with bounded derivative ψ_1'.

N3. The matrix $\mathbf{A} = E\psi'(\epsilon)\ E(\upsilon(\mathbf{x})p(\mathbf{x}, t)\mathbf{z}(\boldsymbol{\beta}_0)\mathbf{z}(\boldsymbol{\beta}_0)^{\mathrm{T}})$ is non-singular.

N4. The matrix $\boldsymbol{\Sigma} = E\psi^2(\epsilon)\ E(\upsilon^2(\mathbf{x})p(\mathbf{x}, t)\mathbf{z}(\boldsymbol{\beta}_0)\mathbf{z}(\boldsymbol{\beta}_0)^{\mathrm{T}})$ is positive definite.

N5. $E(\psi_1'(\epsilon)) \neq 0$ and $E(\psi'(\epsilon)) \neq 0$.

N6. $E(p(\mathbf{x}, t)\upsilon(\mathbf{x})\|\mathbf{z}(\boldsymbol{\beta}_0)\|^2) < \infty$.

Remark 6.1 The convergence requirements in N1 are analogous to those required in condition (7) in Severini and Staniswalis (1994, p. 510) and are needed in order to obtain the desired rate of convergence for the regression estimates. In particular, condition N1 (b) follows from the continuity of $g_{\boldsymbol{\beta}}(\tau) = g(\boldsymbol{\beta}, \tau)$ with respect to $(\boldsymbol{\beta}, \tau)$ and Theorem 3.1 that leads to $\sup_{\boldsymbol{\beta}\in\mathcal{K}} \|\widehat{g}_{\boldsymbol{\beta}} - g_{\boldsymbol{\beta}}\|_\infty \xrightarrow{\text{a.s.}} 0$. Conditions N1 (a) and (d) entail that for any consistent estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$, we have

$$\max_{1\leq j\leq p} \left\| \left.\frac{\partial \widehat{g}_{\boldsymbol{\beta}}}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} - \left.\frac{\partial g_{\boldsymbol{\beta}}}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\|_\infty \xrightarrow{p} 0 \quad \text{and}$$

$$\max_{1\leq j,\ell\leq p} \left\| \left.\frac{\partial^2 \widehat{g}_{\boldsymbol{\beta}}}{\partial \beta_j \partial\beta_\ell}\right|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} - \left.\frac{\partial^2 g_{\boldsymbol{\beta}}}{\partial \beta_j\partial\beta_\ell}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\|_\infty \xrightarrow{p} 0.$$

Remark 6.2 When the kernel K is continuously differentiable with bounded derivative K' and with bounded variation, the uniform convergence required in N1 (d) and (e) can be derived through analogous arguments to those considered in Theorem 3.1 by using that

$$\frac{\partial}{\partial\tau}\widehat{g}_{\boldsymbol{\beta}}(\tau) = \frac{(nh_n{}^2)^{-1}\sum_{i=1}^n K'(\frac{\tau-t_i}{h_n})\delta_i\psi_1(\frac{y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-\widehat{g}_{\boldsymbol{\beta}}(\tau)}{\widehat{s}_{\boldsymbol{\beta}}})\upsilon(\mathbf{x}_i)}{(nh_n)^{-1}\sum_{i=1}^n K(\frac{\tau-t_i}{h_n})\delta_i\psi_1'(\frac{y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-\widehat{g}_{\boldsymbol{\beta}}(\tau)}{\widehat{s}_{\boldsymbol{\beta}}})\upsilon(\mathbf{x}_i)},$$

$$\frac{\partial}{\partial\beta_j}\widehat{g}_{\boldsymbol{\beta}}(\tau) = -\frac{\sum_{i=1}^n K(\frac{\tau-t_i}{h_n})[\delta_i\psi_1'(\frac{y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-\widehat{g}_{\boldsymbol{\beta}}(\tau)}{\widehat{s}_{\boldsymbol{\beta}}})\upsilon(\mathbf{x}_i)](x_{ij} + \frac{y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-\widehat{g}_{\boldsymbol{\beta}}(\tau)}{\widehat{s}_{\boldsymbol{\beta}}}\frac{\partial}{\partial\beta_j}\widehat{s}_{\boldsymbol{\beta}})}{\sum_{i=1}^n K(\frac{\tau-t_i}{h_n})\delta_i\psi_1'(\frac{y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-\widehat{g}_{\boldsymbol{\beta}}(\tau)}{\widehat{s}_{\boldsymbol{\beta}}})\upsilon(\mathbf{x}_i)}$$

and requiring that $u\psi_1'(u)$ is a bounded function and

$$\sup_{\tau\in\mathcal{T}} E\left(\sup_{\boldsymbol{\beta}\in\mathcal{K},\sigma\in\mathcal{K}_\sigma} \left| \psi_1'\left(\frac{y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-g_{\boldsymbol{\beta}}(\tau)}{\sigma}\right)\|\mathbf{x}\| \right| t=\tau \right) < \infty,$$

$$\sup_{\tau\in\mathcal{T}} E\left(\sup_{\boldsymbol{\beta}\in\mathcal{K},\sigma\in\mathcal{K}_\sigma} \left| \psi_1''\left(\frac{y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-g_{\boldsymbol{\beta}}(\tau)}{\sigma}\right)\|\mathbf{x}\| \right| t=\tau \right) < \infty,$$

$$\inf_{\substack{\boldsymbol{\beta}\in\mathcal{K},\sigma\in\mathcal{K}_\sigma\\ \tau\in\mathcal{T}}} \left| E\left(\psi_1'\left(\frac{y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-g_{\boldsymbol{\beta}}(\tau)}{\sigma}\right)|t=\tau \right)\right| > 0.$$

The uniform convergence rates required in N1 (c) are fulfilled when $\widehat{g}_{\boldsymbol{\beta}}$ is defined in Step 1 using kernel weights and a rate-optimal bandwidth is used for the kernel.

