Robust Estimates in Generalized Partially Linear Models [*]

Graciela Boente
Universidad de Buenos Aires and CONICET, Argentina

Xuming He
University of Illinois at Urbana–Champaign, USA

and

Jianhui Zhou
University of Virginia, USA

Abstract

In this paper, we introduce a family of robust estimates for the parametric and nonparametric components under a generalized partially linear model, where the data are modeled by $y_i | (\mathbf{x}_i, t_i) \sim F(\cdot, \mu_i)$ with $\mu_i = H(\eta(t_i) + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})$, for some known distribution function F and link function H. It is shown that the estimates of $\boldsymbol{\beta}$ are root–n consistent and asymptotically normal. Through a Monte Carlo study the performance of these estimators is compared with that of the classical ones.

AMS Subject Classification 1990: Primary 62F35, Secondary 62H25.
Key words and phrases: Kernel Weights; Partly linear models; Rate of convergence; Robust estimation; Smoothing.

Abbreviated Title: Robust Semiparametric Regression

[*] This research was partially supported by Grants PICT # 03-00000-006277 from ANPCYT, PID 5505 from CONICET, X-094 from the Universidad de Buenos Aires, 13900-6 from the Fundación Antorchas at Buenos Aires, Argentina, and the U.S. National Science Foundation Grant DMS-0102411. The research began while the first author was granted with a Guggenheim fellowship.

1 Introduction

Semiparametric models contain both a parametric and a nonparametric component. Sometimes the nonparametric component plays the role of a nuisance parameter. A lot of research has been done on estimators of the parametric component in a general framework, aiming to obtain asymptotically efficient estimators. The aim of this paper is to consider semiparametric versions of the generalized linear models where the response y is to be predicted by covariates $(\mathbf{x}, t)$, where $\mathbf{x} \in I\!R^p$ and $t \in \mathcal{T} \subset I\!R$. It will be assumed that the conditional distribution of $y|(\mathbf{x}, t)$ belongs to the canonical exponential family $\exp\left[y\theta(\mathbf{x}, t) - B\left(\theta(\mathbf{x}, t)\right) + C(y)\right]$, for known functions B and C. Then, $\mu\left(\mathbf{x}, t\right) = \mathrm{E}\left(y|(\mathbf{x}, t)\right) = B'\left(\theta(\mathbf{x}, t)\right)$, with B' as the derivative of B. In generalized linear models (McCullagh and Nelder, 1989), which is a popular approach for modeling a wide variety of data, it is often assumed that the mean is modeled linearly through a known inverse link function, g, i.e.,

$$g(\mu\left(\mathbf{x}, t\right)) = \beta_0 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \alpha t\ .$$

For instance, an ordinary logistic regression model assumes that the observations $(y_i, \mathbf{x}_i, t_i)$ are such that the response variables are independent binomial variables $y_i|(\mathbf{x}_i, t_i) \sim Bi(1, p_i)$ whose success probabilities depend on the explanatory variables through the relation $g(p_i) = \beta_0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \alpha t_i$, with $g(u) = \ln(u/(1-u))$.

The influence function of the classical estimates based on the quasi–likelihood is unbounded. Large deviations of the response from its mean, as measured by the Pearson residuals, or outlying points in the covariate space can have large influence on the estimators. Those outliers or potential outliers for the generalized linear regression model are to be detected and controlled by robust procedures such as those considered by Stefanski, Carroll and Ruppert (1986), Künsch, Stefanski and Carroll (1989), Bianco and Yohai (1995) and Cantoni and Ronchetti (2001a).

In some applications, the linear model is insufficient to explain the relationship between the response variable and its associated covariates. A natural generalization, which suffers from the *curse of dimensionality*, is to model the mean nonparametrically in the covariates. An alternative strategy is to allow most predictors to be modeled linearly while one or a small number of predictors enter the model nonparametrically. This is the approach we will follow, so that the relationship will be given by the semiparametric generalized partially linear model

$$\mu\left(\mathbf{x},t\right) = H\left(\eta(t) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right) \tag{1}$$

where $H = g^{-1}$ is a known link function, $\boldsymbol{\beta} \in I\!\!R^p$ is an unknown parameter and η is an unknown continuous function.

Severini and Wong (1992) introduced the concept of generalized profile likelihood, which was later applied to this model by Severini and Staniswalis (1994). In this method, the nonparametric component is viewed as a function of the parametric component, and $\sqrt{n}-$consistent estimates for the parametric component can be obtained when the usual optimal rate for the smoothing parameter is used. Such estimates do not deal with outlying observations. In a semiparametric setting, outliers can have a devastating effect, since the extreme points can easily affect the scale and the shape of the function estimate of η, leading to possibly wrong conclusions on β. The basic ideas from robust smoothing and from robust regression estimation have been adapted to partly linear regression models where $H(t) = t$; we refer to Gao and Shi (1997), He, Zhu and Fung (2002) and Bianco and Boente (2004). A robust generalized estimating equations approach, for generalized partially linear models with clustered data, using regression splines and Pearson residuals is given in He, Fung and Zhu (2005).

In Section 2 of the present paper, we introduce a two–step robust procedure to estimate the parameter $\boldsymbol{\beta}$ and the function η, under the generalized partly linear model (1). In Section 3, we give conditions under which the proposed method will lead to strongly consistent estimators, and in Section 4, we derive the asymptotic distribution of those estimators. In Section 5 simulation studies are carried out to assess the robustness and efficiency of the proposals. All the proofs are given in the Appendix.

2 The Proposal

2.1 The estimators

Let $(y_i, \mathbf{x}_i, t_i)$ be independent observations such that $y_i|\left(\mathbf{x}_i, t_i\right) \sim F\left(\cdot, \mu_i\right)$ with $\mu_i = H\left(\eta(t_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$ and $\mathrm{V{\scriptsize AR}}\left(y_i|(\mathbf{x}_i, t_i)\right) = V\left(\mu_i\right)$. Let $\eta_0(t)$ and $\boldsymbol{\beta}_0$ denote the true parameter values, and E_0 the expected value under the true model, so that $\mathrm{E}_0(y|(\mathbf{x},t)) = H\left(\eta_0(t) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0\right)$. Letting $\rho(y,u)$ be a loss function to be specified in the next subsection, we define

$$\begin{aligned} S_n(a,\boldsymbol{\beta},t) &= \sum_{i=1}^{n} W_i(t)\rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + a\right) w_1(\mathbf{x}_i) & (2)\\ S(a,\boldsymbol{\beta},\tau) &= \mathrm{E}_0\left[\rho\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + a\right) w_1(\mathbf{x})|t=\tau\right] \ , & (3)\end{aligned}$$

where $W_i(t)$ are the kernel (or nearest–neighbor with kernel) weights on t_i, and $w_1(\cdot)$ is a function that downweights high leverage points in the $\mathbf{x}$ space. Note that $S_n(a,\boldsymbol{\beta},t)$ is an estimate of $S(a,\boldsymbol{\beta},t)$, which is a continuous function of $(a,\boldsymbol{\beta},t)$ if $(y,\mathbf{x})|t=\tau$ has a distribution function that is continuous with respect to τ.

The Fisher–consistency states that $\eta_0(t) = \operatorname{argmin}_a S(a,\boldsymbol{\beta}_0,t)$. This is a key point in order to get asymptotically unbiased estimates for the nonparametric component. In many situations, a stronger condition holds, that is, under general conditions it can be verified that

$$S(\eta_0(t),\boldsymbol{\beta}_0,t) < S(a,\boldsymbol{\beta},t) \quad \forall \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \quad a \neq \eta_0(t), \tag{4}$$

which entails the Fisher–consistency. Moreover, it is clear that in this case, $\boldsymbol{\beta}_0$ can be estimated by minimizing $S_n(a,\boldsymbol{\beta},t)$, over a and $\boldsymbol{\beta}$. However, this procedure will not lead to a root–n estimate.

Following the ideas of Severini and Staniswalis (1994), we define the function $\eta_{\boldsymbol{\beta}}(t)$ as the minimizer of $S(a,\boldsymbol{\beta},t)$ that will be estimated by the minimizer $\widehat{\eta}_{\boldsymbol{\beta}}(t)$ of $S_n(a,\boldsymbol{\beta},t)$.

To provide an estimate of $\boldsymbol{\beta}$ with the root-n convergence rate, we denote

$$\begin{aligned} F_n(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \widehat{\eta}_{\boldsymbol{\beta}}(t_i)\right) w_2(\mathbf{x}_i) & (5)\\ F(\boldsymbol{\beta}) &= \mathrm{E}_0\left[\rho\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(t)\right) w_2(\mathbf{x})\right] \ , & (6)\end{aligned}$$

where $w_2(\cdot)$ plays the same role (and can be taken to be the same) as $w_1(\cdot)$. We will assume that $\boldsymbol{\beta}_0$ is the unique minimizer of $F(\boldsymbol{\beta})$. This assumption is a standard condition in M–estimation in order to get consistent estimators of the parametric component and is analogous to condition (A-4) of Huber (1981, p.129).

A two–step robust proposal is now given as follows

- **Step 1**: For each value of t and $\boldsymbol{\beta}$, let

$$\widehat{\eta}_{\boldsymbol{\beta}}(t) \;=\; \underset{a\in I\!\!R}{\operatorname{argmin}}\, S_n(a,\boldsymbol{\beta},t)\ , \tag{7}$$

- **Step 2**: Define the estimate of $\boldsymbol{\beta}_0$ as

$$\widehat{\boldsymbol{\beta}} \;=\; \underset{\boldsymbol{\beta}\in I\!\!R^p}{\operatorname{argmin}}\, F_n(\boldsymbol{\beta})\ , \tag{8}$$

  and the estimate of $\eta_0(t)$ as $\widehat{\eta}_{\widehat{\boldsymbol{\beta}}}(t)$.

2.2 Loss function ρ

We propose two classes of loss functions. The first one aims to bound the deviances, while the second one introduced by Cantoni and Ronchetti (2001a) bounds the Pearson residuals.

The first class of loss function takes the form of

$$\rho(y,u) = \phi[-\ln\, f(y,H(u)) + A(y)] + G(H(u))\ , \tag{9}$$

where ϕ is a bounded nondecreasing function with continuous derivative φ, and $f(\cdot,s)$ is the density of the distribution function $F(\cdot,s)$ with $y|(\mathbf{x},t) \sim F\left(\cdot, H\left(\eta_0(t)+\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0\right)\right)$. To avoid triviality, we also assume that ϕ is non–constant in a positive probability set. Typically, ϕ is a function performing like the identity function in a neighborhood of 0. The function $A(y)$ is typically used to remove a term from the log–likelihood that is independent of the parameter, and can be defined as $A(y) = \ln\left(f(y,y)\right)$ in order to get the deviance. The correction term G is used to guarantee the Fisher–consistency, and satisfies

$$\begin{aligned} G'(s) &= \int \varphi[-\ln\, f(y,s) + A(y)]\, f'(y,s) d\mu(y) \\ &= \mathrm{E}_s\left(\varphi[-\ln\, f(y,s)+A(y)]\, f'(y,s)/f(y,s)\right)\ , \end{aligned}$$

where E_s indicates expectation taken under $y \sim F(\cdot,s)$ and $f'(y,s)$ is shorthand for $\partial\, f(y,s)/\partial s$. With this class of ρ functions, we call the resulting estimator a *modified likelihood estimator.*

In a logistic regression setting, Bianco and Yohai (1995) considered the score function

$$\phi(t) = \begin{cases} t - t^2/2c & \text{if } t \le c \\ c/2 & \text{otherwise,} \end{cases}$$

while Croux and Haesbroeck (2002) proposed using the score function

$$\phi(t) = \begin{cases} t\exp(-\sqrt{c}) & \text{if } t \le c \\ -2(1+\sqrt{t})\exp(-\sqrt{t}) + (2(1+\sqrt{c})+c)\exp(-\sqrt{c}) & \text{otherwise.} \end{cases}$$

Both score functions can be used in the general setting. Explicit forms of the correction term $G(s)$, for the Binomial and Poisson families, are given in Bianco and Boente (1996). It is worth noticing that, when considering the deviance and a continuous family of distributions with strongly unimodal density function, the correction term G can be avoided, as discussed in Bianco, García Ben and Yohai (2005).

