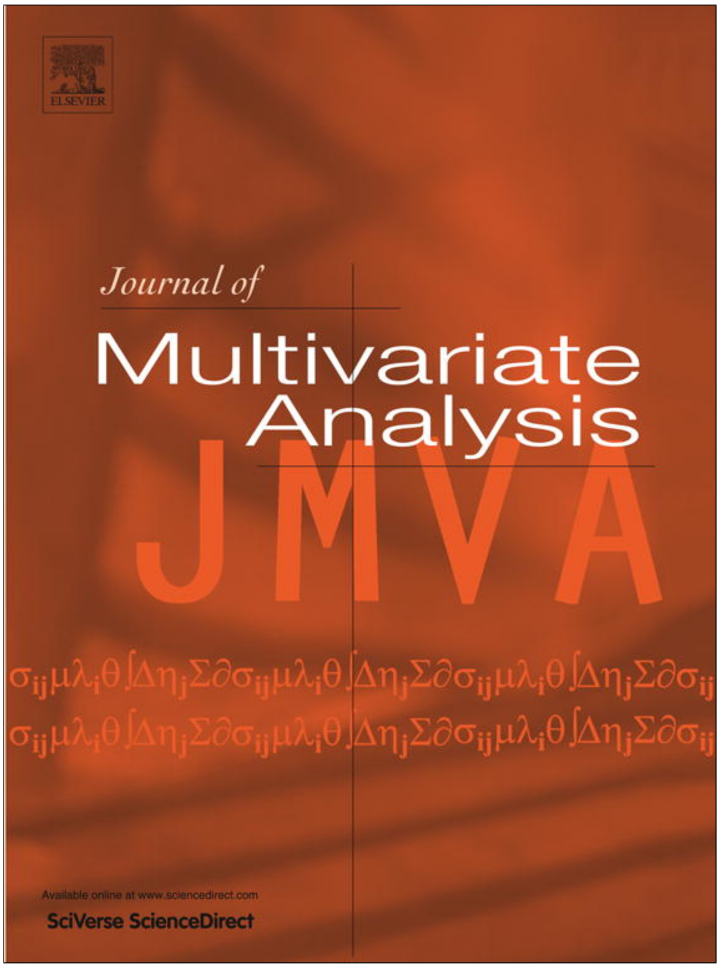
ELSEVIER
Journal of
Multivariate Analysis
JMVA
Available online at www.sciencedirect.com
SciVerse ScienceDirect

Journal of Multivariate Analysis 114 (2013) 209–226

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Resistant estimators in Poisson and Gamma models with missing responses and an application to outlier detection

Ana M. Bianco [a], Graciela Boente [a,*], Isabel M. Rodrigues [b]

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

[b] *Departamento de Matemática and CEMAT, Instituto Superior Técnico, Technical University of Lisbon (TULisbon), Lisboa, Portugal*

ARTICLE INFO

Article history:
Received 29 August 2011
Available online 16 August 2012

AMS subject classifications:
62F35
62F10

Keywords:
Fisher-consistency
Generalized linear models
Missing data
Outliers
Robust estimation

ABSTRACT

When dealing with situations in which the responses are discrete or show some type of asymmetry, the linear model is not appropriate to establish the relation between the responses and the covariates. Generalized linear models serve this purpose, since they allow one to model the mean of the responses through a link function, linearly on the covariates. When atypical observations are present in the sample, robust estimators are useful to provide fair estimations as well as to build outlier detection rules. The focus of this paper is to define robust estimators for the regression parameter when missing data possibly occur in the responses. The estimators introduced turn out to be consistent under mild conditions. In particular, resistant methods for Poisson and Gamma models are given. A simulation study allows one to compare the behaviour of the classical and robust estimators, under different contamination schemes. The robustness of the proposed procedures is studied through the influence function, while asymptotic variances are derived from it. Besides, outlier detection rules are defined using the influence function. The procedure is also illustrated by analysing a real data set.

1. Introduction

The generalized linear model (GLM), introduced by Nelder and Wedderburn [21], underlies most of the statistical analyses that are used in applied and social research. The fact that the standard linear model does not handle nonnormal responses, y, motivated in recent decades a great development of GLM, which can represent discrete or continuous responses even with an asymmetric behaviour. The GLM provides a unified approach that enables one to model categorical, binary, Poisson and other response types.

The GLM assumes that the observations $(y_i, \mathbf{x}_i^{\mathrm{T}})$, $1 \le i \le n$, $\mathbf{x}_i \in \mathbb{R}^k$, are independent with the same distribution as $(y, \mathbf{x}^{\mathrm{T}}) \in \mathbb{R}^{k+1}$ such that the conditional distribution of $y|\mathbf{x}$ belongs to the canonical exponential family $\exp\{[y\theta(\mathbf{x}) - B(\theta(\mathbf{x}))]/A(\tau) + C(y, \tau)\}$, for known functions A, B and C. In this situation, if we denote by B' the derivative of B, the mean $\mu(\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = B'(\theta(\mathbf{x}))$ is modelled linearly through a known link function, g, i.e., $g(\mu(\mathbf{x})) = \theta(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. For more details, see [20].

In this setting, the classical estimators are based on the minimization of the deviance, which is equivalent to the maximum likelihood method. It is very well known that these procedures can be affected by anomalous observations. To overcome this problem, robust methods have been developed and among others we can cite those proposed by Stefanski et al. [25], Künsch et al. [18], Bianco and Yohai [4], Cantoni and Ronchetti [6,7], Croux and Haesbroeck [8] and Bianco et al. [3], see also, [19]. Even though developing robust methods for GLM has been an active research area in recent decades, all these methods were designed for complete data sets.

* Correspondence to: Instituto de Cálculo, FCEyN, UBA, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina.
E-mail addresses: abianco@dm.uba.ar (A.M. Bianco), gboente@dm.uba.ar (G. Boente), irodrig@math.ist.utl.pt (I.M. Rodrigues).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2012.08.008

In practice, missing data can arise and thus, standard procedures are no longer a useful tool. Indeed, missing responses may be introduced just by design, as in the case of two-stage studies, or simply by chance. In some cases, the responders may refuse to answer, for instance about some private issues, or the responses y's may be an expensive measure to be obtained. In other cases, missing data may be caused by some loss of information due to uncontrollable factors or by failure on recording the correct information. In this paper, we will focus our attention on robust inference when the response variable has randomly missing observations, but the covariate $\mathbf{x}$ is totally observed.

We introduce robust procedures to estimate the parameter $\boldsymbol{\beta}$ under a GLM. These estimators include, when there are no missing data, some of the estimators previously studied. The proposed robust estimators of $\boldsymbol{\beta}$ are consistent under mild assumptions.