Remark 6.3 Note that if $P(\upsilon(\mathbf{x})>0)=1$ and $E\psi'(\epsilon)\neq 0$, N3 holds, i.e., $\mathbf{A}$ will be non-singular unless $P(\mathbf{a}^{\mathrm{T}}\mathbf{z}(\boldsymbol{\beta}_0)=0)=1$, for some $\mathbf{a}\in\mathbb{R}^d$, that is, unless there is a linear combination of $\mathbf{x}$ which can be completely determined by t, in which case the model is fully nonparametric instead of partly linear. The condition $E\psi'(\epsilon)\neq 0$ is a standard requirement in robust regression in order to get root-n estimators of $\boldsymbol{\beta}$. Again, if N4 is fulfilled the columns of $\mathbf{x}+(\partial g_{\boldsymbol{\beta}}(t)/\partial\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ will not be collinear. It is necessary not to allow $\mathbf{x}$ to be predicted by t to get root-n regression estimates. N5 is a standard condition in robustness in order to get root-n estimators. It is worth noticing that N5 entails that

$$E\left[\left(\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}g_{\boldsymbol{\beta}}(\tau)\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)\upsilon(\mathbf{x})p(\mathbf{x},\tau)|t=\tau\right]=0. \tag{13}$$

Effectively, since $g_{\boldsymbol{\beta}}(\tau)$ satisfies (2) for each τ differentiating with respect to $\boldsymbol{\beta}$, we get

$$E\left[\delta\psi_1'\left(\frac{y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-g_{\boldsymbol{\beta}}(\tau)}{\sigma_{\boldsymbol{\beta}}}\right)\left(\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}g_{\boldsymbol{\beta}}(\tau)+\frac{y-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}-g_{\boldsymbol{\beta}}(\tau)}{\sigma_{\boldsymbol{\beta}}}\frac{\partial}{\partial\boldsymbol{\beta}}\sigma_{\boldsymbol{\beta}}\right)\upsilon(\mathbf{x})\bigg|t=\tau\right]$$
$$=0\quad\forall\boldsymbol{\beta}.$$

Thus, specializing at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ and using that the errors ϵ are independent of $(\mathbf{x},t)$, we obtain

$$\begin{aligned}0&=E\big(\psi_1'(\epsilon)\big)E\left[p(\mathbf{x},t)\left(\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}g_{\boldsymbol{\beta}}(\tau)\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)\upsilon(\mathbf{x})|t=\tau\right]\\ &\quad+E\big(\epsilon\psi_1'(\epsilon)\big)E\big[p(\mathbf{x},t)\upsilon(\mathbf{x})|t=\tau\big]\frac{\partial}{\partial\boldsymbol{\beta}}\sigma_{\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\\ &=E\big(\psi_1'(\epsilon)\big)E\left[p(\mathbf{x},t)\left(\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}g_{\boldsymbol{\beta}}(\tau)\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)\upsilon(\mathbf{x})|t=\tau\right],\end{aligned}$$

where the last equality holds since ψ_1' is an even function and ϵ has a symmetric distribution. Thus, (13) holds.

Assumption N6 is used to ensure the consistency of the estimates of $\mathbf{A}$ based on preliminary estimates of the regression parameter $\boldsymbol{\beta}$ and of the functions $g_{\boldsymbol{\beta}}$. Besides, note that the continuous differentiability of $g_{\boldsymbol{\beta}}(\tau)$ with respect to $(\boldsymbol{\beta}, \tau)$ required in N1 implies that $(\partial g_{\boldsymbol{\beta}}(t)/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is bounded and so, $E(\|(\partial g_{\boldsymbol{\beta}}(t)/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\|^2) < \infty$. Thus, under N1, N6 holds if $E(\upsilon(\mathbf{x})\|\mathbf{x}\|^2) < \infty$. When considering $\upsilon \equiv 1$, this is a second moment condition on the regression carriers. On the other hand, when choosing a weight function to control leverage points, this condition is fulfilled without any moment condition, taking for instance a Tukey's biweight function since it has compact support.

Remark 6.4 As mentioned in Sect. 2.1, Bianco and Boente (2004) introduced robust estimation under a partly linear model when there are no missing observations. Therein, conditions analogous to N2, N3 and N6 were considered to derive the asymptotic distribution of the estimators of $\boldsymbol{\beta}$. Besides, a rate requirement to the nonparametric estimators analogous to N1 (c) was also needed together with some smoothness of the estimators as in N1 (a), (d) and (e). The main difference between our assumption N3 and their assumption N2 is that the missing probability is introduced. On the other hand, N4 ensures that the limiting distribution is non-degenerate and this was also a requirement in the complete setting. Note that N5 is needed to obtain (13) which allows to ensure that $\widehat{g}_{\boldsymbol{\beta}}$ and its first derivative with respect to $\boldsymbol{\beta}$ can be replaced by the true functions. This assumption is analogous to condition N4 in Bianco and Boente (2004).