The second class of loss function is based on Cantoni and Ronchetti (2001a), where they consider a general class of M–estimators of Mallows type, by bounding separately the influence of deviations on y and $(\mathbf{x}, t)$. Their approach is based on robustifying the quasi–likelihood, which is an alternative to the generalizations given for generalized linear regression models by Stefanski, Carroll and Ruppert (1986) and Künsch, Stefanski and Carroll (1989). Let $r(y,\mu) = (y-\mu)\, V^{-1/2}(\mu)$ be the Pearson residuals with $\text{VAR}\,(y_i|(\mathbf{x}_i, t_i)) = V\,(\mu_i)$. Denote $\nu(y,\mu) = V^{-1/2}(\mu)\psi_c\,(r(y,\mu))$, with ψ_c an odd nondecreasing score function with tunning constant c, such as the Huber function, and

$$\rho(y,u) = -\left[\int_{s_0}^{H(u)} \nu(y,s)ds + G(H(u))\right], \tag{10}$$

where s_0 is such that $\nu(y, s_0) = 0$ and the correction term here to ensure Fisher–consistency, also denoted as $G(s)$, satisfies $G'(s) = -\text{E}_s\,(\nu(y,s))$. With such a ρ function, we call the resulting estimator a *quasi-likelihood estimator*. For the Binomial and Poisson families, explicit forms of the correction term $G(s)$ are given in Cantoni and Ronchetti (2001a).

2.2.1 General comments

a) Fisher–Consistency

Under a logistic partly linear regression model, if

$$P\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = \alpha | t = \tau\right) < 1, \qquad \forall(\boldsymbol{\beta},\alpha) \neq 0 \quad \text{and} \quad \tau \in \mathcal{T}\,, \tag{11}$$

and if we consider the score function given by (9) with ϕ satisfying the regularity conditions given in Bianco and Yohai (1995), it is easy to see that (4) holds, and the Fisher–consistency for the nonparametric component is attained under this model. Moreover, it is easy to verify that $\boldsymbol{\beta}_0$ is the unique minimizer of $F(\boldsymbol{\beta})$ in this case.

Condition (11) does not allow $\boldsymbol{\beta}_0$ to include an intercept, so that the model will be identifiable. This means that only the "slope" coefficients can be estimated. Moreover, we do not allow any linear combination of $\mathbf{x}$ to be predicted by t (see Robinson (1988)).

Under a generalized partly linear model with response having a gamma distribution with fixed shape parameter, Theorem 1 of Bianco, García Ben and Yohai (2005) allows us to verify (4) and the Fisher–consistency for the nonparametric and parametric component, if the score function ϕ is bounded and strictly increasing on the set where it is not constant and if (11) holds.

For any generalized partly linear model, conditions similar to those considered in Cantoni and Ronchetti (2001a) will lead to the uniqueness of the differentiated equations, which entail (4). Note that this condition is quite similar to Condition (E) of Severini and Staniswalis (1994, p. 511). When considering the classical quasi–likelihood, the assumption $\boldsymbol{\beta}_0 = \operatorname{argmin}_{\boldsymbol{\beta}} F(\boldsymbol{\beta})$ is related to Condition (7.e.) of Severini and Staniswalis (1994, p. 510), but for the robust quasi–likelihood, this assumption is fulfilled, for instance, for a gamma family with a fixed shape parameter when (11) holds and ψ_c is bounded and increasing.

b) Differentiated equations

If the function $\rho(y,u)$ is continuously differentiable and we denote $\Psi(y,u) = (\partial\rho(y,u))/\partial u$, the estimates will be a solution to the differentiated equations. More precisely, $\eta_{\boldsymbol{\beta}}(t)$ and $\widehat{\eta}_{\boldsymbol{\beta}}(t)$ will be solutions to $S^1(a,\boldsymbol{\beta},t) = 0$ and $S^1_n(a,\boldsymbol{\beta},t) = 0$ respectively, with

$$\begin{aligned} S^1(a,\boldsymbol{\beta},\tau) &= E\left(\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+a\right)w_1(\mathbf{x})|t=\tau\right) & (12)\\ S^1_n(a,\boldsymbol{\beta},t) &= \sum_{i=1}^{n} W_i(t)\Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+a\right)w_1(\mathbf{x}_i)\,. & (13)\end{aligned}$$

Under regularity conditions on the kernel K and on the function Ψ, the implicit function theorem entails that $\eta_{\boldsymbol{\beta}}(t)$ and $\widehat{\eta}_{\boldsymbol{\beta}}(t)$ are continuous functions of both $\boldsymbol{\beta}$

and t, which is a condition that will be required later both for consistency and asymptotic normality.

Besides, $\widehat{\boldsymbol{\beta}}$ is a solution of $F_n^1(\boldsymbol{\beta}) = 0$, and the Fisher consistency states that $F^1(\boldsymbol{\beta}_0) = 0$ and $S^1(\eta_0(t), \boldsymbol{\beta}_0, t) = 0$, where

$$\begin{aligned} F^1(\boldsymbol{\beta}) &= E\left(\Psi\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(t)\right) w_2(\mathbf{x}) \left[\mathbf{x} + \frac{\partial}{\partial \boldsymbol{\beta}} \eta_{\boldsymbol{\beta}}(t)\right]\right) & (14) \\ F_n^1(\boldsymbol{\beta}) &= n^{-1} \sum_{i=1}^{n} \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \widehat{\eta}_{\boldsymbol{\beta}}(t_i)\right) w_2(\mathbf{x}_i) \left[\mathbf{x}_i + \frac{\partial}{\partial \boldsymbol{\beta}} \widehat{\eta}_{\boldsymbol{\beta}}(t_i)\right] . & (15) \end{aligned}$$

Note that these first order equations may have multiple solutions and therefore, we may need the values of the objective functions (2) and (5) to select the final estimator. For a family of distributions with positive and finite information number, Bianco and Boente (1996) give conditions that entail the following: for each t there exists a neighborhood of $\eta_0(t)$ where $S^1(\eta_0(t), \boldsymbol{\beta}_0, t) = 0$ and $S^1(a, \boldsymbol{\beta}_0, t) \neq 0$ for $a \neq \eta_0(t)$. Moreover, $\eta_0(t)$ corresponds to a local minimum of $S(a, \boldsymbol{\beta}_0, t)$. The asymptotic results in this paper are derived by assuming existence of a unique minimum; otherwise, one can only ensure that there exists a solution to the estimating equations that is consistent.

In the modified likelihood approach, the derivative of (9) is given by $\Psi(y, u) = H'(u)\left[\Psi_1(y, H(u)) + G'(H(u))\right]$ where

$$\Psi_1(y, u) \quad = \quad \varphi[-\ln f(y, H(u)) + A(y)]\left[-f'(y, H(u))/f(y, H(u))\right] .$$

On the other hand, for the proposal based on the robust quasi–likelihood, we have the following expression for the derivative of (10)

$$\begin{aligned} \Psi(y, u) &= -\left[\nu(y, H(u)) + G'(H(u))\right] H'(u) \\ &= -\left[\psi_c\left(r(y, H(u))\right) V^{-1/2}(H(u)) + G'(H(u))\right] H'(u) \\ &= -\left[\psi_c\left(r(y, H(u)) - E_{H(u)}\left\{\psi_c\left(r(y, H(u)\right)\right\}\right] H'(u) V^{-1/2}(H(u)) . \end{aligned}$$

For the uniqueness of the minimizer of (3), we note that if (11) holds, and if ϕ is bounded and nondecreasing, the modified likelihood proposal that bounds the deviance satisfies this condition. The same assertion can be verified for the robust quasi–likelihood proposal not only for the logistic case but for other families such as the gamma distribution with a fixed shape parameter, if ψ_c is bounded and increasing. In general, a condition on the behavior of the variance with respect to the mean will be needed to guarantee the uniqueness of the minimum. This uniqueness condition implies that we will be able to solve $S_n^1(a, \boldsymbol{\beta}, t) = 0$ and $F_n^1(\boldsymbol{\beta}) = 0$ to avoid the numerical integration involved in the loss function (10). Moreover, when using the score function of Croux and Haesbroeck (2002), the function $G(s)$ in (9) has an explicit expression without any need for numerical integration.

c) Some robustness issues

It is clear that for unbounded response variables y, a bounded score function allows us to deal with large residuals. For models with a bounded response, e.g., under a logistic model, the advantage of a bounded score function is mainly to guard against outliers with large Pearson residuals. If a binary response y is contaminated, the Pearson residuals are large only when the variances at the contaminated points are close to 0. These points are made more specific in the simulation study in Section 5.

It is also worth noting that our robust procedures are effective only if at least one non-constant covariate $\mathbf{x}$ is present. To consider a case without any covariate, we may take $y_i \sim Bi(1, p)$ as a random sample, then easy calculations show that the minimizer $\widehat{a}$ of $S_n(a) = n^{-1} \sum_{i=1}^n \rho\left(y_i, a\right)$ equals the classical estimator, i.e., $\widehat{a} = H^{-1}\left(\sum_{i=1}^n y_i/n\right)$, with $H(u) = 1/(1+\exp(-u))$, when using either the modified likelihood or the robust quasi–likelihood proposals. The same happens if, $y_i|t_i \sim Bi(1, p(t_i))$, where the resulting estimate of $p(t)$ will be the local mean. In the present paper with a semiparametric model where the covariate $\mathbf{x}$ plays a role, both downweighting the leverage points and controlling outlying responses work towards robustness.

3 Consistency

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

In this section, we will show that the estimates defined through (7) and (8) are consistent under mild conditions, when the smoother weights are the kernel weights $W_i(t) = \left(\sum_{j=1}^n K((t-t_j)/h_n)\right)^{-1} K((t-t_i)/h_n)$. Analogous results can be obtained for the weights based on nearest neighbors using similar arguments to those considered in Boente and Fraiman (1991). In this paper, we will use the following set of assumptions

C1. The function $\rho(y,a)$ is continuous and bounded, and the functions $\Psi(y,a) = \partial\rho(y,a)/\partial a$, $w_1(.)$ and $w_2(.)$ are bounded.

C2. The kernel $K : I\!R \to I\!R$ is an even, nonnegative, 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, and given any compact set $\mathcal{T}_0 \subset \mathcal{T}$ there exists a positive constant $A_1\left(\mathcal{T}_0\right)$ such that $A_1\left(\mathcal{T}_0\right) < f_T(t)$ for all $t \in \mathcal{T}_0$.

C5. The function $S(a,\boldsymbol{\beta},t)$ satisfies the following equicontinuity condition: for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $t_1, t_2 \in \mathcal{T}_0$ and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{K}$, a compact set in R^p,

$$|t_1 - t_2| < \delta \text{ and } \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| < \delta \Rightarrow \sup_{a \in I\!R} |S(a,\boldsymbol{\beta}_1,t_1) - S(a,\boldsymbol{\beta}_2,t_2)| < \epsilon\,.$$

C6. The function $S(a,\boldsymbol{\beta},t)$ is continuous, and $\eta_{\boldsymbol{\beta}}(t)$ is a continuous function of $(\boldsymbol{\beta},t)$.

Remark 3.1. If the conditional distribution of $\mathbf{x}|t=\tau$ is continuous with respect to τ, the continuity and boundness of ρ stated in **C1** entail that $S(a,\boldsymbol{\beta},\tau)$ 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. In the classical case it corresponds to Condition (D) of Severini and Staniswalis (1994, p. 511). It is also considered in nonparametric regression when the uniform consistency results on the t-space are needed; it allows us to deal with the denominator in the definition of the kernel weights, which is in fact an estimate of the marginal density f_T.

Assumption **C5** is fulfilled under **C1** if the following equicontinuity condition holds: for any $\epsilon > 0$ there exist compact sets $\mathcal{K}_1 \subset I\!R$ and $\mathcal{K}_p \subset I\!R^p$ such that for any $\tau \in \mathcal{T}_0$ $P\left((y, \mathbf{x}) \in \mathcal{K}_1 \times \mathcal{K}_p | t = \tau\right) > 1 - \epsilon$, which holds for instance if, for $1 \le j \le p$, $x_{ij} = \phi_j(t_i) + u_{ij}$, $1 \le i \le n$, where ϕ_j are continuous functions and u_{ij} are *i.i.d* and independent of t_i.