The paper is organized as follows. The robust proposals are given in Section 2 and consistency results are provided in Section 3. To measure robustness with respect to single outliers, the influence function is studied in Section 4, where we include some comments regarding the breakdown point. Also, a diagnostic measure to detect outliers is defined using the influence function. The results of a Monte Carlo study are summarized in Section 5, while a real data set is analysed in Section 6. Proofs are relegated to the Appendix.

2. Robust inference

2.1. The robust estimators

Let us consider a random sample of incomplete data $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i\right)$, $1 \leq i \leq n$, following a generalized linear model where $\delta_i = 1$ if y_i is observed, $\delta_i = 0$ if y_i is missing. The responses and covariates $(y_i, \mathbf{x}_i^{\mathrm{T}}) \in \mathbb{R}^{k+1}$ are such that $y_i|\mathbf{x}_i \sim F(\cdot, \mu_i, \tau)$ with $\mu_i = H(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\mathrm{VAR}(y_i|\mathbf{x}_i) = A^2(\tau)V^2(\mu_i) = A^2(\tau)B''(\theta(\mathbf{x}_i))$ with B'' the second derivative of B. The parameter τ usually lies on a subset of $\mathbb{R}$, for that reason we will assume that $\tau \in \mathcal{T}$, where $\mathcal{T} \subset \mathbb{R}$ stands for an open set. Let $(\boldsymbol{\beta}, \tau) \in \mathbb{R}^k \times \mathcal{T}$ denote the true parameter values and $\mathbb{E}_F$ the expectation under the true model, thus $\mathbb{E}_F(y|\mathbf{x}) = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. In a more general situation, we will think of τ as a nuisance parameter such as the tuning constant for the score function to be considered. For instance, under a Gamma regression model τ is related to the shape parameter, while for Poisson and logistic regression $\tau = 1$.

Let $(y, \mathbf{x}^{\mathrm{T}}, \delta)$ be a random vector with the same distribution as $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i\right)$. As mentioned in the Introduction, our aim is to define robust estimators of the regression parameter when missing responses occur. 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 $\mathbf{x}$, i.e.,

$$P\left(\delta = 1|(y, \mathbf{x})\right) = P\left(\delta = 1|\mathbf{x}\right) = p\left(\mathbf{x}\right). \tag{1}$$

Usually, when considering propensity estimators, it is assumed that $\inf_{\mathbf{x}} p\left(\mathbf{x}\right) > 0$, which means that at any values of the covariate response variables are observed. This assumption can be avoided by introducing a weight function with bounded support at the cost of some loss of efficiency.

Let $w_1 : \mathbb{R}^k \to \mathbb{R}$ be a weight function to control leverage points on the carriers $\mathbf{x}$ and $\rho : \mathbb{R}^2 \times \mathcal{T} \to \mathbb{R}$ a loss function. Define

$$S_n(\mathbf{b}, t) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i), \tag{2}$$

$$S(\mathbf{b}, t) = \mathbb{E}_F\left[\delta \rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right] = \mathbb{E}_F\left[p(\mathbf{x})\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right]. \tag{3}$$

Let us assume that $w_1(\cdot)$ and $\rho(\cdot)$ are such that $S(\boldsymbol{\beta}, \tau) = \min_{\mathbf{b}} S(\mathbf{b}, \tau)$, $\boldsymbol{\beta}$ being the unique minimum (see Remark 2.1 below), i.e., we are assuming Fisher-consistency. Then in order to estimate $\boldsymbol{\beta}$ one can minimize $S_n(\mathbf{b}, \tau)$ that provides a consistent estimator of $S(\mathbf{b}, \tau)$. Moreover, note that, except from the multiplicative factor $p(\mathbf{x})$, $S(\mathbf{b}, t)$ corresponds to the asymptotic version of a general M-estimator in the complete data set situation. This suggests that the presence of the missing probability may introduce some bias in the estimation of the regression parameter when considering the finite-sample version $S_n(\mathbf{b}, t)$.

Let $\widehat{\tau} = \widehat{\tau}_n$ be robust consistent estimators of τ. The *robust simplified estimator*, $\widehat{\boldsymbol{\beta}}$, of the regression parameter uses the portion of the sample with complete information and is defined as

$$\widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}}{\operatorname{argmin}}\, S_n(\mathbf{b}, \widehat{\tau}). \tag{4}$$

When ρ is continuously differentiable, if we denote by $\Psi\left(y, u, t\right) = \partial\rho(y, u, t)/\partial u$, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ satisfy the differentiated equations $S^{(1)}(\boldsymbol{\beta}, \tau) = 0$ and $S_n^{(1)}(\mathbf{b}, \widehat{\tau}) = 0$, respectively, where

$$S^{(1)}(\mathbf{b}, t) = \mathbb{E}_F\left(\Psi\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})p(\mathbf{x})\mathbf{x}\right) \quad \text{and} \quad S_n^{(1)}(\mathbf{b}, t) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i)\mathbf{x}_i.$$

When $S_n(\mathbf{b}, \widehat{\tau})$ has only one critical point, i.e., when the equation $S_n^{(1)}(\mathbf{b}, \widehat{\tau}) = 0$ has only one root, corresponding to the minimum of $S_n(\mathbf{b}, \widehat{\tau})$, the estimator $\widehat{\boldsymbol{\beta}}$ can be computed using a Newton–Raphson approach.

To improve the finite-sample bias caused in the estimation by the missing mechanism, robust propensity score estimators may be considered including an estimator of the missingness probability. Denote by $\widehat{p}(\mathbf{x})$ any estimator of $p(\mathbf{x})$. For instance, if we assume that the missingness probability is given by the logistic model, i.e., that $p(\mathbf{x}) = G_{\mathrm{L}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda}_0)$ where $G_{\mathrm{L}}(s) = (1+e^{-s})^{-1}$ is the logistic distribution function, we only need to estimate the parameter $\boldsymbol{\lambda}$ to define the estimator $\widehat{p}(\mathbf{x})$. Let $\mathcal{P} = \{q : \mathbb{R}^k \to \mathbb{R}$ such that $0 < q(\mathbf{x}) \le 1\}$, and define $S_{\mathrm{P},n} : \mathbb{R}^k \times \mathcal{T} \times \mathcal{P} \to \mathbb{R}$ and its related functional $S_{\mathrm{P}} : \mathbb{R}^k \times \mathcal{T} \times \mathcal{P} \to \mathbb{R}$ as

$$S_{\mathrm{P},n}(\mathbf{b}, t, q) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{q(\mathbf{x}_i)} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i), \tag{5}$$

$$S_{\mathrm{P}}(\mathbf{b}, t, q) = \mathbb{E}_F\left[\frac{\delta}{q(\mathbf{x})}\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right] = \mathbb{E}_F\left[\frac{p(\mathbf{x})}{q(\mathbf{x})}\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right]. \tag{6}$$