Lemma 6.1 *Let* $(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i)$ *be independent observations satisfying* (1). *Assume that* t_i *are random variables with distribution on a compact set* $\mathcal{T}$ *and that* N1–N3 *and* N6 *hold. Let* $\widetilde{\boldsymbol{\beta}}$ *be such that* $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ *and* $\widehat{\mathbf{z}}_i(\widetilde{\boldsymbol{\beta}}) = \mathbf{x}_i + (\partial \widehat{g}_{\boldsymbol{\beta}}(t_i)/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}$. *Then,* $\mathbf{A}_n \xrightarrow{p} \mathbf{A}$ *where* $\mathbf{A}$ *is given in* N3 *and*

$$\mathbf{A}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\psi'\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \widehat{g}_{\widetilde{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\widehat{\mathbf{z}}_i(\widetilde{\boldsymbol{\beta}})\widehat{\mathbf{z}}_i(\widetilde{\boldsymbol{\beta}})^{\mathrm{T}}\right.$$

$$\left. + \psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \widehat{g}_{\widetilde{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}\widehat{g}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\right)\delta_i\,\upsilon(\mathbf{x}_i).$$

Proof Note that $\mathbf{A}_n$ can be written as $\mathbf{A}_n = \sum_{j=1}^{6}\mathbf{A}_n^{(j)}$ with

$$\mathbf{A}_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\,\psi'\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - g_0(t_i)}{\widehat{\sigma}}\right)\mathbf{z}_i\,\mathbf{z}_i^{\mathrm{T}}\upsilon(\mathbf{x}_i),$$

$$\mathbf{A}_n^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\,\psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - g_0(t_i)}{\widehat{\sigma}}\right)\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}g_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}^{\mathrm{T}}\upsilon(\mathbf{x}_i),$$

$$\mathbf{A}_n^{(3)} = \frac{1}{\widehat{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\delta_i\,\psi''\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \xi_{i,1}}{\widehat{\sigma}}\right)\widehat{\mathrm{w}}_0(t_i)\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\upsilon(\mathbf{x}_i),$$

$$\mathbf{A}_n^{(4)} = \frac{1}{\widehat{\sigma}}\frac{1}{n}\sum_{i=1}^{n} \delta_i \psi'\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \xi_{i,2}}{\widehat{\sigma}}\right)\widehat{\mathrm{w}}_0(t_i)\, \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} g_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}^{\mathrm{T}} \upsilon(\mathbf{x}_i),$$

$$\mathbf{A}_n^{(5)} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \psi'\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \widehat{g}_{\widetilde{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\left[\widehat{\mathbf{w}}(t_i)\mathbf{z}_i^{\mathrm{T}} + \mathbf{z}_i\widehat{\mathbf{w}}(t_i)^{\mathrm{T}} + \widehat{\mathbf{w}}(t_i)\,\widehat{\mathbf{w}}(t_i)^{\mathrm{T}}\right]\upsilon(\mathbf{x}_i),$$

$$\mathbf{A}_n^{(6)} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \widehat{g}_{\widetilde{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\widehat{\mathbf{V}}(t_i)^{\mathrm{T}}\upsilon(\mathbf{x}_i),$$

where $\xi_{i,1}$ and $\xi_{i,2}$ are intermediate points and $\mathbf{z}_i = \mathbf{z}_i(\boldsymbol{\beta}_0)$, $\widehat{\mathrm{w}}_0(t) = \widehat{g}_{\widehat{\boldsymbol{\beta}}}(t) - g_0(t)$ and

$$\widehat{\mathbf{w}}(t) = \frac{\partial}{\partial\boldsymbol{\beta}}\widehat{g}_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} - \frac{\partial}{\partial\boldsymbol{\beta}} g_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0},$$

$$\widehat{\mathbf{V}}(t) = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\widehat{g}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} - \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} g_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

Using N1 (a), (b) and (d), N6, the boundedness of ψ, ψ', ψ'', υ and Υ and the fact that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, it follows easily that $\mathbf{A}_n^{(j)} \xrightarrow{p} 0$ for $3 \leq j \leq 6$. From N6, the consistency of $\widetilde{\boldsymbol{\beta}}$ and the continuity of ψ and ψ', we get easily that $\mathbf{A}_n^{(1)} + \mathbf{A}_n^{(2)} \xrightarrow{p} \mathbf{A}$. □

Proof of Theorem 3.3 Let $\widehat{\boldsymbol{\beta}}$ be a solution of $H_n^{(1)}(\boldsymbol{\beta}) = 0$ defined in (4) and denote $\widehat{\mathbf{z}}_i(\boldsymbol{\beta}) = \mathbf{x}_i + (\partial\widehat{g}_{\boldsymbol{\beta}}(t_i)/\partial\boldsymbol{\beta})|_{\boldsymbol{\beta}}$. Using the Taylor expansion of order one, we get

$$0 = \sum_{i=1}^{n} \delta_i \psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - \widehat{g}_{\boldsymbol{\beta}_0}(t_i)}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i)\widehat{\mathbf{z}}_i(\boldsymbol{\beta}_0) - \frac{1}{\widehat{\sigma}} n\mathbf{A}_n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where