Theorem 3.1. *Let $\mathcal{K} \subset I\!R^p$ 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) K is of bounded variation,

ii) the family of functions $\mathcal{F} = \{f(y, \mathbf{x}) = \rho\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + a\right) w_1(\mathbf{x}), \boldsymbol{\beta} \in \mathcal{K}, a \in I\!R\}$ has covering number $N\left(\epsilon, \mathcal{F}, L^1(I\!\!Q)\right) \le A\epsilon^{-W}$, for any probability $I\!\!Q$ and $0 < \epsilon < 1$.

Then, we have

a) $\sup_{\substack{\boldsymbol{\beta} \in \mathcal{K} \\ a \in I\!R}} \|S_n(a, \boldsymbol{\beta}, \cdot) - S(a, \boldsymbol{\beta}, \cdot)\|_{0,\infty} \xrightarrow{a.s.} 0.$

b) If $\inf_{\substack{\boldsymbol{\beta} \in \mathcal{K} \\ t \in \mathcal{T}_0}} \left[\lim_{|a| \to \infty} S(a, \boldsymbol{\beta}, t) - S(\eta_{\boldsymbol{\beta}}(t), \boldsymbol{\beta}, t)\right] > 0$, *then*

$$\sup_{\boldsymbol{\beta} \in \mathcal{K}} \|\widehat{\eta}_{\boldsymbol{\beta}} - \eta_{\boldsymbol{\beta}}\|_{0,\infty} \xrightarrow{a.s.} 0 \, .$$

Theorem 3.2. *Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of $F_n(\boldsymbol{\beta})$ where $F_n(\boldsymbol{\beta})$ is defined as in (5) with $\widehat{\eta}_{\boldsymbol{\beta}}$ satisfying*

$$\sup_{\boldsymbol{\beta} \in \mathcal{K}} \|\widehat{\eta}_{\boldsymbol{\beta}} - \eta_{\boldsymbol{\beta}}\|_{0,\infty} \xrightarrow{a.s.} 0 \qquad (16)$$

*for any compact set $\mathcal{K}$ in R^p. If **C1** holds, then*

a) $\sup_{\boldsymbol{\beta} \in \mathcal{K}} |F_n(\boldsymbol{\beta}) - F(\boldsymbol{\beta})| \xrightarrow{a.s.} 0,$

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

Remark 3.2. Theorems 3.1 and 3.2 entail that $\|\widehat{\eta}_{\widehat{\boldsymbol{\beta}}} - \eta_0\|_{0,\infty} \xrightarrow{a.s.} 0$, since $\eta_{\boldsymbol{\beta}}(t)$ is continuous.

4 Asymptotic Normality

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

Theorem 4.1. *Assume that the t_i's are random variables with distribution on a compact set $\mathcal{T}$. Assume that **N1** to **N6** hold, then for any consistent solution $\widehat{\boldsymbol{\beta}}$ of (15), we have*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{D} N\left(\mathbf{0}, \mathbf{A}^{-1}\boldsymbol{\Sigma}\left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\right) ,$$

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

Remark 4.1. Theorem 4.1 can be used to construct a Wald-type statistic to make inferences involving only a subset of the regression parameter, that is, when we want to test $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$, with $\boldsymbol{\beta}^{\mathrm{T}} = \left(\boldsymbol{\beta}_{(1)}^{\mathrm{T}}, \boldsymbol{\beta}_{(2)}^{\mathrm{T}}\right)$.

Likelihood ratio-type tests can also be used based on the robust quasi–likelihood introduced in Section 2, as it was done for generalized linear models by Cantoni and Ronchetti (2001a), or on the robustified deviance. A robust measure of discrepancy between the two models is defined as

$$\Lambda = 2\left[\sum_{i=1}^{n} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{\eta}_{\widehat{\boldsymbol{\beta}}}(t_i)\right) w_2(\mathbf{x}_i) - \sum_{i=1}^{n} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_0 + \widehat{\eta}_{\widehat{\boldsymbol{\beta}}_0}(t_i)\right) w_2(\mathbf{x}_i)\right]$$

where $\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}} = \left(\widehat{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\right)$ is the estimate of $\boldsymbol{\beta}$ under the null hypothesis. Both estimates $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}$ need to be computed using the same score function ρ considered in Λ, in order to ensure that Λ will behave asymptotically as a linear combination of independent chi–square random variables with one degree of freedom. As in Cantoni and Ronchetti (2001a), it can be seen that $\Lambda = n\mathbf{U}_{n,(2)}^{\mathrm{T}}\mathbf{A}_{22.1}\mathbf{U}_{n,(2)} + o_p(1)$ with $\mathbf{A}_{22.1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and $\sqrt{n}\mathbf{U}_n \xrightarrow{D} N\left(\mathbf{0}, \mathbf{A}^{-1}\boldsymbol{\Sigma}\left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\right)$.

5 Monte Carlo Study

A small scale simulation study was carried out to assess the performance of the robust estimators considered in this paper. A one-dimensional covariate x and a nonparametric function $\eta(t)$ were considered. The modified likelihood estimator (MOD) used the score function of Croux and Haesbroeck (2002) with $c = 0.5$. With this choice, the function $C(s)$ has an explicit expression so no numerical integration is needed. The weight functions take the following form

$$w_1^2(x_i) = w_2^2(x_i) = \{1 + (x_i - M_n)^2\}^{-1}$$

where $M_n = Median\{x_j : \ j = 1, \cdots, n\}$ is the sample median.

The two competitors considered in the study were the quasi-likelihood estimator (QAL) of Severeni and Staniswalis (1994) and the robust quasi-likelihood estimator (RQL) of Cantoni and Ronchetti (2001a). For the latter, the Huber function $\psi_{1.2}(x) = \max\{-1.2, \min(1.2, x)\}$ was used with the same weight functions as above. The QAL estimator corresponds to $\psi(x) = x$ and $w_1(x) = w_2(x) = 1$. In all cases, the kernel $K(t) = \max\{0, 1 - |t|\}$ was used. In Studies 1 and 3 below, the search for $\boldsymbol{\beta}$ uses a grid of size 0.05, while in Study 2 the grid size is 0.01.

An important issue in any smoothing procedure is the choice of the smoothing parameter. Under a nonparametric regression model with $\boldsymbol{\beta} = 0$ and $H(t) = t$, two commonly used approaches are cross–validation and plug–in. However, these procedures may not be robust and their sensitivity to anomalous data was discussed by several authors, including Leung, Marrot and Wu (1993), Wang and Scott (1994), Boente, Fraiman and Meloche (1997) and Cantoni and Ronchetti (2001b). Wang and Scott (1994) note that, in the presence of outliers, the least squares cross–validation function is nearly constant on its whole domain and thus, essentially worthless for the purpose of choosing a bandwidth. The robustness issue remains for the estimators considered in this paper. With a small bandwidth, a small number of outliers with similar values of t_i could easily drive the estimate of η to dangerous levels. Therefore, we may consider a robust cross-validation approach as follows.

- For each given h, let

$$\begin{aligned} \widehat{\eta}_{\boldsymbol{\beta}}^{(-i)}(t,h) &= \underset{a\in\mathbb{R}}{\operatorname{argmin}} \sum_{j\neq i}^{n} W_j(t,h)\rho\left(y_j, \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}+a\right) w_1(\mathbf{x}_j) \\ \widehat{\boldsymbol{\beta}}^{(-i)}(h) &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{j\neq i}^{n} \rho\left(y_j, \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}+\widehat{\eta}_{\boldsymbol{\beta}}^{(-j)}(t_j,h)\right) w_2(\mathbf{x}_j)\ , \end{aligned}$$

  where $W_i(t,h) = \{\sum_{j=1}^{n} K((t-t_j)/h)\}^{-1} K((t-t_i)/h)$.

- Choose

$$\widehat{h}_n = \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(-i)}(h) + \widehat{\eta}_{\widehat{\boldsymbol{\beta}}^{(-i)}}^{(-i)}(t_i,h)\right) w_2(\mathbf{x}_i)\ .$$

However, this method is computationally expensive. Another approach divides the sample into two subsets by choosing at random $100\,(1-\alpha)\%$ of the sample as training sample and $100\,\alpha\%$ as validating sample. This procedure can be described as follows:

- Select at random a subset of size $100\,(1-\alpha)\%$. Let $\mathcal{J}_{1-\alpha}$ stand for the indexes of these observations and $\mathscr{J}_{1-\alpha}$ for the indexes of the remaining ones.

- For each given h, compute

$$\begin{aligned} \widehat{\eta}_{\boldsymbol{\beta}}^{(-\alpha)}(t,h) &= \underset{a\in\mathbb{R}}{\operatorname{argmin}} \sum_{i\in\mathcal{J}_{1-\alpha}} W_i(t,h)\rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+a\right) w_1(\mathbf{x}_i) \\ \widehat{\boldsymbol{\beta}}^{(-\alpha)}(h) &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in\mathcal{J}_{1-\alpha}} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\widehat{\eta}_{\boldsymbol{\beta}}^{(-\alpha)}(t_i,h)\right) w_2(\mathbf{x}_i)\ , \end{aligned}$$

  where $W_i(t,h)$ are the kernel weights defined above.

- Choose

$$\widehat{h}_n = \underset{h}{\operatorname{argmin}} \sum_{i\in\mathscr{J}_{1-\alpha}} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(-\alpha)}(h) + \widehat{\eta}_{\widehat{\boldsymbol{\beta}}^{(-\alpha)}}^{(-\alpha)}(t_i,h)\right) w_2(\mathbf{x}_i).$$

This is the procedure we have found to be helpful based on our experience with a number of data sets, including some from Study 1 below, but a full evaluation of this method is not yet done.

To measure performance, we use the bias and standard deviation for the $\boldsymbol{\beta}$ estimate as well as the mean square error of the function estimate

$$\mathrm{MSE}(\widehat{\eta}) = n^{-1} \sum_{i=1}^{n} \left[\widehat{\eta}(t_i) - \eta(t_i)\right]^2 .$$

We report the comparisons under three scenarios as follows.

Study 1: Random samples of size $n = 100$ were generated from the following model

$$x \sim \mathcal{U}(-1,1), \quad t \sim \mathcal{U}(\{.1, .2, \cdots, 1.0\}), \quad y|(x,t) \sim Bi(10, p(x,t))$$

where $\log(p(x,t)/(1-p(x,t))) = 3x + e^{2t} - 4$. We summarized the results over 100 runs in Table 1, using three different bandwidths $h_n = 0.1$, $h_n = 0.2$ and $h_n = 0.3$. The three estimates are labelled as QAL(h_n), MOD(h_n) and RQL (h_n). Figure 1 gives the histograms of the estimates of β for each method and bandwidth. It is clear that the robust estimators MOD and RQL have similar performance, and the relative efficiencies of the MOD(h_n) are between 0.69 and 0.80 as compared to QAL (h_n). The MOD method tends to have smaller bias than the RQL method and even than the QAL method. The normality of $\hat{\beta}$ appeared to hold up quite well at this sample size.

We also applied the data-adaptive method described in this section for choosing h_n based on a split of the sample into a training set (80% of the data) and a validation set (20%). On a total of 10 random samples for Study 1, the resulting h_n are mostly between 0.1 and 0.2. From Table 1, we notice that $h_n = 0.2$ is indeed a good choice, but the performance of $\hat{\beta}$ is not very sensitive to the choice of h_n.

Study 2: To see how the robust estimators protect us from gross errors in the data, we generated a data set of $n = 100$ from the following model

$$x \sim N(0,1), \quad t \sim N(1/2, 1/6), \quad y|(x,t) \sim Bi(10, p(x,t))$$

where $\log(p(x,t)/(1-p(x,t))) = 2x + 0.2$. Then, we replaced the first one, two and three observations by gross outliers. Table 2 gives the parameter estimates under the contaminated data, with $h_n = 0.1$, where (x_i, y_i), $1 \leq i \leq 3$ denote the outliers. It is clear that the QAL estimate of β was very sensitive to a single outlier whereas the robust estimators remained stable.