The *robust propensity score estimator* $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$ is defined as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{P}} = \underset{\mathbf{b}}{\operatorname{argmin}}\, S_{\mathrm{P},n}(\mathbf{b}, \widehat{\tau}_{\mathrm{P}}, \widehat{p}), \tag{7}$$

where $\widehat{\tau}_{\mathrm{P}}$ is a robust consistent estimator of τ, possibly different than the one considered in (4). Note that τ and $q(\mathbf{x})$ play now the role of nuisance parameters. Moreover, it is worth noticing that $S_{\mathrm{P}}(\mathbf{b}, t, p) = \mathbb{E}_F\left[\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right]$, so it corresponds to the objective function when the sample contains no missing responses. Throughout this paper we will assume Fisher-consistency, which means that $\boldsymbol{\beta}$ is the unique minimum of $S_{\mathrm{P}}(\mathbf{b}, \tau, p)$.

As above, when ρ is continuously differentiable, if we denote by $\Psi(y, u, t) = \partial\rho(y, u, t)/\partial u$, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$ satisfy the differentiated equations $S_{\mathrm{P}}^{(1)}(\boldsymbol{\beta}, \tau, p) = 0$ and $S_{\mathrm{P},n}^{(1)}(\mathbf{b}, \widehat{\tau}, \widehat{p}) = 0$, respectively, where

$$S_{\mathrm{P}}^{(1)}(\mathbf{b}, t, q) = \mathbb{E}_F\left(\Psi\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\mathbf{x}\right) \quad \text{and} \quad S_{\mathrm{P},n}^{(1)}(\mathbf{b}, t, q) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{q(\mathbf{x}_i)}\Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i)\mathbf{x}_i.$$

2.2. The loss functions used in the estimation procedure

In the frame of GLM, two families of loss functions ρ have been considered in the literature. The first one aims to bound the deviances, while the second one introduced by Cantoni and Ronchetti [6] bounds the Pearson residuals. For the sake of completeness, we recall their definition.

Let ϕ_t be a bounded nondecreasing function with continuous derivative φ_t, t being the tuning constant, and denote by $f(\cdot, s)$ the density of the distribution function $F(\cdot, s)$ with $y|\mathbf{x} \sim F\left(\cdot, H\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)$. For families of distributions that can be transformed to avoid an extra parameter in the model, the first class of loss functions takes the form of

$$\rho(y, u, t) = \phi_t[-\ln f(y, H(u)) + D(y)] + G(H(u)). \tag{8}$$

To avoid triviality, it is assumed that ϕ_t is non–constant in a positive probability set. Typically, ϕ_t is a function performing like the identity function in a neighbourhood of 0. The function $D(y)$ is typically used to remove a term from the log–likelihood that is independent of the parameter, and can be defined as $D(y) = \ln(f(y, y))$ in order to get the deviance. The correction term G is used to guarantee the Fisher-consistency and satisfies

$$G'(s) = \int \varphi_t[-\ln f(y, s) + D(y)]f'(y, s)d\mu(y) = \mathbb{E}_s\left(\varphi_t[-\ln f(y, s) + D(y)]f'(y, s)/f(y, s)\right),$$

where $\mathbb{E}_s$ indicates expectation taken under $y \sim F(\cdot, s)$ and $f'(y, s)$ is a shorthand for $\partial f(y, s)/\partial s$. Note that under a generalized linear model, the maximum likelihood estimator corresponds to the choice $\phi(s) = s$, $D(y) = \ln(f(y, y))$, $G(u) = 0$ and $w_1 \equiv 1$.

For the Poisson and logistic regression models, we have that $\tau = 1$ so, τ does not need to be estimated. For the logistic model, Bianco and Yohai [4] introduced the score function

$$\phi_c(s) = \begin{cases} s - s^2/(2c) & \text{if } s \le c \\ c/2 & \text{otherwise.} \end{cases} \tag{9}$$

In order to guarantee existence of solution, Croux and Haesbroeck [8] proposed using the score function

$$\phi_c(s) = \begin{cases} s\exp(-\sqrt{c}) & \text{if } s \le c \\ -2(1+\sqrt{s})\exp(-\sqrt{s}) + \left(2(1+\sqrt{c}) + c\right)\exp(-\sqrt{c}) & \text{otherwise.} \end{cases}$$

It is worth noting that, when considering the deviance and a continuous family of distributions with strongly unimodal density function, the correction term G is not needed, as discussed in [3]. In this case, τ may play the role of the tuning constant and, for instance, for the Gamma distribution it depends on the shape parameter, so initial estimators need to be considered.

The second class of loss functions is based on the proposal given by Cantoni and Ronchetti [6] for generalized linear models, where they consider a general class of Mallows type M-estimators, by bounding separately the influence of deviations on y and $\mathbf{x}$. Their approach is based on robustifying the quasi–likelihood, which is an alternative to the generalizations given for the GLM by Stefanski et al. [25] and Künsch et al. [18]. Let $r(y, \mu, \tau) = (y - \mu)/(V(\mu)A(\tau))$ be the Pearson residuals with $\text{VAR}(y_i|\mathbf{x}_i) = A^2(\tau)V^2(\mu_i)$. Denote $\nu(y, \mu, \tau) = \psi_c(r(y, \mu, \tau))/(V(\mu)A(\tau))$, with ψ_c an odd nondecreasing score function with tuning constant c, such as the Huber function, and $\rho(y, u, t) = -[\int_{s_0}^{H(u)} \nu(y, s, t)ds + G(H(u))]$, where s_0 is such that $\nu(y, s_0, \tau) = 0$. To ensure Fisher-consistency, the correction term $G(s)$ satisfies $G'(s) = -\mathbb{E}_s(\nu(y, s, \tau))$. For the Binomial and Poisson families, explicit forms of the correction term $G(s)$ are given in [6], while for the Gamma family with log link, an expression for $G(s)$ is provided in [7]. The classical counterpart of this approach corresponds to the choice $\psi_c(u) = u$, $w_1 \equiv 1$.