$$\begin{aligned}\mathbf{A}_n &= -\frac{\widehat{\sigma}}{n}\sum_{i=1}^{n} \delta_i \frac{\partial}{\partial\boldsymbol{\beta}}\left\{\psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \widehat{g}_{\boldsymbol{\beta}}(t_i)}{\widehat{\sigma}}\right)\widehat{\mathbf{z}}_i(\boldsymbol{\beta})\right\}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} \upsilon(\mathbf{x}_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\psi'\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \widehat{g}_{\widetilde{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\widehat{\mathbf{z}}_i(\widetilde{\boldsymbol{\beta}})\widehat{\mathbf{z}}_i(\widetilde{\boldsymbol{\beta}})^{\mathrm{T}}\right.\\ &\quad \left. - \psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} - \widehat{g}_{\widetilde{\boldsymbol{\beta}}}(t_i)}{\widehat{\sigma}}\right)\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\widehat{g}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}^{\mathrm{T}}\right)\delta_i\upsilon(\mathbf{x}_i),\end{aligned}$$

with $\widetilde{\boldsymbol{\beta}}$ an intermediate point between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. From Lemma 6.1 we have that $\mathbf{A}_n \xrightarrow{p} \mathbf{A}$, where $\mathbf{A}$ is defined in N3. Therefore, in order to obtain the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$, it will be sufficient to derive the asymptotic behavior of

$\widehat{L}_n = n^{-1/2}\sum_{i=1}^n \delta_i \psi((y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - \widehat{g}_{\boldsymbol{\beta}_0}(t_i))/\widehat{\sigma})\upsilon(\mathbf{x}_i)\widehat{\mathbf{z}}_i(\boldsymbol{\beta}_0)$. Let

$$L_n = n^{-1/2}\sum_{i=1}^n \delta_i \psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - g_{\boldsymbol{\beta}_0}(t_i)}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i)\mathbf{z}_i(\boldsymbol{\beta}_0)$$
$$= n^{-1/2}\sum_{i=1}^n \delta_i \psi\left(\frac{\epsilon_i \sigma_0}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i)\mathbf{z}_i(\boldsymbol{\beta}_0),$$

since $g_{\boldsymbol{\beta}_0} = g_0$. Using that ψ is odd and the errors have a symmetric distribution and are independent of the carriers, we have that $E[\psi(\epsilon_i\sigma_0/\sigma)|(\mathbf{x}_i, t_i)] = E\psi(\epsilon_i\sigma_0/\sigma) = 0$, for all σ. Then, the consistency of $\widehat{\sigma}$ and standard tightness arguments entail that L_n is asymptotically normally distributed with covariance matrix $\boldsymbol{\Sigma}$. Therefore, it remains to show that $L_n - \widehat{L}_n \xrightarrow{p} 0$.

We have the following expansion: $\widehat{L}_n - L_n = -\widehat{\sigma}^{-2}L_n^1 + \widehat{\sigma}^{-1}L_n^2 - \widehat{\sigma}^{-1}L_n^3 + \widehat{\sigma}^{-2}L_n^4$, with

$$L_n^1 = n^{-1/2}\widehat{\sigma}\sum_{i=1}^n \delta_i \psi'\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - g_{\boldsymbol{\beta}_0}(t_i)}{\widehat{\sigma}}\right)\mathbf{z}_i(\boldsymbol{\beta}_0)\upsilon(\mathbf{x}_i)\widehat{\gamma}_0(t_i),$$

$$L_n^2 = n^{-1/2}\widehat{\sigma}\sum_{i=1}^n \delta_i \psi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - g_{\boldsymbol{\beta}_0}(t_i)}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i)\widehat{\mathbf{v}}_0(t_i),$$

$$L_n^3 = n^{-1}\sum_{i=1}^n \delta_i \psi'\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - g_{\boldsymbol{\beta}_0}(t_i)}{\widehat{\sigma}}\right)\upsilon(\mathbf{x}_i)\left(n^{1/4}\widehat{\mathbf{v}}_0(t_i)\right)\left(n^{1/4}\widehat{\gamma}_0(t_i)\right),$$

$$L_n^4 = (2n)^{-1}\sum_{i=1}^n \delta_i \psi''\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - \xi_i(t_i)}{\widehat{\sigma}}\right)\mathbf{z}_i(\boldsymbol{\beta}_0)\upsilon(\mathbf{x}_i)\left(n^{1/4}\widehat{\gamma}_0(t_i)\right)^2,$$

where $\widehat{\gamma}_0(\tau) = \widehat{g}_{\boldsymbol{\beta}_0}(\tau) - g_0(\tau)$, $\widehat{\mathbf{v}}_0(\tau) = (\widehat{\mathrm{v}}_{1,0}(\tau), \ldots, \widehat{\mathrm{v}}_{p,0}(\tau))^{\mathrm{T}} = \partial\widehat{\gamma}(\boldsymbol{\beta}, \tau)/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is defined in (12), $\widehat{\gamma}$ is defined in (11) and $\xi(t_i)$ an intermediate point between $\widehat{g}_{\boldsymbol{\beta}_0}(t_i)$ and $g_0(t_i)$. It is easy to see that $L_n^3 \xrightarrow{p} 0$ and $L_n^4 \xrightarrow{p} 0$ follow from N1 (c) and N2.