Study 3: We considered a data set of size $n = 200$. We first generated data from a bivariate normal distribution $(x_i, t_i) \sim N((0, 1/2), \Sigma)$ truncated to $t \in [1/4, 3/4]$ with

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

The response variable was then generated as

$$y_i = \begin{cases} 1, & \beta_0 x_i + \eta_0(t_i) + \epsilon_i \geq 0 \\ 0, & \beta_0 x_i + \eta_0(t_i) + \epsilon_i < 0 \end{cases}$$

where $\beta_0 = 2, \eta_0(t) = 2\sin(4\pi t)$, ϵ_i was a standard logistic variate. For each data set generated from this model, we also created three contaminated data sets denoted C_1, C_2 and C_3 in Table 3, respectively. The purpose of the first two contaminations is to see how the robust methods work when one has contamination in y only.

- **Contamination 1**. The contaminated data points were generated as follows: $u_i \sim \mathcal{U}(0, 1)$, $x_i = x_i$, and

$$y_i = \begin{cases} y_i & \text{if } u_i \leq 0.90 \\ \text{a new observation from } Bi(1, 0.5) & \text{if } u_i > 0.90 \end{cases}$$

- **Contamination 2**. For each generated data set, we chose 10 “design points” with $H(\beta_0 x_i + \eta_0(t_i)) > 0.99$, where $H(u) = 1/(1 + \exp(-u))$, so at those points the conditional mean of y given the covariates is not close to 0.5. Then, we contaminate y as in Contamination 1 but only at those 10 points. At those 10 points, about half of them are expected to be outliers with large Pearson residuals.

- **Contamination 3**. Here, we considered a contamination with bad leverage points by using $u_i \sim \mathcal{U}(0, 1)$,

$$x_i = \begin{cases} x_i & \text{if } u_i \leq 0.90 \\ \text{a new observation from N}(10, 1) & \text{if } u_i > 0.90 \end{cases}$$

$$y_i = \begin{cases} y_i & \text{if } u_i \leq 0.90 \\ \text{a new observation from } Bi(1, 0.05) & \text{if } u_i > 0.90\ . \end{cases}$$

Both the original and the contaminated data sets were analyzed using the three competing estimators. Using a bandwidth of $h_n = 0.1$, we summarized the results in Tables 3 and 4 based on 100 Monte Carlo samples. The bandwidth was chosen to be smaller than that in Study 1, because we have 200 distinct observed values of t here as compared to 10 in the earlier study. Table 3 shows the poor performance of the classical estimates of β, specially under contamination C_3. It is worth noticing, that our contamination framework in Study 3, shows that the bounded score becomes more helpful if outliers are present in the sense that the Pearson residuals are large, which could happen when y is contaminated to 1 (or 0) when the expected value of y from the model is nearly 0 (or 1). Under C_1, most contaminated y do no result in large Pearson residuals, and the robust estimators RQL and MOD can improve the non-robust estimator somewhat, but not as significantly as under C_2 and specially under C_3, where high leverage points are downweighted in the robust procedure introduced. With respect to the estimation of η, all procedures seem to be stable, because the magnitude of outlying y is very limited in this case.

Our studies show the good performance of the two families of robust estimators considered here in the presence of outliers. The MOD method often shows smaller bias for estimating $\boldsymbol{\beta}$ but its mean squared error is usually similar to that of RQL.

Appendix

A.1 Proof of the consistency results

Proof of Theorem 3.1. a) Let $Z_i(a, \boldsymbol{\beta}) = \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + a\right) w_1(\mathbf{x}_i)$,

$$\begin{aligned} R_{1n}(a, \boldsymbol{\beta}, t) &= (nh_n)^{-1} \sum_{i=1}^{n} Z_i(a, \boldsymbol{\beta}) K((t - t_i)/h_n) \\ R_{0n}(t) &= (nh_n)^{-1} \sum_{i=1}^{n} K((t - t_i)/h_n) \ . \end{aligned}$$

Then, $S_n(a,\boldsymbol{\beta},t)=R_{1n}(a,\boldsymbol{\beta},t)/R_{0n}(t)$ which entails

$$\begin{aligned}\sup_{\substack{\boldsymbol{\beta}\in\mathcal{K}\\ a\in\boldsymbol{R}}}\|S_n(a,\boldsymbol{\beta},\cdot)-S(a,\boldsymbol{\beta},\cdot)\|_{0,\infty} &\leq \left[\sup_{\substack{\boldsymbol{\beta}\in\mathcal{K}\\ a\in\boldsymbol{R}}}\|R_{1n}(a,\boldsymbol{\beta},\cdot)-E\left(R_{1n}(a,\boldsymbol{\beta},\cdot)\right)\|_{0,\infty}\right.\\ &+ \sup_{\substack{\boldsymbol{\beta}\in\mathcal{K}\\ a\in\boldsymbol{R}}}\|E\left(R_{1n}(a,\boldsymbol{\beta},\cdot)\right)-S(a,\boldsymbol{\beta},\cdot)E\left(R_{0n}(\cdot)\right)\|_{0,\infty}\\ &+ \left.\|\rho\|_\infty\|w_1\|_\infty\left\|R_{0n}-E\left(R_{0n}\right)\right\|_{0,\infty}\right]\left\{\inf_{t\in\mathcal{T}_0}R_{0n}(t)\right\}^{-1},\end{aligned}$$

where $\|\rho\|_\infty=\sup_{(y,a)}|\rho(y,a)|$ and $\|w_1\|_\infty=\sup_{\mathbf{x}}|w_1(\mathbf{x})|$.

Since, for n large enough,

$$\begin{aligned}\inf_{t\in\mathcal{T}_0}R_{0n}(t) &\geq \inf_{t\in\mathcal{T}_0}E\left(R_{0n}(t)\right)-\|R_{0n}-E\left(R_{0n}\right)\|_{0,\infty}\\ E\left(R_{0n}(t)\right) &= \int K(u)f_T(t-uh_n)du>A_1\left(\mathcal{T}_\delta\right)\ ,\end{aligned}$$

it is enough to show that

$$\sup_{\substack{\boldsymbol{\beta}\in\mathcal{K}\\ a\in\boldsymbol{R}}}\|R_{1n}(a,\boldsymbol{\beta},\cdot)-E\left(R_{1n}(a,\boldsymbol{\beta},\cdot)\right)\|_{0,\infty}\xrightarrow{a.s.}0\ ,\qquad(A.1)$$

$$\|R_{0n}-E\left(R_{0n}\right)\|_{0,\infty}\xrightarrow{a.s.}0\ ,\qquad(A.2)$$

$$\sup_{\substack{\boldsymbol{\beta}\in\mathcal{K}\\ a\in\boldsymbol{R}}}\|E\left(R_{1n}(a,\boldsymbol{\beta},\cdot)\right)-S(a,\boldsymbol{\beta},\cdot)E\left(R_{0n}(\cdot)\right)\|_{0,\infty}\to 0\ .\qquad(A.3)$$

Assumptions **C2** to **C4** entail (A.2) (see, for instance, Pollard, pp. 35, 1984). On the other hand, since

$$\begin{aligned}|E\left(R_{1n}(a,\boldsymbol{\beta},t)\right)\ -\ &S(a,\boldsymbol{\beta},t)E\left(R_{0n}(t)\right)|\\ &= \left|h^{-1}E\left[\left(S(a,\boldsymbol{\beta},t_1)-S(a,\boldsymbol{\beta},t)\right)K((t-t_1)/h_n)\right]\right|\\ &= \left|\int\left(S(a,\boldsymbol{\beta},t-uh_n)-S(a,\boldsymbol{\beta},t)\right)K\left(u\right)f_T(t-uh_n)du\right|\\ &\leq \|f_T\|_\infty\int|S(a,\boldsymbol{\beta},t-uh_n)-S(a,\boldsymbol{\beta},t)|\,K\left(u\right)du\ ,\end{aligned}$$

(A.3) follows easily from the boundness of ρ, the integrability of the kernel, the equicontinuity condition **C5** and the fact that $h_n\to 0$.

It remains to prove (A.1). Let us consider the class of functions

$$\begin{aligned}\mathcal{F}_n=\{f_{t,a,\boldsymbol{\beta},h_n}(y,\mathbf{x},v) &= B^{-1}\rho(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+a)w_1(\mathbf{x})K\left((t-v)/h_n\right)\\ &= B^{-1}\rho(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+a)w_1(\mathbf{x})K_{t,h_n}(v)\}\end{aligned}$$

with $B = \|\rho\|_\infty \|w_1\|_\infty \|K\|_\infty$. From Problem 27 in Pollard (1984) the graphs of translated kernels K_{t,h_n} have polynomial discrimination and $0 \leq K_{t,h_n} \leq \|K\|_\infty$, which together with assumption ii) entails that $N\left(\epsilon, \mathcal{F}_n, L^1(\mathcal{Q})\right) \leq A_1 \epsilon^{-W_1}$, for all probability $\mathcal{Q}$ and $0 < \epsilon < 1$, where A_1 and W_1 do not depend on n. For any $f_{t,a,\boldsymbol{\beta},h_n} \in \mathcal{F}_n$, $|f_{t,a,\boldsymbol{\beta},h}| \leq 1$ and $E\left(f^2_{t,a,\boldsymbol{\beta},h_n}(y,\mathbf{x},v)\right) \leq h_n \|K\|^{-1}_\infty \|f_T\|_\infty$. Then, Theorem 37 in Pollard (1984) (with $\alpha_n = 1$, $\delta_n^2 = h_n$) and **C4** entail that

$$(h_n)^{-1} \sup_{\mathcal{F}_n} \left| n^{-1} \sum_{i=1}^n f_{t,a,\boldsymbol{\beta},h_n}(y_i, \mathbf{x}_i, t_i) - E f_{t,a,\boldsymbol{\beta},h_n}(y_1, \mathbf{x}_1, t_1) \right| \xrightarrow{a.s.} 0 \ ,$$

which concludes the proof of (A.1).

b) Since $\eta_{\boldsymbol{\beta}}(t)$ is a continuous function of both $(\boldsymbol{\beta}, t)$, we get that $\eta_{\boldsymbol{\beta}}(t)$ is bounded for $t \in \mathcal{T}_0$ and $\boldsymbol{\beta} \in \mathcal{K}$ and thus there exists a compact set $\mathcal{A}(\mathcal{T}_0, \mathcal{K})$ such that $\eta_{\boldsymbol{\beta}}(t) \in \mathcal{A}(\mathcal{T}_0, \mathcal{K})$ for any $t \in \mathcal{T}_0$ and $\boldsymbol{\beta} \in \mathcal{K}$. Assume that $\sup_{\boldsymbol{\beta} \in \mathcal{K}} \left\| \widehat{\eta}_{\boldsymbol{\beta}} - \eta_{\boldsymbol{\beta}} \right\|_{0,\infty}$ does not converge to 0 in a set Ω_0 with $P(\Omega_0) > 0$. Then, for each $\omega \in \Omega_0$ we have that there exists a sequence $(\boldsymbol{\beta}_k, t_k)$ such that $t_k \in \mathcal{T}_0$, $\boldsymbol{\beta}_k \in \mathcal{K}$ and $\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k) - \eta_{\boldsymbol{\beta}_k}(t_k) \to c \neq 0$. Since $\mathcal{T}_0$ and $\mathcal{K}$ are compact without loss of generality we can assume $t_k \to t_L \in \mathcal{T}_0$ and $\boldsymbol{\beta}_k \to \boldsymbol{\beta}_L \in \mathcal{K}$. From the continuity of $\eta_{\boldsymbol{\beta}}(t)$, we get that $\eta_{\boldsymbol{\beta}_k}(t_k) \to \eta_{\boldsymbol{\beta}_L}(t_L)$, which entails that $\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k) - \eta_{\boldsymbol{\beta}_L}(t_L) \to c$.

Assume first that $c < \infty$. Then, the proof follow the same steps as that of Lemma A1 of Carroll, Fan, Gijbels and Wand (1997).