Remark 2.1. The correction factor, denoted $G(s)$, is included to guarantee Fisher-consistency under the true model. Otherwise, one can only ensure that the estimators will be consistent to the solution $\boldsymbol{\beta}(F)$ of the related functional equations, i.e., to $\boldsymbol{\beta}(F) = \text{argmin}_{\mathbf{b}} S(\mathbf{b}, \tau)$, where $S(\mathbf{b}, t)$ is defined in (3). On the other hand, as is well known, when $H(u) = u$, i.e., under the linear regression model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$, Fisher-consistency holds if, for instance, the errors ϵ_i have a symmetric distribution and the score function ψ_c is odd.

Under a logistic regression model, Fisher-consistency can easily be derived for the loss function given by (8), when ϕ satisfies the regularity conditions stated in [4] and

$$P\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = \alpha\right) < 1, \quad \forall(\boldsymbol{\beta}, \alpha) \neq 0. \tag{10}$$

Moreover, it is easy to verify that $\boldsymbol{\beta}$ is the unique minimizer of $S(\mathbf{b}, \tau)$ in this case. The same assertion can be verified for the robust quasi–likelihood proposal if ψ_c is bounded and increasing.

For a Gamma regression model with a fixed shape parameter, Fisher-consistency for the regression parameter will be derived below when (10) holds and the score function ϕ is bounded and strictly increasing on the set where it is not constant.

2.3. The Poisson regression model

In the case of the Poisson distribution with parameter μ the density can be written as

$$f(y, \mu) = \begin{cases} \exp(-\mu)\mu^y/y! & y \in N \cup \{0\} \\ 0 & \text{in other case} \end{cases}$$

with $\mathbb{E}_\mu(y) = \mu$, $\text{VAR}_\mu(y) = \mu$ and so $\tau = 1$. Hence, $\rho(y, u, \tau)$ given in (8) is given by $\rho(y, u, 1) = \phi(-y + y\ln y + H(u) - y\ln(H(u))) + G(H(u))$, where

$$G'(t) = -\varphi(t)\exp(-t) - \sum_{j=1}^{\infty} \varphi(j\ln j - j + t - j\ln t)\left(\frac{t-j}{t}\right)\exp(-t)t^j/j!.$$

In the particular case of the canonical link function, that is, when $\log\mu = u$, i.e. $H(u) = \exp(u)$, we have that $\rho(y, u, 1) = \phi(-y + y\ln y + H(u) - yu) + G(H(u))$.

2.4. The log–Gamma regression model

An important application among generalized linear models is the Gamma distribution with a log-link. This model is called log–Gamma regression and is introduced in Chapter 8 of [20]. For the log–Gamma model we have that $y_i|\mathbf{x}_i \sim \Gamma(\tau, \mu_i)$, where $\mu_i = \mathbb{E}(y_i|\mathbf{x}_i)$ and the link function is $\log(\mathbb{E}(y_i)) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$, where, for any $\tau > 0$ and $\mu > 0$, we denote by $\Gamma(\tau, \mu)$ the parametrization of the Gamma distribution given by the density

$$f(y, \tau, \mu) = \tau^\tau\, y^{\tau-1}\, \exp(-(\tau/\mu)y)\{\mu^\tau\, \Gamma(\tau)\}^{-1} I_{y\geq 0}.$$

It is well known (see for instance, [7]) that in this case, the responses can be transformed so that they are modelled through a linear regression model with asymmetric errors. Indeed, let $z_i = \log(y_i)$ be the transformed responses, then $z_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + u_i$, where u_i and $\mathbf{x}_i$ are independent. Besides, $u_i \sim \log(\Gamma(\tau, 1))$ with density

$$g(u, \tau) = \frac{\tau^\tau}{\Gamma(\tau)}\, \exp(\tau(u - \exp(u))). \tag{11}$$

This density is asymmetric and unimodal with maximum at $u_0 = 0$. We refer to [3] for a description on the robust estimators based on deviances for complete data sets and to [14] for a description on M-type estimators. For the sake of completeness, we will describe how to adapt the estimators based on deviances to the situation with missing responses since this is one of the models used in the simulation study considered in the technical report by Bianco et al. [2].

Denote by $d_i(\boldsymbol{\beta}, \tau)$ the deviance component of the i-th observation, i.e., $d_i(\boldsymbol{\beta}, \tau) = 2\tau\; d^*(y_i, \mathbf{x}_i, \boldsymbol{\beta})$, where $d^*(y, \mathbf{x}, \boldsymbol{\beta}) = -1 - (\log(y) - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + y\exp(-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. Consider the transformed responses $z_i = \log(y_i)$. Let us now assume that we are dealing with the situation in which some of the responses may be missing, with $\delta_i = 1$ if z_i is observed, and 0 otherwise. Moreover, (1) entails that δ_1 and $\underset{\sim}{z}_1$ are conditionally independent given $\mathbf{x}_1$, so δ_1 and u_1 are independent. Note that $d^*(y, \mathbf{x}, \mathbf{b}) = \underset{\sim}{d}(u, \mathbf{x}, \boldsymbol{\beta} - \mathbf{b})$, where $\underset{\sim}{d}(u, \mathbf{x}, \mathbf{b}) = -1 - u - \mathbf{x}^{\mathrm{T}}\mathbf{b} + \exp(u)\exp(\mathbf{x}^{\mathrm{T}}\mathbf{b})$. The maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ is, thus, obtained as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i d^*(y_i, \mathbf{x}_i, \mathbf{b}).$$

As in [3], a three-step procedure can be considered to compute the estimators when missing responses are present. First note that, since the tuning constant of the loss function depends on the unknown parameter τ, Maronna et al. [3] introduce an adaptive sequence of tuning constants $\widehat{c}_{\mathrm{M},n}$ to define a sequence of M-estimators, $\widehat{\boldsymbol{\beta}}_{\mathrm{M},n}$, for data sets with no missing observations. These estimators, which satisfy $\widehat{\boldsymbol{\beta}}_{\mathrm{M},n} = \operatorname{argmin}_{\mathbf{b}} \sum_{i=1}^{n} \phi\left(\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}/\widehat{c}_{\mathrm{M},n}\right)$, have as asymptotic covariance matrix $\left(B(\phi, \tau, c_0)/A^2(\phi, \tau, c_0)\right) \mathbb{E}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)^{-1}$, where $\widehat{c}_{\mathrm{M},n} \xrightarrow{p} c_0$. The constants $B(\phi, \tau, c_0)$ and $A^2(\phi, \tau, c_0)$ depend only on the derivative of the score function ϕ and the shape parameter τ, but not on the covariates. Hence, the estimators can be calibrated to attain a given efficiency. From now on, denote $C_e(\tau)$ as the value of the tuning constant c_0 such that the M-estimator has efficiency e with respect to the maximum likelihood estimator.

In our modification, we consider the following three-step algorithm to compute a generalized MM-estimator.