To complete the proof, it remains show that $L_n^j \xrightarrow{p} 0$ for $j = 1, 2$ which follow from N1 (c)–(e) and (13), using similar arguments to those considered in Bianco and Boente (2004). Details can be seen in Bianco et al. (2010b). □

6.2 Proof of the weak consistency of the marginal estimators

Proof of Proposition 4.1 Let $\Gamma(u)$ stand for either ψ_2 or ρ_2. Then, Γ is bounded and continuously differentiable and $\Lambda(u) = u\Gamma'(u)$ is bounded. Therefore,

$$C_n(a, \widehat{\varsigma}, \varsigma) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{p(\mathbf{x}_i, t_i)}\Gamma\left(\frac{y_i - a}{\widehat{\varsigma}}\right) - \frac{\delta_i}{p(\mathbf{x}_i, t_i)}\Gamma\left(\frac{y_i - a}{\varsigma}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{p(\mathbf{x}_i, t_i)} \Lambda\left(\frac{y_i - a}{\widehat{\xi}}\right) \widehat{\xi} \left(\frac{1}{\widehat{\varsigma}} - \frac{1}{\varsigma}\right)$$

where $\widehat{\xi}$ is an intermediate point between $\widehat{\varsigma}$ and ς. This implies that $|C_n(a, \widehat{\varsigma}, \varsigma)| \leq \|\Lambda\|_\infty |\widehat{\xi}||\widehat{\varsigma}^{-1} - \varsigma^{-1}| L_n$ with $L_n = (1/n)\sum_{i=1}^{n} \delta_i / p(\mathbf{x}_i, t_i)$ and so, using that $\widehat{\varsigma} \xrightarrow{p} \varsigma_0$, we get that $\sup_a |C_n(a, \widehat{\varsigma}, \varsigma_0| \xrightarrow{p} 0$. Hence, $U_n(p, \varsigma_0, \widehat{\theta}^{(1)}_{\psi_2}) \xrightarrow{p} 0$ and $D_n(p, \varsigma_0, \widehat{\theta}^{(1)}_{\rho_2}) - \inf_a D_n(p, \varsigma_0, a) \xrightarrow{p} 0$.

(a) The consistency of $\widehat{\theta}^{(1)}_{\psi_2}$ follows now as in Theorem 10.5 from Maronna et al. (2006) using the fact that ψ_2 is increasing implies that θ is the unique solution of $E\psi_2((y - a)/\varsigma) = 0$, for all ς.

(b) Using Theorem 2.2 in Huber (1981), we obtain that $\widehat{\theta}^{(1)}_{\rho_2} \xrightarrow{p} \theta$. Note that in this case, assumption (A-5) in Huber (1981) is fulfilled with $b(\theta) = \|\rho_2\|_\infty$.

□

Proof of Proposition 4.2 Using (i) and (ii) and the boundedness of ψ_2 and ρ_2, it is easy to see that $\sup_a |U_n(p, \widehat{\varsigma}, a) - U_n(p_n, \widehat{\varsigma}, a)| \xrightarrow{p} 0$ and $\sup_a |D_n(p, \widehat{\varsigma}, a) - D_n(p_n, \widehat{\varsigma}, a)| \xrightarrow{p} 0$. Therefore, $\widehat{\theta}_{\psi_2}$ and $\widehat{\theta}_{\rho_2}$ satisfy that $U_n(p, \widehat{\varsigma}, \widehat{\theta}_{\psi_2}) \xrightarrow{p} 0$ and $D_n(p, \widehat{\varsigma}, \widehat{\theta}_{\rho_2}) - \inf_a D_n(p, \widehat{\varsigma}, a) \xrightarrow{p} 0$ and the proof follows as in the proof of Proposition 4.1 □

6.3 Proof of the asymptotic distribution of the marginal estimators

From now on, when estimating the marginal location, we will assume, without loss of generality, that the marginal scale ς_0 is known and so we will replace $\widehat{\varsigma}$ by ς_0. Recall that $u = (y - \theta)/\varsigma_0$ and denote $u_i = (y_i - \theta)/\varsigma_0$. The following assumptions are needed to obtain the asymptotic distribution of the weighted simplified marginal M-estimator, under two of the scenarios to be considered, i.e., when the missing probability is assumed to be known or when it is estimated parametrically.

NM1. The function ψ_2 is twice continuously differentiable with bounded derivatives.

NM2. $A(\psi_2) = E[\delta\psi_2'(u)/p(\mathbf{x}, t)] = E\psi_2'(u) \neq 0$.

NM3. $\inf_{(\mathbf{x},t)} p(\mathbf{x}, t) = \iota(p) > 0$.

NM4. The missing probability $p(\mathbf{x}, t) = G(\mathbf{x}, t, \boldsymbol{\lambda}_0)$, $\boldsymbol{\lambda}_0 \in \mathbb{R}^q$, is such that:

- (a) The family of functions $\mathcal{G} = \{G(\mathbf{x}, t, \boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathbb{R}^q\}$ has finite entropy.
- (b) $G(\mathbf{x}, t, \boldsymbol{\lambda})$ is twice continuously differentiable with respect to $\boldsymbol{\lambda}$. We will denote by $G'(\mathbf{x}, t, \boldsymbol{\lambda})$ and $G''(\mathbf{x}, t, \boldsymbol{\lambda})$ the gradient and Hessian matrix of $G(\mathbf{x}, t, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$.
- (c) $E(|G_j'(\mathbf{x}, t, \boldsymbol{\lambda}_0)|\psi_2'(u)/p(\mathbf{x}, t)) < \infty$ for $1 \leq j \leq q$.
- (d) For some $\Lambda > 0$, $E(\sup_{\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_0| < \Lambda} |G_{j\ell}''(\mathbf{x}, t, \boldsymbol{\lambda})\psi_2'(u)|/p(\mathbf{x}, t)) < \infty$ for $1 \leq j, \ell \leq q$.