If $c = \infty$, we have that $\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k) \to \infty$. By assumption, we have that

$$0 < i = \inf_{\substack{\boldsymbol{\beta} \in \mathcal{K} \\ t \in \mathcal{T}_0}} \left[\lim_{|a| \to \infty} S(a, \boldsymbol{\beta}, t) - S(\eta_{\boldsymbol{\beta}}(t), \boldsymbol{\beta}, t) \right] \ ,$$

and so $\lim_{|a|\to\infty} S(a, \boldsymbol{\beta}_L, t_L) - S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_L, t_L) \geq i$, thus for k large enough $S(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_L, t_L) > S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_L, t_L) + i/2$. The equicontinuity condition entails that given $\epsilon > 0$ for k large enough, $S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_k, t_k) \leq S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_L, t_L) + \epsilon/4$ and $S(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_L, t_L) \leq S(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_k, t_k) + \epsilon/4$ which from (a) and the definition of $\widehat{\eta}_{\boldsymbol{\beta}}$ entails

$$S(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_L, t_L) \leq S_n(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_k, t_k) + \epsilon/2 \leq S_n(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_k, t_k) + \epsilon/2 \ .$$

Using again (a), we get $S(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_L, t_L) \leq S_n(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_k, t_k) + \epsilon/2 \leq S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_k, t_k) + 3\epsilon/4 \leq S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_L, t_L) + \epsilon$. Hence, for k large enough

$S(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_L, t_L) \leq S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_L, t_L) + \epsilon$, which contradicts the fact that $S(\widehat{\eta}_{\boldsymbol{\beta}_k}(t_k), \boldsymbol{\beta}_L, t_L) > S(\eta_{\boldsymbol{\beta}_L}(t_L), \boldsymbol{\beta}_L, t_L) + i/2$. □

The following Proposition states a general uniform convergence result which will be helpful in proving Theorem 3.2. and Theorem 4.1.

We will begin by fixing some notation. Denote $\mathcal{C}^1(\mathcal{T})$ the set of continuously differentiable functions in $\mathcal{T}$. Note that if $S^1(a, \boldsymbol{\beta}, \tau)$ defined in (12) is continuously differentiable with respect to (a, τ) then $\eta_{\boldsymbol{\beta}} \in \mathcal{C}^1(\mathcal{T})$. $\mathcal{V}(\boldsymbol{\beta})$ and $\mathcal{H}_\delta(\boldsymbol{\beta})$ stand for neighborhoods of $\boldsymbol{\beta} \in \mathcal{K}$ and $\eta_{\boldsymbol{\beta}}$ such that $\mathcal{V}(\boldsymbol{\beta}) \subset \mathcal{K}$ and

$$\mathcal{H}_\delta(\boldsymbol{\beta}) = \left\{ u \in \mathcal{C}^1(\mathcal{T}) : \quad \|u - \eta_{\boldsymbol{\beta}}\|_\infty \leq \delta \ , \quad \left\| \frac{\partial}{\partial t} u - \frac{\partial}{\partial t} \eta_{\boldsymbol{\beta}} \right\|_\infty \leq \delta \right\} .$$

Proposition A.1.1. *Let $(y_i, \mathbf{x}_i, t_i)$ be independent observations such that $y_i| \left(\mathbf{x}_i, t_i\right) \sim F\left(\cdot, \mu_i\right)$ with $\mu_i = H\left(\eta_0(t_i) + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right)$ and $\mathrm{VAR}\left(y_i|(\mathbf{x}_i, t_i)\right) = V\left(\mu_i\right)$. Assume that t_i are random variables with distribution on $\mathcal{T}$. Let $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a continuous and bounded function, $W(\mathbf{x}, t) : \mathbb{R}^{p+1} \rightarrow \mathbb{R}$ be such that $E(|W(\mathbf{x}, t)|) < \infty$ and $\eta_{\boldsymbol{\beta}}(t) = \eta(\boldsymbol{\beta}, t) : \mathbb{R}^{p+1} \rightarrow \mathbb{R}$ be a continuous function of $(\boldsymbol{\beta}, t)$. Define $L\left(y, \mathbf{x}, t, \boldsymbol{\beta}, v\right) = g\left(y, \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta} + v(t)\right) W(\mathbf{x}, t)$ and $\mathbf{E}(\boldsymbol{\beta}) = E_0\left(L\left(y, \mathbf{x}, t, \boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}\right)\right)$. Then,*

- *a)* $E\left(n^{-1} \sum_{i=1}^n L\left(y_i, \mathbf{x}_i, t_i, \boldsymbol{\theta}, v\right)\right) \rightarrow \mathbf{E}(\boldsymbol{\beta})$ *when* $\|\boldsymbol{\theta} - \boldsymbol{\beta}\| + \left\|v - \eta_{\boldsymbol{\beta}}\right\|_\infty \rightarrow 0$.
- *b)* $\sup_{\boldsymbol{\theta} \in \mathcal{K}} \left|n^{-1} \sum_{i=1}^n L\left(y_i, \mathbf{x}_i, t_i, \boldsymbol{\theta}, \eta_{\boldsymbol{\theta}}\right) - E\left(L\left(y_i, \mathbf{x}_i, t_i, \boldsymbol{\theta}, \eta_{\boldsymbol{\theta}}\right)\right)\right| \xrightarrow{a.s.} 0$.
- *c)* $\sup_{\boldsymbol{\theta} \in \mathcal{K}, v \in \mathcal{H}_1(\boldsymbol{\beta})} \left|n^{-1} \sum_{i=1}^n L\left(y_i, \mathbf{x}_i, t_i, \boldsymbol{\theta}, v\right) - E\left(L\left(y_i, \mathbf{x}_i, t_i, \boldsymbol{\theta}, v\right)\right)\right| \xrightarrow{a.s.} 0$, *if in addition, $\mathcal{T}$ is compact and $\eta_{\boldsymbol{\beta}} \in \mathcal{C}^1(\mathcal{T})$.*

PROOF. a) follows from the Dominated Convergence Theorem. The proof of (b) and (c) follows using the continuity of $\eta_{\boldsymbol{\beta}}$ and g, Theorem 3 in Chapter 2 of Pollard (1984), the compactness of $\mathcal{K}$ and $\mathcal{H}_1(\boldsymbol{\beta})$ and analogous arguments to those considered in Bianco and Boente (2002). □

Remark A.1.1. Proposition A.1.1. entails that for any weakly consistent estimate

$\widehat{\eta}_{\boldsymbol{\beta}}$ of $\eta_{\boldsymbol{\beta}}$ such that

$$\begin{aligned}\sup_{t\in\mathcal{T}}\left|\widehat{\eta}_{\boldsymbol{\beta}}(t)-\eta_{\boldsymbol{\beta}}(t)\right| &\xrightarrow{a.s.} 0,\\ \sup_{t\in\mathcal{T}}\left|\frac{\partial}{\partial t}\widehat{\eta}_{\boldsymbol{\beta}}(t)-\frac{\partial}{\partial t}\eta_{\boldsymbol{\beta}}(t)\right| &\xrightarrow{a.s.} 0\ ,\end{aligned}$$

we have $(1/n)\sum_{i=1}^n \mathbf{H}\left(y_i,\mathbf{x}_i,t_i,\boldsymbol{\beta},\widehat{\eta}_{\boldsymbol{\beta}}\right)\xrightarrow{a.s.}\mathbf{E}(\boldsymbol{\beta})$.

An analoguous result can be obtained replacing $\xrightarrow{a.s.}$ by $\xrightarrow{p}$.

Proof of Theorem 3.2. a) Define

$$\widetilde{F}_n(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^n\rho\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\eta_{\boldsymbol{\beta}}(t_i)\right)w_2(\mathbf{x}_i)\ .$$

For any $\epsilon>0$, let $\mathcal{T}_0$ be a compact set such that $P\left(t_i\notin\mathcal{T}_0\right)<\epsilon$. We have that

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathcal{K}}|F_n(\boldsymbol{\beta})-\widetilde{F}_n(\boldsymbol{\beta})| &\le \|\Psi\|_\infty\|w_2\|_\infty\sup_{\boldsymbol{\beta}\in\mathcal{K}}\left\|\widehat{\eta}_{\boldsymbol{\beta}}-\eta_{\boldsymbol{\beta}}\right\|_{0,\infty}\\ &+ 2\|\rho\|_\infty\|w_2\|_\infty\frac{1}{n}\sum_{i=1}^n I\left(t_i\notin\mathcal{T}_0\right)\end{aligned}$$

and hence, using (16) and the Strong Law of Large numbers we get easily that

$$\sup_{\boldsymbol{\beta}\in\mathcal{K}}|F_n(\boldsymbol{\beta})-\widetilde{F}_n(\boldsymbol{\beta})|\xrightarrow{a.s.}0\ . \tag{A.4}$$

Moreover, Proposition A.1.1.b) with $W(\mathbf{x},t)=w_2(\mathbf{x})$ and $g(y,u)=\rho(y,u)$, implies that $\sup_{\boldsymbol{\beta}\in\mathcal{K}}|\widetilde{F}_n(\boldsymbol{\beta})-F(\boldsymbol{\beta})|\xrightarrow{a.s.}0$ which together with (A.4) concludes the proof of a).

b) Note that (a) entails that

$$\begin{aligned}F_n(\widehat{\boldsymbol{\beta}})=\inf_{\boldsymbol{\beta}\in\mathcal{K}}F_n(\boldsymbol{\beta}) &\xrightarrow{a.s.} \inf_{\boldsymbol{\beta}\in\mathcal{K}}F(\boldsymbol{\beta})=F(\boldsymbol{\beta}_0)\\ F_n(\widehat{\boldsymbol{\beta}})-F(\widehat{\boldsymbol{\beta}}) &\xrightarrow{a.s.} 0\end{aligned}$$

and so, $F(\widehat{\boldsymbol{\beta}})\xrightarrow{a.s.}F(\boldsymbol{\beta}_0)$. Since F has a unique minimum at $\boldsymbol{\beta}_0$, b) follows easily. □

A.2 Proof of the asymptotic normality of the regression estimates

For the sake of simplicity, we denote

$$\begin{aligned}\chi(y,a) &= \frac{\partial}{\partial a}\Psi\left(y,a\right)\\ \chi_1(y,a) &= \frac{\partial^2}{\partial a^2}\Psi\left(y,a\right)\end{aligned}$$

$$\widehat{\upsilon}(\boldsymbol{\beta},t)=\widehat{\eta}_{\boldsymbol{\beta}}(t)-\eta_{\boldsymbol{\beta}}(t) \qquad \widehat{\upsilon}_0(t)=\widehat{\upsilon}(\boldsymbol{\beta}_0,t) \tag{A.5}$$

$$\widehat{\mathrm{v}}_j(\boldsymbol{\beta},t)=\frac{\partial\widehat{\upsilon}(\boldsymbol{\beta},t)}{\partial\beta_j} \qquad \widehat{\mathrm{v}}_{j,0}(t)=\widehat{\mathrm{v}}_j(\boldsymbol{\beta}_0,t)\ . \tag{A.6}$$

We list the following conditions needed for the asymptotic normality theorem, followed by general comments on those conditions. The first condition is on the preliminary estimate of $\eta_{\boldsymbol{\beta}}(t)$, and the rest are on the score functions and the underlying model distributions.

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

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

b) $\left\|\widehat{\eta}_{\widehat{\boldsymbol{\beta}}}-\eta_0\right\|_\infty\xrightarrow{p}0$, for any consistent estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$.

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

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

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

N2. The functions Ψ, χ, χ_1, w_2 and $\psi_2(\mathbf{x})=\mathbf{x}w_2(\mathbf{x})$ are bounded and continuous.