- *Step 1.* Consider the complete data subset, that is, the portion of data at hand, i.e., $\{(y_i, \mathbf{x}_i)\}_{i:\delta_i=1}$. We first compute an initial S-estimate $\widetilde{\boldsymbol{\beta}}_n$ and the corresponding scale estimate $\widehat{\sigma}_n$ taking $b = \frac{1}{2}\sup\phi$ with the complete data subset. To be more precise, for each value of $\mathbf{b}$ let $\sigma_n(\mathbf{b})$ be the M-scale estimate of $\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}$ given by

$$\sum_{i=1}^{n} \delta_i \phi\left(\frac{\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}}{\sigma_n(\mathbf{b})}\right) = b \sum_{i=1}^{n} \delta_i,$$

where ϕ is the Tukey bisquare function, $\phi(u) = \min\left\{u^2/2 \min\left(1 - u^2 + u^4/3\right), 1\right\}$.

The S-estimate of $\boldsymbol{\beta}$ for the considered model is defined by $\widetilde{\boldsymbol{\beta}}_n = \operatorname{argmin}_{\mathbf{b}} \sigma_n(\mathbf{b})$ and the corresponding scale estimate by $\widehat{\sigma}_n = \min_{\mathbf{b}} \sigma_n(\mathbf{b})$. The functional related to this S-estimator is defined by $\widetilde{\boldsymbol{\beta}}(F) = \operatorname{argmin}_{\mathbf{b}} \sigma(\tau, \mathbf{b})$, where $\sigma(\tau, \mathbf{b})$ is the solution of

$$\mathbb{E}_F \delta\phi\left(\frac{\sqrt{d^*(y, \mathbf{x}, \mathbf{b})}}{\sigma(\tau, \mathbf{b})}\right) = b\, \mathbb{E}p(\mathbf{x}). \tag{12}$$

Let u be a random variable with density (11) and write $\sigma^*(\tau)$ for the solution of

$$\mathbb{E}_G\left[\phi\left(\frac{\sqrt{h(u)}}{\sigma^*(\tau)}\right)\right] = b,$$

where $h(u) = 1 - u - \exp(u)$. Note that since u and δ are independent, we have that $\sigma^*(\tau) = \sigma(\tau, \boldsymbol{\beta})$. Similar arguments to those considered in Theorem 5 in [3] allow one to show that under mild conditions $\widetilde{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}$ and that $\widehat{\sigma}_n \xrightarrow{a.s.} \sigma^*(\tau)$. Moreover, as in [3], $\sigma^*(\tau)$ is a continuous and strictly decreasing function, thus an estimator of τ can be defined as $\widehat{\tau}_n = \sigma^{*-1}(\widehat{\sigma}_n)$, leading to a a strongly consistent estimator for τ.
- *Step 2.* In the second step, we compute $\widehat{\tau}_n = \sigma^{*-1}(\widehat{\sigma}_n)$ and

$$\widehat{c}_n = \max(\widehat{\sigma}_n, C_e(\widehat{\tau}_n)) = \max(\widehat{\sigma}_n, C_e(\sigma^{*-1}(\widehat{\sigma}_n))).$$

We then have that $\widehat{c}_n \xrightarrow{p} c_0 = \max\{\sigma^*(\tau), C_e(\tau)\}$.
- *Step 3.* Let $\widehat{\boldsymbol{\beta}}_n$ be the adaptive general M-estimator of $\boldsymbol{\beta}$ defined by

$$\widehat{\boldsymbol{\beta}}_n = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i \phi\left(\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}/\widehat{c}_n\right) w_1(\mathbf{x}_i). \tag{13}$$

Note that in this case, $\rho(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t) = \phi\left(\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}/c(t)\right)$, where $c(t) = \max\{\sigma^*(t), C_e(t)\}$, so that $\rho(y, v, t) = \phi\left(\sqrt{-1 - \log(y) + v + y\exp(-v)}/c(t)\right)$.

The following Lemma states the Fisher-consistency of the functionals related to the estimators $\widetilde{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}_n$. Its proof can be found in the Appendix.

Lemma 2.1. *If the score function ϕ is bounded and strictly increasing on the set where it is not constant and if (10) holds, we have that the functionals defined as $\widetilde{\boldsymbol{\beta}}(F) = \operatorname{argmin}_{\mathbf{b}} \sigma(\tau, \mathbf{b})$ and $\boldsymbol{\beta}(F) = \operatorname{argmin}_{\mathbf{b}} \mathbb{E}_F\left[\delta\phi\left(\sqrt{d^*(y, \mathbf{x}, \mathbf{b})}/c_0\right) w_1(\mathbf{x})\right]$ are Fisher-consistent.*

Propensity score estimators are defined in an analogous way.

3. Consistency results

As mentioned in [13], under Fisher-consistency and under assumptions **A1**–**A3** below, standard arguments allow one to show that the simplified estimators introduced in Section 2.1 are consistent (see Huber, [16], for instance). For the sake of completeness, we state here these results.

A1. The functions $w_1(\mathbf{x})$ is bounded.
A2. $\rho(y, u, v)$ is a continuous function. Moreover, the function $S(\mathbf{b}, t)$ satisfies the following equicontinuity condition: for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $t_1, t_2 \in \mathcal{K}$, a compact set in $\mathcal{T}$,

$$|t_1 - t_2| < \delta \Rightarrow \sup_{\mathbf{b}\in\mathbb{R}^k} |S(\mathbf{b}, t_1) - S(\mathbf{b}, t_2)| < \epsilon.$$

A3. The class of functions $\mathcal{F} = \{f_{\mathbf{b},t}(y, \mathbf{x}, \delta) = \delta\rho\,(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t)\,w_1(\mathbf{x}), t \in \mathcal{T}, \mathbf{b} \in \mathbb{R}^k\}$ satisfies either (a) or (b) where:
(a) The bracketing number $N_{[\,]}\left(\epsilon, \mathcal{F}, L^1(P)\right) < \infty$, for any $0 < \epsilon < 1$, where P is the distribution of $(y, \mathbf{x})$.
(b) The covering number $N\left(\epsilon, \mathcal{F}, L^1(P_n)\right)$ is such that $\log N\left(\epsilon, \mathcal{F}, L^1(P_n)\right) = o_P(n)$, where P_n is the empirical distribution.

Proposition 3.1. *Assume that* **A1**, **A2** *and* **A3** *hold, then* $\widehat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}$.

Remark 3.1. Assumptions **A1** and **A2** are standard requirements since they state that the weight function controls large values of the covariates and that the score function bounds large residuals, respectively. Assumption **A3** is fulfilled for the score functions described in Section 2.2 under mild conditions. For instance, if ϕ and ψ_c are functions with bounded variation, **A3** holds for most distribution families.