NM5. $\widehat{\boldsymbol{\lambda}}$ admits a Bahadur expansion given by $\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) = (1/\sqrt{n}) \times \sum_{i=1}^{n} \boldsymbol{\eta}(\delta_i, \mathbf{x}_i, t_i) + o_p(1)$ where $E\boldsymbol{\eta}(\delta_i, \mathbf{x}_i, t_i) = \mathbf{0}$ and $E\|\boldsymbol{\eta}(\delta_i, \mathbf{x}_i, t_i)\|^2 < \infty$. We will denote by $\boldsymbol{\Sigma} = E\boldsymbol{\eta}(\delta, \mathbf{x}, t)\boldsymbol{\eta}(\delta, \mathbf{x}, t)^{\mathrm{T}}$ the asymptotic covariance matrix of $\widehat{\boldsymbol{\lambda}}$.

Remark 6.5 NM1 and NM2 are standard conditions in robustness; in particular, NM2 is required in order to get root-n estimators. Assumption NM3 is a common assumption in the literature meaning that at any value of the covariate, response variables are observed. Assumption NM4 holds in most parametric situations such as the logistic missing model. Maximum likelihood estimators usually fulfill NM5. Moreover, many classes of robust estimators such as the weighted Bianco and Yohai estimators considered by Croux and Haesbroeck (2003) for a logistic model admit a Bahadur expansion; in fact, their asymptotic normality is usually derived by showing that NM5 holds.

Proof of Theorem 4.1 Since NM1 holds, the Taylor expansion of order two leads to

$$0=\sqrt{n}U_n\big(p,\varsigma_0,\widehat{\theta}^{(1)}\big)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{p(\mathbf{x}_i,t_i)}\psi_2(u_i)-\sqrt{n}\big(\widehat{\theta}^{(1)}-\theta\big)A_n(\psi_2),$$

where

$$A_n(\psi_2)=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{p(\mathbf{x}_i,t_i)}\psi_2'(u_i)+\frac{1}{2}\big(\widehat{\theta}^{(1)}-\theta\big)\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{p(\mathbf{x}_i,t_i)}\psi_2''\left(\frac{y_i-\xi_n}{\varsigma_0}\right)$$

and ξ_n is an intermediate point between $\widehat{\theta}^{(1)}$ and θ. Note that NM3 ensures that $\gamma^{(1)}=E(\delta\psi_2^2(u)/p(\mathbf{x},t)^2)=E\psi_2^2(u)/p(\mathbf{x},t)<\infty$ and so, the Central Limit Theorem implies that $\sqrt{n}\,U_n(p,\varsigma_0,\theta)\xrightarrow{\mathcal{D}}N(0,\gamma^{(1)})$. Using that $A_n(\psi_2)\xrightarrow{p}A(\psi_2)$ and the fact that from NM2 $A(\psi_2)\neq 0$, the proof follows. □

Proof of Theorem 4.2 As in the proof of Theorem 4.1, using the Taylor expansion of order two, we get that $0=(1/\sqrt{n})\sum_{i=1}^{n}(\delta_i/p_{n,\widehat{\boldsymbol{\lambda}}}(\mathbf{x}_i,t_i))\psi_2(u_i)-\sqrt{n}(\widehat{\theta}^{(2)}-\theta)A_n^{(2)}(\psi_2)$, where

$$A_n^{(2)}(\psi_2)=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{p_{n,\widehat{\boldsymbol{\lambda}}}(\mathbf{x}_i,t_i)}\psi_2'(u_i)+\frac{1}{2}(\widehat{\theta}^{(2)}-\theta)\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{p_{n,\widehat{\boldsymbol{\lambda}}}(\mathbf{x}_i,t_i)}\psi_2''\left(\frac{y_i-\xi_n}{\varsigma_0}\right)$$

and ξ_n is an intermediate point between $\widehat{\theta}^{(2)}$ and θ. Using NM3, it follows that $A_n^{(2)}(\psi_2)\xrightarrow{p}A(\psi_2)$.

Therefore, it is sufficient to show that $B_n=(1/\sqrt{n})\sum_{i=1}^{n}(\delta_i/p_{n,\widehat{\boldsymbol{\lambda}}}(\mathbf{x}_i,t_i))\psi_2(u_i)\xrightarrow{\mathcal{D}}N(0,\upsilon^2)$. Note that

$$B_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{p(\mathbf{x}_i,t_i)}\psi_2(u_i)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{p(\mathbf{x}_i,t_i)}{p_{n,\widehat{\boldsymbol{\lambda}}}(\mathbf{x}_i,t_i)}-1\right)\frac{\delta_i}{p(\mathbf{x}_i,t_i)}\psi_2(u_i).$$