N3. The matrix $\mathbf{A}$ is non-singular, where

$$\begin{aligned}\mathbf{A} &= \mathrm{E}_0\left[\left\{\chi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(t)\right)\left[\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]\left[\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]^{\mathrm{T}}\right.\right.\\ &+\left.\left.\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(t)\right)\;\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\eta_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}^{\mathrm{T}}\right\}w_2(\mathbf{x})\right].\end{aligned}$$

N4. The matrix $\boldsymbol{\Sigma}$ is positive definite with

$$\boldsymbol{\Sigma}=\mathrm{E}_0\left\{\Psi^2\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(t)\right)w_2^2(\mathbf{x})\left[\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]\left[\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]^{\mathrm{T}}\right\}.$$

N5. a) $\mathrm{E}_0\left\{\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(t)\right)|(\mathbf{x},t)\right\}=0$.

b) $\mathrm{E}_0\left[\left\{\chi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(\tau)\right)\left(\mathbf{x}+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(\tau)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)\right\}w_2(\mathbf{x})|t=\tau\right]=0\,.$

N6. $\mathrm{E}_0\left(w_2(\mathbf{x})\left\|\mathbf{x}+(\partial\eta_{\boldsymbol{\beta}}(\tau))/\partial\boldsymbol{\beta}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\|^2\right)<\infty.$

Remark A.2.1. **Conditions N1**a) and d) entail that for any consistent estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$, we have $\Delta_n\xrightarrow{p}0$ and $\Lambda_n\xrightarrow{p}0$ with

$$\begin{aligned}\Delta_n &= \max_{1\le j\le p}\left\|\frac{\partial}{\partial\beta_j}\widehat{\eta}_{\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}-\frac{\partial}{\partial\beta_j}\eta_{\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\|_\infty\\ \Lambda_n &= \max_{1\le j,\ell\le p}\left\|\frac{\partial^2}{\partial\beta_j\partial\beta_\ell}\widehat{\eta}_{\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}-\frac{\partial^2}{\partial\beta_j\partial\beta_\ell}\eta_{\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right\|_\infty.\end{aligned}$$

Condition **N1**b) follows from the continuity of $\eta_{\boldsymbol{\beta}}(t)=\eta(\boldsymbol{\beta},t)$ with respect to $(\boldsymbol{\beta},t)$ and Theorem 3.1 that leads to $\sup_{\boldsymbol{\beta}\in\mathcal{K}}\left\|\widehat{\eta}_{\boldsymbol{\beta}}-\eta_{\boldsymbol{\beta}}\right\|_\infty\xrightarrow{a.s.}0$.

Remark A.2.2. When the kernel K is continuously differentiable with derivative K' bounded and with bounded variation, the uniform convergence required in **N1**b) to e) can be derived through analogous arguments to those considered in Theorem

3.1 by using that

$$\begin{aligned}
\frac{\partial}{\partial t}\widehat{\eta}_{\boldsymbol{\beta}}(t) &= -\frac{\left(n{h_n}^2\right)^{-1}\sum_{i=1}^{n} K'\left((t-t_i)/h_n\right)\Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\widehat{\eta}_{\boldsymbol{\beta}}(t)\right)}{\left(nh_n\right)^{-1}\sum_{i=1}^{n} K\left((t-t_i)/h_n\right)\chi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\widehat{\eta}_{\boldsymbol{\beta}}(t)\right)} \\
\frac{\partial}{\partial \beta_j}\widehat{\eta}_{\boldsymbol{\beta}}(t) &= -\frac{\left(nh_n\right)^{-1}\sum_{i=1}^{n} K\left((t-t_i)/h_n\right)\chi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\widehat{\eta}_{\boldsymbol{\beta}}(t)\right)x_{ij}}{\left(nh_n\right)^{-1}\sum_{i=1}^{n} K\left((t-t_i)/h_n\right)\chi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\widehat{\eta}_{\boldsymbol{\beta}}(t)\right)}
\end{aligned}$$

and requiring

$$\begin{aligned}
\sup_{t\in\mathcal{T}} E\left(\sup_{\boldsymbol{\beta}\in\mathcal{K}} |\chi\left(y_1, \mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}+\eta_{\boldsymbol{\beta}}(t)\right)\|\mathbf{x}_1\| \,|t_1=t\right) &< \infty \\
\sup_{t\in\mathcal{T}} E\left(\sup_{\boldsymbol{\beta}\in\mathcal{K}} |\chi_1\left(y_1, \mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}+\eta_{\boldsymbol{\beta}}(t)\right)\|\mathbf{x}_1\| \,|t_1=t\right) &< \infty \\
\inf_{\substack{\boldsymbol{\beta}\in\mathcal{K}\\ t\in\mathcal{T}}} E\left(\chi\left(y_1, \mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}+\eta_{\boldsymbol{\beta}}(t)\right)|t_1=t\right) &> 0\,.
\end{aligned}$$

The uniform convergence rates required in **N1**c) are fulfilled when $\widehat{\eta}_{\boldsymbol{\beta}}$ is defined through (7) and a rate-optimal bandwidth is used for the kernel. 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. More precisely, assumption **N1**c) avoids the bias term and ensures that $G_n(\widehat{\eta}_{\boldsymbol{\beta}_0})$ will behave asymptotically as $G_n(\eta_{\boldsymbol{\beta}_0})$, where for any $\boldsymbol{\beta}\in I\!R^p$ and any differentiable function $\upsilon_{\boldsymbol{\beta}}(t)=\upsilon(\boldsymbol{\beta},t): I\!R^{p+1}\to I\!R$

$$G_n\left(\upsilon_{\boldsymbol{\beta}}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\upsilon_{\boldsymbol{\beta}_0}(t_i)\right)\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\upsilon_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]w_2(\mathbf{x}_i)\,.$$

Remark A.2.3. If **N4** is fulfilled then the columns of $\mathbf{x}+(\partial\eta_{\boldsymbol{\beta}}(t))/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ will not be collinear. It is necessary not to allow $\mathbf{x}$ to be predicted by t to get root-n regression estimates.

Note that for the Ψ functions considered by Bianco and Yohai (1995), Croux and Haesbroeck (2002) and Cantoni and Ronchetti (2001a), **N5**a) is satisfied. This condition is the conditional Fisher consistency property as stated in the generalized linear regression model by Künsch, Stefanski and Carroll (1989).

Note also that **N5**b) is fulfilled if $w_2 \equiv w_1$. Effectively, since $\eta_{\boldsymbol{\beta}}(\tau)$ minimizes $S(a, \boldsymbol{\beta}, \tau)$ for each τ, it satisfies

$$\mathrm{E}_0 \left[\Psi \left(y, \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(\tau) \right) w_1(\mathbf{x}) \middle| \, t = \tau \right] = 0 \, ,$$

thus, differentiating with respect to $\boldsymbol{\beta}$, we get

$$\mathrm{E}_0 \left[\chi \left(y, \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta} + \eta_{\boldsymbol{\beta}}(\tau) \right) \left(\mathbf{x} + \frac{\partial}{\partial \boldsymbol{\beta}} \eta_{\boldsymbol{\beta}}(\tau) \right) w_1(\mathbf{x}) \middle| \, t = \tau \right] = 0.$$

Moreover, either if $w_2 \equiv w_1$ or if **N5**a) holds

$$\mathbf{A} = \mathrm{E}_0 \left\{ \chi \left(y, \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}_0 + \eta_0(t) \right) \left[\mathbf{x} + \frac{\partial}{\partial \boldsymbol{\beta}} \eta_{\boldsymbol{\beta}}(t) \Big|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0} \right] \left[\mathbf{x} + \frac{\partial}{\partial \boldsymbol{\beta}} \eta_{\boldsymbol{\beta}}(t) \Big|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0} \right]^{\mathrm{T}} w_2(\mathbf{x}) \right\} .$$

Therefore, if $\Psi(y, u)$ is strictly monotone in u and $P(w_2(\mathbf{x}) > 0) = 1$, **N3** holds, i.e., $\mathbf{A}$ will be non- singular unless

$$P \left(\mathbf{a}^{\mathrm{T}} \left[\mathbf{x} + \frac{\partial}{\partial \boldsymbol{\beta}} \eta_{\boldsymbol{\beta}}(t) \Big|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0} \right] = 0 \right) = 1 \, ,$$

for some $\mathbf{a} \in I\!\!R^p$, that is, unless there is a linear combination of $\mathbf{x}$ which can be completely determined by t.

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 $\eta_{\boldsymbol{\beta}}$.

Lemma A.2.1. *Let $(y_i, \mathbf{x}_i, t_i)$ be independent observations such that $y_i | (\mathbf{x}_i, t_i) \sim F(\cdot, \mu_i)$ with $\mu_i = H(\eta_0(t_i) + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0)$ and $\mathrm{V{\scriptsize AR}}(y_i | (\mathbf{x}_i, t_i)) = V(\mu_i)$. Assume that t_i are random variables with distribution on a compact set $\mathcal{T}$ and that **N1** to **N3** and **N6** hold. Let $\widetilde{\boldsymbol{\beta}}$ be such that $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$. Then , $\mathbf{A}_n \xrightarrow{p} \mathbf{A}$ where $\mathbf{A}$ is given in **N3** and*

$$\begin{aligned} \mathbf{A}_n &= n^{-1} \sum_{i=1}^{n} \chi \left(y_i, \mathbf{x}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} + \widehat{\eta}_{\widetilde{\boldsymbol{\beta}}}(t_i) \right) \widehat{\mathbf{z}}_i \left(\widetilde{\boldsymbol{\beta}} \right) \widehat{\mathbf{z}}_i \left(\widetilde{\boldsymbol{\beta}} \right)^{\mathrm{T}} w_2(\mathbf{x}_i) \\ &+ n^{-1} \sum_{i=1}^{n} \Psi \left(y_i, \mathbf{x}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} + \widehat{\eta}_{\widetilde{\boldsymbol{\beta}}}(t_i) \right) \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \widehat{\eta}_{\boldsymbol{\beta}}(t_i) \Big|_{\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}}^{\mathrm{T}} w_2(\mathbf{x}_i) \\ \widehat{\mathbf{z}}_i \left(\widetilde{\boldsymbol{\beta}} \right) &= \mathbf{x}_i + \frac{\partial}{\partial \boldsymbol{\beta}} \widehat{\eta}_{\boldsymbol{\beta}}(t_i) \Big|_{\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}} . \end{aligned}$$

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

$$
\begin{aligned}
\mathbf{A}_n^{(1)} &= n^{-1}\sum_{i=1}^{n} \chi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} + \eta_0(t_i)\right) \mathbf{z}_i\, \mathbf{z}_i^{\mathrm{T}} w_2(\mathbf{x}_i)\\
\mathbf{A}_n^{(2)} &= n^{-1}\sum_{i=1}^{n} \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} + \eta_0(t_i)\right)\ \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}^{\mathrm{T}} w_2(\mathbf{x}_i)\\
\mathbf{A}_n^{(3)} &= n^{-1}\sum_{i=1}^{n} \chi_1\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} + \xi_{i,1}\right) \widehat{\mathrm{w}}_0(t_i)\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}} w_2(\mathbf{x}_i)\\
\mathbf{A}_n^{(4)} &= n^{-1}\sum_{i=1}^{n} \chi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} + \xi_{i,2}\right) \widehat{\mathrm{w}}_0(t_i)\ \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}^{\mathrm{T}} w_2(\mathbf{x}_i)\\
\mathbf{A}_n^{(5)} &= n^{-1}\sum_{i=1}^{n} \chi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} + \widehat{\eta}_{\widetilde{\boldsymbol{\beta}}}(t_i)\right) [\widehat{\mathbf{w}}(t_i)\mathbf{z}_i^{\mathrm{T}} + \mathbf{z}_i\widehat{\mathbf{w}}(t_i)^{\mathrm{T}} + \widehat{\mathbf{w}}(t_i)\ \widehat{\mathbf{w}}(t_i)^{\mathrm{T}}]\, w_2(\mathbf{x}_i)\\
\mathbf{A}_n^{(6)} &= n^{-1}\sum_{i=1}^{n} \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} + \widehat{\eta}_{\widetilde{\boldsymbol{\beta}}}(t_i)\right)\ \widehat{\mathbf{V}}(t_i)^{\mathrm{T}} w_2(\mathbf{x}_i)\ ,
\end{aligned}
$$

where $\xi_{i,1}$ and $\xi_{i,2}$ are intermediate points and

$$
\begin{aligned}
\mathbf{z}_i &= \mathbf{x}_i + \frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\\
\widehat{\mathrm{w}}_0(t) &= \widehat{\eta}_{\widetilde{\boldsymbol{\beta}}}(t) - \eta_0(t)\\
\widehat{\mathbf{w}}(t) &= \frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\eta}_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} - \frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\\
\widehat{\mathbf{V}}(t) &= \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} - \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\ .
\end{aligned}
$$

Using **N1**a), b) and d), **N6**, the boundness of Ψ, χ, χ_1, w_2 and ψ_2 and the fact that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, it follows easily that $\mathbf{A}_n^{(j)} \xrightarrow{p} 0$ for $3 \le j \le 6$.