Using similar arguments under the following additional conditions, we obtain the consistency of the propensity estimators, which is stated in the next proposition.

A4. $\inf_{\mathbf{x}\in\mathcal{S}_{w_1}\cap\mathcal{S}_{\mathbf{x}}} p(\mathbf{x}) = A > 0$, where $\mathcal{S}_{w_1}$ and $\mathcal{S}_{\mathbf{x}}$ stand for the support of w_1 and $\mathbf{x}$, respectively.
A5. The estimator $\widehat{p}(\mathbf{x})$ of $p(\mathbf{x})$ satisfies either (a) or (b)
(a) $\sup_{\mathbf{x}\in\mathcal{S}_{w_1}\cap\mathcal{S}_{\mathbf{x}}} |\widehat{p}(\mathbf{x}) - p(\mathbf{x})| \xrightarrow{a.s.} 0$ or
(b) $p(\mathbf{x}) = p_{\boldsymbol{\lambda}}(\mathbf{x}) = G_{\mathrm{P}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda})$ for some continuous function $G_{\mathrm{P}} : \mathbb{R} \to (0, 1]$ with bounded variation, $\boldsymbol{\lambda} \in \boldsymbol{\Lambda} \subset \mathbb{R}^k$ and $\widehat{p}(x) = p_{\widehat{\boldsymbol{\lambda}}}(\mathbf{x})$, where $\widehat{\boldsymbol{\lambda}} \xrightarrow{a.s.} \boldsymbol{\lambda}$.
A6. The function $S_{\mathrm{P}}(\mathbf{b}, t, p)$ satisfies the following equicontinuity condition:
(a) under **A5**(a), for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $t_1, t_2 \in \mathcal{K}$, a compact set in $\mathcal{T}$,
$$|t_1 - t_2| < \delta \Rightarrow \sup_{\mathbf{b}\in\mathbb{R}^k} |S_{\mathrm{P}}(\mathbf{b}, t_1, p) - S_{\mathrm{P}}(\mathbf{b}, t_2, p)| < \epsilon.$$
(b) under **A5**(b), for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $t_1, t_2 \in \mathcal{K}$ and $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{K}_{\boldsymbol{\Lambda}}$, compact sets in $\mathcal{T}$ and $\mathbb{R}^k$, respectively,
$$|t_1 - t_2| < \delta \text{ and } \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| < \delta \Rightarrow \sup_{\mathbf{b}\in\mathbb{R}^k} |S_{\mathrm{P}}(\mathbf{b}, t_1, p_{\boldsymbol{\lambda}_1}) - S_{\mathrm{P}}(\mathbf{b}, t_2, p_{\boldsymbol{\lambda}_2})| < \epsilon.$$

Proposition 3.2. *Assume that* **A1** –**A6** *hold, then* $\widehat{\boldsymbol{\beta}}_{\mathrm{P}} \xrightarrow{a.s.} \boldsymbol{\beta}$.

Remark 3.2. Note that **A6** holds if $\Psi\,(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t)$ and $w_1(\mathbf{x})\|\mathbf{x}\|$ are bounded, which holds for the usual functions considered in robustness. Besides, if w_1 has compact support, as is the case for the Tukey weight function, **A4** holds for any continuous missingness probability such that $p(\mathbf{x}) > 0$. This includes, for instance, a logistic model for $p(\mathbf{x})$. On the other hand, if $\mathcal{S}_{\mathbf{x}} = \mathbb{R}^k$ and $w_1 \equiv 1$, i.e., if high leverage points are not downweighted, **A4** restricts the family of missing probabilities to be considered.

4. Influence functions and outlier detection

Usually, in robustness, there are two popular measures of the resistance to outliers of a given estimator: the breakdown point and the influence function of the related functional.

Loosely speaking, the breakdown point of an estimator is the smallest fraction of outliers that can take the estimate beyond any bound. Hampel [12] defined the asymptotic breakdown point, while Donoho and Huber [9] introduced the finite sample version. Consider a sample of n complete observations $\mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, where $\mathbf{z}_i \in \mathbb{R}^\ell$, and let $T_n(\mathcal{Z})$ be an estimate of a parameter $\boldsymbol{\eta} \in \mathbb{R}^d$ defined on all possible datasets. Let $\mathcal{Z}_s$ be the set of all the samples $\mathcal{Z}^* = \{\mathbf{z}_1^*, \dots, \mathbf{z}_n^*\}$ such that $\#\{i : \mathbf{z}_i^* = \mathbf{z}_i\} = n - s$. The finite sample breakdown point is defined as $\epsilon^*(T, \mathcal{Z}) = \min\{s/n : \sup_{\mathcal{Z}^*\in\mathcal{Z}_s} \|T(\mathcal{Z}^*)\| = \infty\}$.

Sued and Yohai [26] extended the notion of finite sample breakdown point to the case where missing responses can occur. Let $\mathcal{W} = \{(y_1, \mathbf{x}_1^{\mathrm{T}}, \delta_1), \dots, (y_n, \mathbf{x}_n^{\mathrm{T}}, \delta_n)\}$ be the set of all the observations. If $\Delta = \{i : 1 \le i \le n, \delta_i = 1\}$, let $m = \#\Delta$. Denote by $\mathcal{W}_{ts}$ the set of all samples obtained from $\mathcal{W}$ by replacing at most t points by outliers, with at most s of these replacements corresponding to the non missing responses. Then, $\mathcal{W}^* = \{(y_1^*, \mathbf{x}_1^{*\mathrm{T}}, \delta_1), \dots, (y_n^*, \mathbf{x}_n^{*\mathrm{T}}, \delta_n)\}$ belongs to $\mathcal{W}_{ts}$ if

$$\#\{i \in \Delta : (y_i^*, \mathbf{x}_i^{*\mathrm{T}}) \neq (y_i, \mathbf{x}_i)\} + \#\{i \in \Delta^c : \mathbf{x}_i^* \neq \mathbf{x}_i\} \le t$$

and

$$\#\{i \in \Delta : (y_i^*, \mathbf{x}_i^{*\mathrm{T}}) \neq (y_i, \mathbf{x}_i)\} \leq s.$$

Given the estimator T_n of $\boldsymbol{\eta}$, denote $M_{ts} = \sup_{\mathcal{W}^* \in \mathcal{W}_{ts}} \{T_n(\mathcal{W}^*)\}$ and $\kappa(t, s) = \max(t/n, s/m)$. Sued and Yohai [26] defined the finite breakdown point of T_n at $\mathcal{W}$ as $\epsilon^* = \min\{\kappa(t, s) : M_{ts} = \infty\}$. In this sense, ϵ^* is the minimum fraction of outliers in the complete sample or in the set of non missing responses that is required to take the estimate beyond any limit.