Denote

$$R_n(\boldsymbol{\lambda})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{G(\mathbf{x}_i,t_i,\boldsymbol{\lambda}_0)}{G(\mathbf{x}_i,t_i,\boldsymbol{\lambda})}-1\right)\frac{\delta_i}{p(\mathbf{x}_i,t_i)}\big[\psi_2(u_i)-r(\mathbf{x}_i,t_i)\big],$$

where $r(\mathbf{x},t) = E\psi_2(u)|(\mathbf{x},t)$. Then, using NM4(a), the fact that $\widehat{\boldsymbol{\lambda}} \xrightarrow{p} \boldsymbol{\lambda}_0$ and standard empirical processes arguments, we get easily that $R_n(\widehat{\boldsymbol{\lambda}}) \xrightarrow{p} 0$ and so, $B_n = B_{1,n} + B_{2,n} + B_{3,n} + o_p(1)$ where

$$B_{1,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{p(\mathbf{x}_i, t_i)} \psi_2(u_i),$$

$$B_{2,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\boldsymbol{\lambda}_0 - \widehat{\boldsymbol{\lambda}})^{\mathrm{T}} \frac{G'(\mathbf{x}_i, t_i, \boldsymbol{\lambda}_0)}{G(\mathbf{x}_i, t_i, \widehat{\boldsymbol{\lambda}})} \frac{\delta_i}{p(\mathbf{x}_i, t_i)} r(\mathbf{x}_i, t_i),$$

$$B_{3,n} = \frac{1}{2} (\boldsymbol{\lambda}_0 - \widehat{\boldsymbol{\lambda}})^{\mathrm{T}} \frac{1}{n} \sum_{i=1}^{n} G''(\mathbf{x}_i, t_i, \boldsymbol{\xi}) \frac{1}{G(\mathbf{x}_i, t_i, \widehat{\boldsymbol{\lambda}})} \frac{\delta_i}{p(\mathbf{x}_i, t_i)} r(\mathbf{x}_i, t_i) \sqrt{n} (\boldsymbol{\lambda}_0 - \widehat{\boldsymbol{\lambda}}).$$

The Bahadur expansion given in NM5 implies that $\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) = O_p(1)$, thus, using NM4(d) we obtain that $B_{3,n} \xrightarrow{p} 0$. Therefore, since $B_{2,n} = -\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)^{\mathrm{T}} E((G'(\mathbf{x}, t, \boldsymbol{\lambda}_0)/G(\mathbf{x}, t, \boldsymbol{\lambda}_0)) r(\mathbf{x}, t)) + o_p(1)$, to derive the asymptotic distribution of B_n it suffices to study the asymptotic behavior of

$$\begin{aligned} & B_{1,n} + B_{2,n} + o_p(1) \\ &\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{p(\mathbf{x}_i, t_i)} \psi_2(u_i) - \sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)^{\mathrm{T}} E\left(\frac{G'(\mathbf{x}, t, \boldsymbol{\lambda}_0)}{G(\mathbf{x}, t, \boldsymbol{\lambda}_0)} r(\mathbf{x}, t) \right) + o_p(1) \\ &\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\delta_i}{p(\mathbf{x}_i, t_i)} \psi_2(u_i) - \boldsymbol{\eta}(\delta_i, \mathbf{x}_i, t_i)^{\mathrm{T}} E\left(\frac{G'(\mathbf{x}, t, \boldsymbol{\lambda}_0)}{G(\mathbf{x}, t, \boldsymbol{\lambda}_0)} r(\mathbf{x}, t) \right) \right] + o_p(1), \end{aligned}$$

where the last equality follows from NM5. The proof follows now from the Central Limit Theorem. □

To derive the asymptotic distribution of $\widehat{\theta}^{(3)}$, we will need the following additional assumptions. For the sake of simplicity, we will denote $\mathbf{w} = (\mathbf{x}^{\mathrm{T}}, t)^{\mathrm{T}} \in \mathbb{R}^{d_1}$, with $d_1 = d + 1$. We state the assumptions in terms of $\mathbf{w}$ and d since, in some situations, due to prior knowledge, the researcher may use a kernel estimator depending on some and not all the covariates; in this case, $\mathbf{w}$ plays the role of the covariates to be considered and d_1 may be lower than $d + 1$.

NM6. The missing probability $p(\mathbf{w})$ is a smooth function of $\mathbf{w}$, rth continuously differentiable.

NM7. The bandwidth b_n satisfies that $\rho_n^2 = \{n b_n^{2r} + (n b_n^{2d_1})^{-1}\} \to 0$.

NM8. The kernel $K_1 : \mathbb{R}^{d_1} \to \mathbb{R}$ is bounded, has compact support and $\int K_1(\mathbf{u})\, d\mathbf{u} > 0$, $\int u_j^m K_1(\mathbf{u})\, d\mathbf{u} = 0$, for $1 \le j \le d_1$, $1 \le m \le r - 1$, $\int u_j^r K_1(\mathbf{u})\, d\mathbf{u} > 0$, for $1 \le j \le d_1$, and $\int u_j^2 K_1(\mathbf{u})\, d\mathbf{u} > 0$.

For the sake of simplicity, we will assume that $\int K_1(\mathbf{u})\, d\mathbf{u} = 1$.

Remark 6.6 Condition NM7 is related to the bias term appearing when replacing the true missing probability by a kernel estimator and so, to the degree of smoothness required to the missing probability, see the discussion in Wang et al. (1997, p. 514). It states that more smoothness is needed as the dimension of the covariates $\mathbf{w}$ increases. For instance, if $b_n = n^{-\alpha}$, NM7 states that $1/(2r) < \alpha < 1/(2d_1)$ and so the missing probability needs to be at least $(d_1 + 1)$th continuously differentiable.

Condition NM8 is a standard condition for kernel estimators and was required also in Wang et al. (1997, p. 514). The conditions $\int u_j^m K_1(\mathbf{u})\, d\mathbf{u} = 0$, for $1 \le j \le d_1$, $1 \le m \le r-1$, $\int u_j^r K_1(\mathbf{u})\, d\mathbf{u} > 0$, for $1 \le j \le d_1$, state that K_1 is an rth order kernel and they allow to expand the bias term of the kernel estimator in terms of the rth derivatives of the missing probability.