It remains to show that $\mathbf{A}_n^{(1)} + \mathbf{A}_n^{(2)} \xrightarrow{p} \mathbf{A}$. which follows from Proposition A.1.1, using **N6**, the consistency of $\widetilde{\boldsymbol{\beta}}$ and the continuity of Ψ and χ. $\square$

Proof of Theorem 4.1. Let $\widehat{\boldsymbol{\beta}}$ is a solution of $F_n^1(\boldsymbol{\beta}) = 0$ defined in (15). Using a Taylor's expansion of order one we get

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{\eta}_{\widehat{\boldsymbol{\beta}}}(t_i)\right) w_2(\mathbf{x}_i)\left[\mathbf{x}_i + \frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\right]\\
&= \sum_{i=1}^{n} \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \widehat{\eta}_{\boldsymbol{\beta}_0}(t_i)\right) w_2(\mathbf{x}_i)\left[\mathbf{x}_i + \frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] + n\mathbf{A}_n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)
\end{aligned}
$$

where

$$\begin{aligned}\mathbf{A}_n &= n^{-1}\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\beta}}\left\{\Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\right)\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\right]\right\}\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} w_2(\mathbf{x}_i)\\ &=n^{-1}\sum_{i=1}^n \chi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}+\widehat{\eta}_{\widetilde{\boldsymbol{\beta}}}(t_i)\right)\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}\right]\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}\right]^{\mathrm{T}} w_2(\mathbf{x}_i)\\ &+ n^{-1}\sum_{i=1}^n \Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}+\widehat{\eta}_{\widetilde{\boldsymbol{\beta}}}(t_i)\right)\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\Big|^{\mathrm{T}}_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} w_2(\mathbf{x}_i)\,,\end{aligned}$$

with $\widetilde{\boldsymbol{\beta}}$ an intermediate point between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$. Note that in the partly linear regresion model, only the first term in $\mathbf{A}_n$ is different from 0, since $\widehat{\eta}_{\boldsymbol{\beta}}(t)$ is linear in $\boldsymbol{\beta}$.

From Lemma A.2.1, 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 enough to derive the asymptotic behaviour of

$$\widehat{L}_n = n^{-1/2}\sum_{i=1}^n \Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\widehat{\eta}_{\boldsymbol{\beta}_0}(t_i)\right)\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\eta}_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] w_2(\mathbf{x}_i)\,.$$

Let

$$L_n = n^{-1/2}\sum_{i=1}^n \Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_{\boldsymbol{\beta}_0}(t_i)\right)\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] w_2(\mathbf{x}_i)\,.$$

Using that $\eta_{\boldsymbol{\beta}_0}=\eta_0$ and since **N5** entails that $E\left[\Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_{\boldsymbol{\beta}_0}(t_i)\right)|(\mathbf{x}_i,t_i)\right]=0$, it follows 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 $\widehat{L}_n - L_n = L_n^1+L_n^2+L_n^3+L_n^4$ where

$$\begin{aligned}L_n^1 &= n^{-1/2}\sum_{i=1}^n \chi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(t_i)\right)\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] w_2(\mathbf{x}_i)\widehat{\upsilon}_0(t_i)\\ L_n^2 &= n^{-1/2}\sum_{i=1}^n \Psi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_{\boldsymbol{\beta}_0}(t_i)\right) w_2(\mathbf{x}_i)\widehat{\mathbf{v}}_0(t_i)\\ L_n^3 &= n^{-1}\sum_{i=1}^n \chi\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_0(t_i)\right) w_2(\mathbf{x}_i)\left(n^{1/4}\widehat{\mathbf{v}}_0(t_i)\right)\left(n^{1/4}\widehat{\upsilon}_0(t_i)\right)\\ L_n^4 &= (2n)^{-1}\sum_{i=1}^n \chi_1\left(y_i,\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\xi(t_i)\right)\left[\mathbf{x}_i+\frac{\partial}{\partial\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] w_2(\mathbf{x}_i)\left(n^{1/4}\widehat{\upsilon}_0(t_i)\right)^2\end{aligned}$$

where $\widehat{\upsilon}_0(t) = \widehat{\eta}_{\boldsymbol{\beta}_0}(t) - \eta_0(t)$, $\widehat{\mathbf{v}}_0(t) = (\widehat{\mathrm{v}}_{1,0}(t), \dots, \widehat{\mathrm{v}}_{p,0}(t))^{\mathrm{T}} = \partial \widehat{\upsilon}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is defined in (A.6), $\widehat{\upsilon}$ is defined in (A.5) and $\xi(t_i)$ an intermediate point between $\widehat{\eta}_{\boldsymbol{\beta}_0}(t_i)$ and $\eta_0(t_i)$. 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, we will show that $L_n^j \xrightarrow{p} 0$ for $j = 1, 2$ which will follow from **N1**c) to e) and **N5**, using similar arguments to those considered in Bianco and Boente (2004).

Effectively, fix the coordinate j, $1 \le j \le p$. For any function υ, if $x_{i,j}$ and β_j denote the $j-$th coordinate of $\mathbf{x}_i$ and $\boldsymbol{\beta}$ respectively, we define

$$\begin{aligned} J_{n,1}(\upsilon) &= n^{-1/2} \sum_{i=1}^n \chi\left(y_i, \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \eta_0(t_i)\right) \left[x_{i,j} + \frac{\partial}{\partial \beta_j} \eta_{\boldsymbol{\beta}}(t_i)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] w_2(\mathbf{x}_i)\, \upsilon(t_i) \\ J_{n,2}(\upsilon) &= n^{-1/2} \sum_{i=1}^n \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \eta_0(t_i)\right) w_2(\mathbf{x}_i)\, \upsilon(t_i) \end{aligned}$$

where we have omitted the subscript j for the sake of simplicity.

Let $\mathcal{V} = \{\upsilon \in \mathcal{C}^1(\mathcal{T}) : \|\upsilon\|_\infty \le 1 \quad \|\upsilon'\|_\infty \le 1\}$. Note that, for any probability measure $\mathcal{Q}$, the bracketing number $N_{[\,]}\left(\epsilon, \mathcal{V}, L^2(\mathcal{Q})\right)$, and so the covering number $N\left(\epsilon, \mathcal{V}, L^2(\mathcal{Q})\right)$, satisfy

$$\log N\left(\epsilon/2, \mathcal{V}, L^2(\mathcal{Q})\right) \le \log N_{[\,]}\left(\epsilon, \mathcal{V}, L^2(\mathcal{Q})\right) \le K \epsilon^{-1} ,$$

for $0 < \epsilon < 2$, where the constant K is independent of the probability measure $\mathcal{Q}$ (see Corollary 2.7.2 in van der Vaart and Wellner (1996)).

Consider the classes of functions

$$\begin{aligned} &\mathcal{F}_1 = \left\{ f_{1,\upsilon}(y, \mathbf{x}, t) = \chi\left(y, \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}_0 + \eta_0(t)\right) \left[x_j + \frac{\partial}{\partial \beta_j} \eta_{\boldsymbol{\beta}}(t)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] w_2(\mathbf{x})\, \upsilon(t) \ , \ \upsilon \in \mathcal{V} \right\} \\ &\mathcal{F}_2 = \left\{ f_{2,\upsilon}(y, \mathbf{x}, t) = \Psi\left(y, \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}_0 + \eta_{\boldsymbol{\beta}_0}(t)\right) w_2(\mathbf{x})\, \upsilon(t) \ , \ \upsilon \in \mathcal{V} \right\} \ . \end{aligned}$$

$\mathcal{F}_1$ and $\mathcal{F}_2$ have enveloppes the constants

$$A_1 = \|\chi\|_\infty \left[\|\psi_2\|_\infty + \left\| (\partial \eta_{\boldsymbol{\beta}})/\partial \beta_j |_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right\|_\infty \|w_2\|_\infty \right]$$

and $A_2 = \|\Psi\|_\infty \|w_2\|_\infty$, respectively. On the other hand, **N5** implies that, for any $f \in \mathcal{F}_1 \cup \mathcal{F}_2$, $E\, f\left(y_i, \mathbf{x}_i, t_i\right) = 0$.

Denote $\|f\|_{\mathbb{Q},2} = \left(E_{\mathbb{Q}}(f^2)\right)^{1/2}$. It is easy to see that, given $\upsilon \in \mathcal{V}$, for any $0 < \epsilon < 2$, $\|\upsilon_s - \upsilon\|_{\mathbb{Q},2} < \epsilon$ entail that

$$\begin{aligned} \|f_{1,\upsilon_s} - f_{1,\upsilon}\|_{\mathbb{Q},2} &\leq A_1\,\epsilon \\ \|f_{2,\upsilon_s} - f_{2,\upsilon}\|_{\mathbb{Q},2} &\leq A_2\,\epsilon \end{aligned}$$

and so,

$$\begin{aligned} N\left(\epsilon\, A_1, \mathcal{F}_1, L^2(\mathcal{Q})\right) &\leq N\left(\epsilon, \mathcal{V}, L^2(\mathcal{Q})\right) \\ N\left(\epsilon\, A_2, \mathcal{F}_2, L^2(\mathcal{Q})\right) &\leq N\left(\epsilon, \mathcal{V}, L^2(\mathcal{Q})\right) \,. \end{aligned}$$

Therefore, these classes of functions have finite uniform–entropy.

For any class of functions $\mathcal{F}$, denote $\mathcal{J}(\delta, \mathcal{F})$ the integral

$$\mathcal{J}(\delta, \mathcal{F}) = \sup_{\mathcal{Q}} \int_0^\delta \sqrt{1 + \log\left(N\left(\epsilon\,\|F\|_{\mathbb{Q},2}, \mathcal{F}, L^2(\mathcal{Q})\right)\right)}\, d\epsilon \,,$$

where the supremum is taken over all discrete probability measures $\mathcal{Q}$ with $\|F\|_{\mathbb{Q},2} > 0$ and F is the enveloppe of $\mathcal{F}$. The function $\mathcal{J}$ is increasing, $\mathcal{J}(0, \mathcal{F}) = 0$ and $\mathcal{J}(1, \mathcal{F}) < \infty$ and $\mathcal{J}(\delta, \mathcal{F}) \to 0$ as $\delta \to 0$ for classes of functions $\mathcal{F}$ which satisfies the uniform–entropy condition. Moreover, if $\mathcal{F}_0 \subset \mathcal{F}$ and the enveloppe F is used for $\mathcal{F}_0$, then $\mathcal{J}(\delta, \mathcal{F}_0) \leq \mathcal{J}(\delta, \mathcal{F})$.

For any $\epsilon > 0$ and $0 < \delta < 1$, consider the subclasses

$$\begin{aligned} \mathcal{F}_{1,\delta} &= \{f_{1,\upsilon}(y, \mathbf{x}, t) \in \mathcal{F}_1 \text{ with } \|\upsilon\|_\infty < \delta\} \subset \mathcal{F}_1 \\ \mathcal{F}_{2,\delta} &= \{f_{2,\upsilon}(y, \mathbf{x}, t) \in \mathcal{F}_2 \text{ with } \|\upsilon\|_\infty < \delta\} \subset \mathcal{F}_2 \,. \end{aligned}$$

Remind that $\widehat{\upsilon}_0(t) = \widehat{\eta}_{\boldsymbol{\beta}_0}(t) - \eta_0(t)$ and $\widehat{\mathrm{v}}_{j,0}(t) = (\partial \widehat{\upsilon}(\boldsymbol{\beta}, t))/\partial \beta_j|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0}$. Using that **N1**c) and e) entail that

$$\begin{aligned} &\sup_{t \in \mathcal{T}} |\widehat{\upsilon}_0(t)| \xrightarrow{p} 0, &&\sup_{t \in \mathcal{T}} |\frac{\partial}{\partial t} \widehat{\upsilon}_0(t)| \xrightarrow{p} 0 \,, \\ &\sup_{t \in \mathcal{T}} |\widehat{\mathrm{v}}_{j,0}(t)| \xrightarrow{p} 0, &&\sup_{t \in \mathcal{T}} |\frac{\partial}{\partial t} \widehat{\mathrm{v}}_{j,0}(t)| \xrightarrow{p} 0 \,, \end{aligned}$$

we have that, for n large enough, $P\left(\widehat{\upsilon}_0 \in \mathcal{V} \text{ and } \|\widehat{\upsilon}_0\|_\infty < \delta\right) > 1 - \delta/2$ and $P\left(\widehat{\mathrm{v}}_{j,0} \in \mathcal{V} \text{ and } \|\widehat{\mathrm{v}}_{j,0}\|_\infty < \delta\right) > 1 - \delta/2$, for $1 \leq j \leq p$.