The simplified estimators, introduced in Section 2.1, depend only on the complete observations in the sample, so the procedure preserves the finite breakdown point of the regression estimator for a sample of size m, as defined by Donoho and Huber [9].

On the other hand, the influence function is a measure of robustness with respect to single outliers that allows us to study the local robustness and the asymptotic efficiency of the estimators, providing a rationale for choosing appropriate weight functions and tuning parameters. It can be thought of as the first derivative of the functional version of the estimator which, under mild conditions, enables the derivation of the asymptotic covariance matrix of the corresponding estimator. The influence function of a functional $T(F)$ is defined as $\mathrm{IF}(\mathbf{z}_0, T, F) = \lim_{\epsilon \to 0}(T(F_{\mathbf{z}_0,\epsilon}) - T(F))/\epsilon$, where $F_{\mathbf{z}_0,\epsilon} = (1-\epsilon)F + \epsilon\Delta_{\mathbf{z}_0}$ and $\Delta_{\mathbf{z}_0}$ denotes the probability measure which puts mass 1 at the point $\mathbf{z}_0 = (y_0, \mathbf{x}_0^{\mathrm{T}}, \delta_0)$ and represents the contaminated model. Under mild conditions, see [10], we have the expansion

$$\sqrt{n}\{T(F_n) - T(F)\} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathrm{IF}(\mathbf{z}_i, T, F) + o_p(1),$$

where F_n denotes the empirical distribution of the observations $\mathbf{z}_i$, $1 \leq i \leq n$. Therefore, the asymptotic variance of the estimates can be evaluated as

$$\mathrm{ASVAR}(T, F) = \mathbb{E}_F \left\{ \mathrm{IF}(\mathbf{z}_1, T, F)\, \mathrm{IF}(\mathbf{z}_1, T, F)^{\mathrm{T}} \right\}. \tag{14}$$

Besides being of theoretical interest and helpful to calibrate the efficiency of the robust estimates, measuring the influence of an observation on the classical estimates can be used as a diagnostic tool to detect influential observations.

Let F_1 stand for any distribution on $\mathbb{R}^{k+1} \times \{0, 1\}$ and denote by $\boldsymbol{\beta}(F_1)$ and $\tau(F_1)$ the functionals related to the simplified estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\tau}$, respectively. That is, $\boldsymbol{\beta}(F_1)$ is the solution of $S^{(1)}(\mathbf{b}, \tau(F_1), F_1) = \mathbf{0}_k$ with $S^{(1)}(\mathbf{b}, t, F_1) = \mathbb{E}_{F_1}(\delta\Psi(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t)\, w_1(\mathbf{x})\mathbf{x})$. Assume that $\boldsymbol{\beta}(F_1)$ is a Fisher-consistent functional at F, i.e., $\boldsymbol{\beta}(F) = \boldsymbol{\beta}$.

Theorem 4.1 gives the influence function for the simplified functional $\boldsymbol{\beta}(F_1)$ at $F_1 = F$ under the assumptions given below.

H1. $\Psi(y, u, v)$ is a continuous differentiable function of (u, v) and $\chi(y, u, v) = \partial\Psi(y, u, v)/\partial u$ is a continuous function.
H2. The matrix $\mathbf{A} = \mathbb{E}_F(\chi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tau)\, w_1(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}})$ is non–singular.
H3. $\mathbb{E}_F(\Psi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, t)\,|\mathbf{x}) = \mathbf{0}_k$ for any fixed $t \in \mathcal{T}$.

Remark 4.1. Assumption **H1** is a standard requirement to ensure the existence of an influence function without additional smoothness requirements on the underlying distribution. Assumption **H2** is a standard condition in the robustness literature to guarantee that the regression estimators will be root-n consistent. Note that **H3** holds for the usual functions considered in robustness, it is the conditional Fisher-consistency defined by Künsch et al. [18].

Theorem 4.1. *Let* **A** *be the symmetric matrix defined in* **H2** *and assume that* $IF(\mathbf{z}_0, \tau, F)$ *exists. Then, under* **H1** *–***H3***,* $IF(\mathbf{z}_0, \boldsymbol{\beta}, F)$ *exists and when* $\tau(F) = \tau$, *then* $IF(\mathbf{z}_0, \boldsymbol{\beta}, F) = -\Psi\left(y_0, \mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}, \tau\right) w_1(\mathbf{x}_0)\delta_0\mathbf{A}^{-1}\mathbf{x}_0$.

It is worth noticing that the influence function depends on the indicator of the missing response δ_0, so it will be 0 if no responses arise. For that reason, a more reliable function to measure the sensitivity to outliers of a given functional $T(F_1)$ under a missing scheme needs to be considered.

We define the *expected influence function* at an observed data $\mathbf{z}_0^* = (y_0, \mathbf{x}_0^{\mathrm{T}})^{\mathrm{T}}$, denoted $EIF(\mathbf{z}_0^*, T, F)$, as $EIF(\mathbf{z}_0^*, T, F) = \mathbb{E}(\mathrm{IF}(\mathbf{z}_0, T, F)|(y_0, \mathbf{x}_0^{\mathrm{T}}))$. There are some differences between the influence function and the expected influence function to be pointed out. The most important one is that the former allows one, under regularity conditions, to provide a Bahadur expansion of the statistic under study. In particular, the asymptotic variance of the estimator can be heuristically obtained through (14). This is not true if one replaces IF by *EIF* in (14) since the missing indicator δ_i cannot be replaced by its conditional expected value when providing the Bahadur expansion of the estimator. However, *EIF* provides a measure of the conditional expected influence of a data point, regardless of the presence or not of a response. In this sense, it also gives an idea on how the missingness probability will affect the functional and, therefore, the related estimator.