Proof of Theorem 4.3 As in the proof of Theorem 4.2, using the Taylor expansion of order two, we get that $0 = (1/\sqrt{n}) \sum_{i=1}^n (\delta_i / p_{n,b_n}(\mathbf{x}_i, t_i)) \psi_2((y_i - \theta)/\varsigma_0) - \sqrt{n}(\widehat{\theta}^{(3)} - \theta) A_n^{(3)}(\psi_2)$, where

$$A_n^{(3)}(\psi_2) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{p_{n,b_n}(\mathbf{x}_i, t_i)} \psi_2'(u_i) + \frac{1}{2}(\widehat{\theta}^{(3)} - \theta) \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{p_{n,b_n}(\mathbf{x}_i, t_i)} \psi_2''\left(\frac{y_i - \xi_n}{\varsigma_0}\right)$$

and ξ_n is an intermediate point between $\widehat{\theta}^{(3)}$ and θ. Using NM3, it is easy to see that $A_n^{(3)}(\psi_2) \xrightarrow{p} A(\psi_2)$. Therefore, it is sufficient to show that $B_n = (1/\sqrt{n}) \sum_{i=1}^n (\delta_i / p_{n,b_n}(\mathbf{x}_i, t_i)) \psi_2(u_i) \xrightarrow{\mathcal{D}} N(0, \gamma^{(3)})$. Note that

$$B_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{p(\mathbf{w}_i)} \psi_2(u_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{p(\mathbf{w}_i)}{p_{n,b_n}(\mathbf{w}_i)} - 1 \right) \frac{\delta_i}{p(\mathbf{w}_i)} \psi_2(u_i) = B_n^{(1)} + B_n^{(2)},$$

where $p_{n,b_n}(\mathbf{w})$ is defined in (7).

Denote $f_n(\mathbf{w}) = \sum_{i=1}^n K_1((\mathbf{w}_i - \mathbf{w})/b_n)/(n b_n^{d_1})$. Arguing as in Wang et al. (1997) and using standard U-statistics arguments, we get that $B_{1,n} = O_p(\rho_n)$ and $B_{2,n} = (1/\sqrt{n}) \sum_{j=1}^n r(\mathbf{w}_j)(\delta_j - p(\mathbf{w}_j))/p(\mathbf{w}_j) + O_p(\rho_n)$, where ρ_n is defined in NM7, see Bianco et al. (2010b) for details. Hence, we obtain that

$$B_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{\delta_j}{p(\mathbf{w}_j)} \psi_2\left(\frac{y_j - \theta}{\varsigma_0}\right) - \frac{(\delta_j - p(\mathbf{w}_j))}{p(\mathbf{w}_j)} r(\mathbf{w}_j) + O_p(\rho_n)$$

and so, the Central Limit Theorem entails that $B_n \xrightarrow{\mathcal{D}} N(0, \gamma^{(3)})$ concluding the proof. □

References

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

Bianco A, Martínez E (2009) Robust testing in the logistic regression model. Comput Stat Data Anal 53:4095–4105

Bianco A, Yohai V (1996) Robust estimation in the logistic regression model. Lecture notes in statistics, vol 109. Springer, New York, pp 17–34

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

Bianco A, Boente G, González-Manteiga W, Pérez-González A (2010b). Asymptotic behavior of robust estimators in partially linear models with missing responses: the effect of estimating the missing probability on the simplified marginal estimators. Report 10-01. Reports in statistics and operations research. Universidad de Santiago de Compostela. http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/REPORTS/10-01.pdf

Boente G, He X, Zhou J (2006) Robust estimates in generalized partially linear models. Ann Stat 34:2856–2878

Boente G, González-Manteiga W, Pérez-González A (2009) Robust nonparametric estimation with missing data. J Stat Plan Inference 139:571–592

Croux C, Haesbroeck G (2003) Implementing the Bianco and Yohai estimator for logistic regression. Comput Stat Data Anal 44:273–295

Härdle W, Liang H, Gao J (2000) Partially linear models. Springer, Berlin

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

Pierce D (1982) The asymptotic effect of substituting estimators for parameters in certain types of statistics. Ann Stat 10:475–478

Pollard D (1984) Convergence of stochastic processes. Springer, New York

Robins J, Rotnitzky A, Zhao L (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–866

Robins J, Rotnitzky A, Zhao L (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90:106–121

Rosenbaum P (1987) Model-based direct adjustment. J Am Stat Assoc 82:387–394

Severini T, Staniswalis J (1994) Quasi-likelihood estimation in semiparametric models. J Am Stat Assoc 89:501–511

van der Vaart A, Wellner J (1996) Weak convergence and empirical processes. With applications to statistics. Springer, New York

Wang Q, Sun Z (2007) Estimation in partially linear models with missing responses at random. J Multivar Anal 98:1470–1493

Wang C, Wang S, Zhao L, Ou S (1997) Weighted semiparametric estimation in regression analysis regression with missing covariates data. J Am Stat Assoc 92:512–525

Wang C, Wang S, Gutierrez R, Carroll R (1998) Local linear regression for generalized linear models with missing data. Ann Stat 26:1028–1050

Wang Q, Linton O, Härdle W (2004) Semiparametric regression analysis with missing response at random. J Am Stat Assoc 99(466):334–345