It is clear that

$$\begin{aligned}
\sup_{f\in\mathcal{F}_{1,\delta}} n^{-1}\sum_{i=1}^n f^2(r_i,\mathbf{z}_i,t_i) &\leq A_1^2\delta^2 \\
\sup_{f\in\mathcal{F}_{2,\delta}} n^{-1}\sum_{i=1}^n f^2(r_i,\mathbf{z}_i,t_i) &\leq A_2^2\,\delta^2\ .
\end{aligned}$$

Therefore, the maximal inequality for covering numbers entails that, for any $0\leq \ell\leq p$,

$$\begin{aligned}
P\left(|J_{n,1}\left(\widehat{\upsilon}_0\right)| > \epsilon\right) &\leq P\left(|J_{n,1}\left(\widehat{\upsilon}_0\right)| > \epsilon\,,\ \widehat{\upsilon}_0\in\mathcal{V} \text{ and } \|\widehat{\upsilon}_0\|_\infty < \delta\right)+\delta \\
&\leq P\left(\sup_{f\in\mathcal{F}_{1,\delta}}\left|n^{-1/2}\sum_{i=1}^n f(y_i,\mathbf{x}_i,t_i)\right| > \epsilon\right)+\delta \\
&\leq \epsilon^{-1}E\left(\sup_{f\in\mathcal{F}_{1,\delta}}\left|n^{-1/2}\sum_{i=1}^n f(r_i,\mathbf{z}_i,t_i)\right|\right)+\delta \\
&\leq \epsilon^{-1}D_1\,A_1\,\mathcal{J}\left(\delta,\mathcal{F}_1\right)+\delta\ ,
\end{aligned}$$

where D_1 is a constant not depending on n.

Similarly, $P\left(|J_{n,2}\left(\widehat{\mathrm{v}}_{j,0}\right)| > \epsilon\right) \leq \epsilon^{-1}D_2\,A_2\,\mathcal{J}\left(\delta,\mathcal{F}_2\right)+\delta$. Using that the classes $\mathcal{F}_1$ and $\mathcal{F}_2$ satisfy the uniform–entropy condition, we get $\lim_{\delta\to 0}\mathcal{J}\left(\delta,\mathcal{F}_1\right)=0$ and $\lim_{\delta\to 0}\mathcal{J}\left(\delta,\mathcal{F}_2\right)=0$. Thus, we have that $L_n^1 = J_{n,1}\left(\widehat{\upsilon}_0\right)\xrightarrow{p} 0$ and $L_n^2 = \left(J_{n,2}\left(\widehat{\mathrm{v}}_{1,0}\right),\ldots,J_{n,2}\left(\widehat{\mathrm{v}}_{p,0}\right)\right)^{\mathrm{T}}\xrightarrow{p} 0$, as desired. □

Acknowledgment. We wish to thank the Associate Editor and two anonymous referees for valuable comments which led to an improved version of the original paper.

References

Bianco, A. and Boente, G. (1996). Robust nonparametric generalized regression estimation. *Impresiones Previas del Departamento de Matemática*, FCEN, September 1996.

Bianco, A. and Boente, G. (2002). On the asymptotic behavior of one-step estimates in heteroscedastic regression models. *Statist. Probab. Letters* **60** 33-47.

Bianco, A. and Boente, G. (2004). Robust estimators in semiparametric partly linear regression models. *J. Statist. Planning and Inference* **122** 229-252.

Bianco, A., García Ben, M. and Yohai, V. (2005). Robust estimation for linear regression with asymmetric errors. *Canad. J. Statist.* **33** 1-18.

Bianco, A. and Yohai, V. (1995). Robust estimation in the logistic regression model. *Lecture Notes in Statistics* **109** 17-34. Springer–Verlag, New York.

Boente, G. and Fraiman, R. (1991). Strong order of convergence and asymptotic distribution of nearest neighbor density estimates from dependent observations. *Sankhya* **53**, Serie A, 194-205.

Boente, G., Fraiman, R. and Meloche, J. (1997). Robust plug-in bandwidth estimators in nonparametric regression. *J. Statist. Planning and Inference* **57** 109-142.

Cantoni, E. and Ronchetti, E. (2001a). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96** 1022-1030.

Cantoni, E. and Ronchetti, E. (2001b). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing* **11** 141-146.

Carroll, R., Fan, J., Gijbels, I. and Wand, M. (1997). Generalized partially linear single-index models. *J. Amer. Statis. Assoc.* **92** 477-489.

Croux, C. and Haesbroeck, G. (2002). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis* **44** 273-295.

Gao, J. and Shi, P. (1997). M–type smoothing splines in nonparametric and semiparametric regression models. *Statistica Sinica* **7** 1155-1169.

He, X., Fung, W.K. and Zhu, Z.Y. (2005). Robust estimation in generalized partial linear models for Clustered Data. To appear in *Journal of the American Statistical Association.*

He, X., Zhu, Z. and 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.

Künsch, H., Stefanski, L. and Carroll, R. (1989). Conditionally unbiased bounded-influence estimation in general regression models with applications to generalized linear models. *J. Amer. Statist. Assoc.* **84** 460-466.

Leung, D., Marrot, F. and Wu, E. (1993). Bandwidth selection in robust smoothing. *J. Nonparametr. Statist.* **4** 333–339.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models.* Chapman and Hall, London.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer–Verlag, New York.

Robinson, P. (1988). Root-n-consistent Semiparametric regression. *Econometrica* **56** 931-954.

Severini, T. and Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501-511.

Severini, T. and Wong, W. (1992). Generalized profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768-1802.

Stefanski, L., Carroll, R. and Ruppert, D. (1986). Bounded score functions for generalized linear models. *Biometrika* **73** 413-424.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer–Verlag, New York.

Wang, F. and Scott, D. (1994). The L_1 method for robust nonparametric regression. *J. Amer. Stat. Assoc.* **89** 65-76.

	Bias($\hat{\beta}$)	SD($\hat{\beta}$)	MSE($\hat{\beta}$)	MSE($\hat{\eta}$)
QAL(0.1)	0.059	0.219	0.051	0.111
QAL(0.2)	0.033	0.214	0.047	0.073
QAL(0.3)	0.004	0.220	0.048	0.152
RQL(0.1)	−0.051	0.242	0.061	0.114
RQL(0.2)	−0.054	0.254	0.067	0.089
RQL(0.3)	−0.105	0.262	0.080	0.154
MOD(0.1)	0.030	0.252	0.064	0.143
MOD(0.2)	0.018	0.251	0.063	0.088
MOD(0.3)	−0.001	0.252	0.064	0.135

Table 1: Summary Results for Study 1.

	QAL	RQL	MOD
Original data	2.02	2.08	1.99
$x_1 = 10, y_1 = 0$	0.90	2.07	2.00
$x_2 = -10, y_2 = 10$	0.31	2.06	1.97
$x_3 = -10, y_3 = 10$	0.12	2.05	1.95

Table 2: Estimates of β (true value of 2) in Study 2. (x_i, y_i), $1 \leq i \leq 3$ denote the three contaminating points which replace the first three observations one by one.

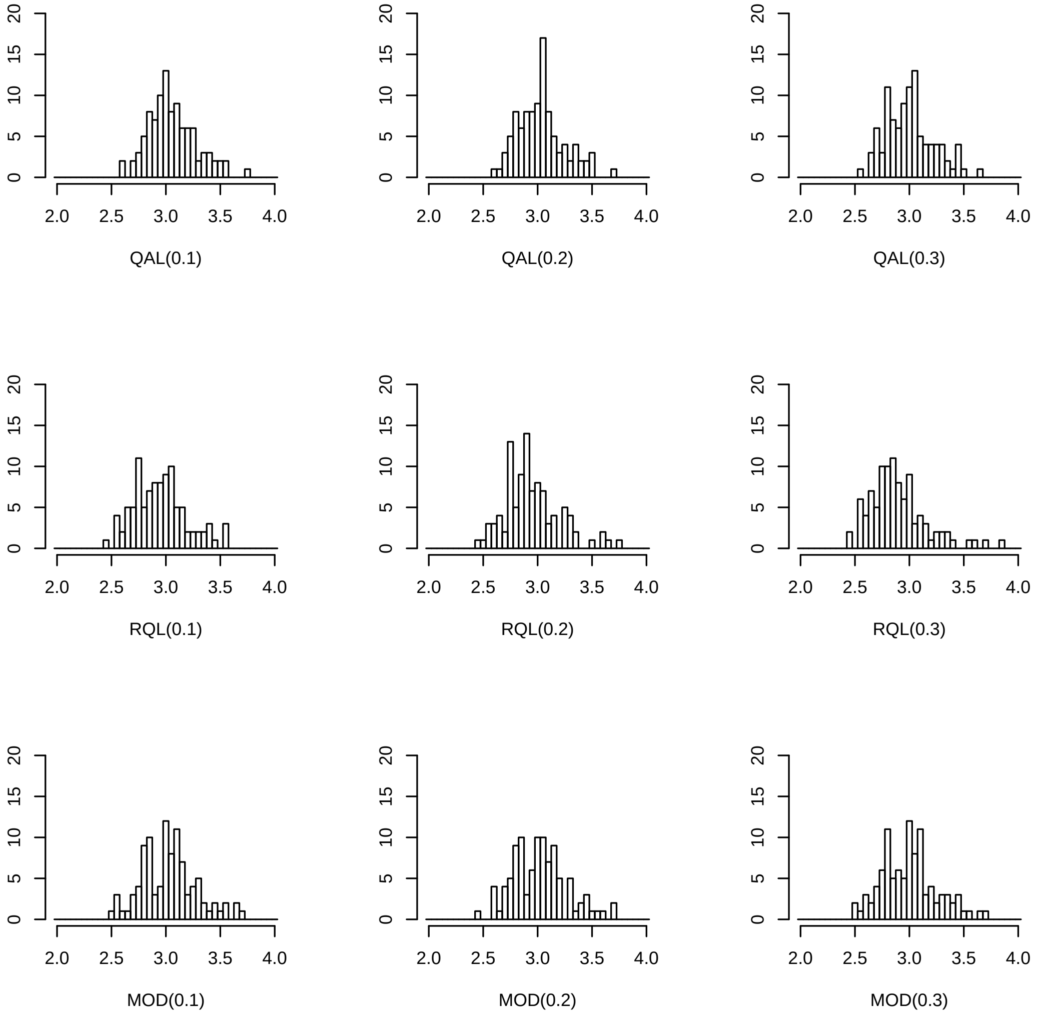

Figure 1: Histograms of $\widehat{\boldsymbol{\beta}}$ for QAL, RQL and MOD using bandwidths $h_n = 0.1$, 0.2 and 0.3 .

Data	Estimator	Bias($\hat{\beta}$)	SD($\hat{\beta}$)	MSE($\hat{\beta}$)	MSE ($\hat{\eta}$)
Original	QAL	0.126	0.357	0.143	0.297
Original	RQL	0.199	0.409	0.207	0.348
Original	MOD	0.158	0.386	0.174	0.317
Contaminated C_1	QAL	−0.393	0.366	0.288	0.378
Contaminated C_1	RQL	−0.171	0.440	0.223	0.378
Contaminated C_1	MOD	−0.245	0.414	0.231	0.365
Contaminated C_2	QAL	−0.935	0.287	0.957	0.446
Contaminated C_2	RQL	0.018	0.545	0.297	0.399
Contaminated C_2	MOD	−0.237	0.436	0.246	0.350
Contaminated C_3	QAL	−2.187	0.071	4.788	0.402
Contaminated C_3	RQL	0.177	0.430	0.216	0.400
Contaminated C_3	MOD	−0.037	0.475	0.227	0.369

Table 3: Summary Results for Study 3.

Data	Ratio	MSE($\hat{\beta}$)	MSE($\hat{\eta}$)
Original	QAL/RQL	0.691	0.853
Original	QAL/MOD	0.822	0.937
Contaminated C_1	QAL/RQL	1.291	1.000
Contaminated C_1	QAL/MOD	1.247	1.036
Contaminated C_2	QAL/RQL	3.222	1.118
Contaminated C_2	QAL/MOD	3.890	1.274
Contaminated C_3	QAL/RQL	22.167	1.005
Contaminated C_3	QAL/MOD	21.093	1.089

Table 4: Summary Results for Study 3. Ratio of MSE