For the functional under study, we have that:

$$EIF(\mathbf{z}_0^*, \boldsymbol{\beta}, F) = -\Psi\left(y_0, \mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}, \tau\right) w_1(\mathbf{x}_0)p(\mathbf{x}_0)\mathbf{A}^{-1}\mathbf{x}_0.$$

To derive the influence function of the propensity score estimators, let us assume a parametric model for the probability of being missing, $p(\mathbf{x}) = G(\mathbf{x}, \boldsymbol{\lambda})$. As above, denote by $\boldsymbol{\beta}_{\mathrm{P}}(F_1)$, $\tau_{\mathrm{P}}(F_1)$ and $\boldsymbol{\lambda}(F_1)$ the functionals related to the estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$, $\widehat{\tau}_{\mathrm{P}}$ and $\widehat{\boldsymbol{\lambda}}$, respectively, where $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$ is defined in (7) with $\widehat{p}(\mathbf{x}) = G(\mathbf{x}, \widehat{\boldsymbol{\lambda}})$ the propensity score estimate and $\widehat{\boldsymbol{\lambda}}$ is a consistent estimator of $\boldsymbol{\lambda}$. Assume that $\boldsymbol{\beta}_{\mathrm{P}}(F_1)$ is a Fisher-consistent functional at F, i.e., $\boldsymbol{\beta}_{\mathrm{P}}(F) = \boldsymbol{\beta}$. Note that $\boldsymbol{\beta}_{\mathrm{P}}(F_1)$ is the solution of $S_{\rm P}^{(1)}(\mathbf{b},\tau_{\rm P}(F_1),\boldsymbol{\lambda}(F_1),F_1)=\mathbf{0}_k$ with $S_{\rm P}^{(1)}(\mathbf{b},\tau,\boldsymbol{\lambda}(F_1),F_1)=\mathbb{E}_{F_1}\left(\delta\Psi\,(y,\mathbf{x}^{\rm T}\mathbf{b},\tau)\,w_1^*(\mathbf{x},F_1)\mathbf{x}\right)$ and $w_1^*(\mathbf{x},F_1)=w(\mathbf{x})/G(\mathbf{x},\boldsymbol{\lambda}(F_1))$.

Theorem 4.2 gives the value of the influence function of the functional $\boldsymbol{\beta}_{\rm P}(F)$ under some additional assumptions.

H4. $p(\mathbf{x})=G(\mathbf{x},\boldsymbol{\lambda})$ for some continuous function $G:\mathbb{R}^{2k}\to(0,1)$ differentiable with respect to $\boldsymbol{\lambda}$.
H5. The matrix $\mathbf{A}_{\rm P}=\mathbb{E}_F\,(\chi\,(y,\mathbf{x}^{\rm T}\boldsymbol{\beta},\tau)\,w_1(\mathbf{x})\mathbf{x}\mathbf{x}^{\rm T})$ is non–singular.
H6. $\mathbb{E}\left(|\Psi\,(y,\mathbf{x}^{\rm T}\mathbf{b},\tau)\,|w_1^*(\mathbf{x},F_1)\|\mathbf{x}\|\|\partial G(\mathbf{x},u)/\partial u|_{u=\boldsymbol{\lambda}}\|\right)<\infty$.

Theorem 4.2. *Let $\mathbf{A}_{\rm P}$ be the symmetric matrix defined in* **H5** *and assume that $IF(\mathbf{z}_0,\tau_{\rm P},F)$ and $IF(\mathbf{z}_0,\lambda,F)$ exist. Then, under* **H1** *and* **H3** *–***H6***, we have that $IF\left(\mathbf{z}_0,\boldsymbol{\beta}_{\rm P},F\right)$ exists and $IF(\mathbf{z}_0,\boldsymbol{\beta}_{\rm P},F)=-\left(\delta_0/p(\mathbf{x}_0)\right)\Psi\left(y_0,\mathbf{x}_0^{\rm T}\boldsymbol{\beta},\tau\right)w_1(\mathbf{x}_0)\mathbf{A}_{\rm P}^{-1}\mathbf{x}_0$.*

It is worth noting that the expected influence function of the propensity estimator $EIF(\mathbf{z}_0^*,\boldsymbol{\beta}_{\rm P},F)=-\Psi\left(y_0,\mathbf{x}_0^{\rm T}\boldsymbol{\beta},\tau\right)w_1(\mathbf{x}_0)\mathbf{A}_{\rm P}^{-1}\mathbf{x}_0$, does not depend on the missing probability due to the fact that $\mathbf{A}_{\rm P}$ is independent of the missing pattern. Thus, they coincide with those of the simplified estimator for the case $p\equiv 1$.

4.1. Expected influence functions for the Poisson model

In this Section, we compute the expected influence functions for the Poisson regression model described in Section 2.3 for the case of the canonical link function, that is

$$y_i|\mathbf{x}\sim\mathcal{P}(\mu(\mathbf{x}))\quad\text{with }\log(\mu(\mathbf{x}))=\beta_0+\beta_1x_1,\quad\mathbf{x}=(1,x_1)^{\rm T},\tag{15}$$

where $x_1\sim N(0,1)$, $\beta_0=0$ and for two values of β_1, $\beta_1=0,0.4$.

For the weighted estimators, we used the Tukey's bisquare weight function with tuning constant $\chi^2_{1,0.975}$. The weights were computed over the robust Mahalanobis distances based on the median and the median of the absolute deviation from the median (MAD). We denote by $\widehat{\boldsymbol{\beta}}_{\rm ML}$, $\widehat{\boldsymbol{\beta}}_{\rm M}$, $\widehat{\boldsymbol{\beta}}_{\rm GM}$, the maximum likelihood estimators, the estimators obtained with $w_1\equiv 1$ and their weighted version with Tukey's weights, respectively. The proposed robust estimators were computed using the loss function introduced by Bianco and Yohai [4] and defined in (9) with tuning constant $c=4$. The choice of $c=4$ is explained in Section 5.

Figs. 1 and 2 give the surface plots of the norm of the expected influence functions, $\|EIF\|$, for the simplified maximum likelihood estimators and robust estimators defined through (4), when $\beta_1=0$ and 0.4, respectively. As mentioned above, the EIF of the propensity estimators equals that of the simplified ones when $p\equiv 1$, so it is avoided. We considered three models for the missing probability, $p\equiv 1$, that is, no missing responses arise, the logistic model $p(\mathbf{x})=1/(1+\exp(-2x-2))$ and $p(\mathbf{x})=0.7+0.2(\cos(2x+0.4))^2$, which corresponds to a missing probability bounded from below. Note that when p equals a constant, which corresponds to a missing completely at random model, the $\|EIF\|$ is the same, except for a scale factor, to that of $p\equiv 1$.

When considering the functional related to maximum likelihood estimator, the plots of the $\|EIF\|$ reveal that it is unbounded due to both leverage points and values of the responses with large deviances. In particular, for a logistic missing probability, the $\|EIF\|$ of the functionals related to all the simplified estimators show an asymmetric behaviour. This can be explained by the fact that large negative values of $\mathbf{x}_0$ lead to values of the missing probability close to 0. On the other hand, for a cosine missing probability model, the fluctuation effect introduced by the cosine is observed in the plot. As one can guess, the norms of the expected influence functionals corresponding to $\widehat{\boldsymbol{\beta}}_{\rm M}$ and to its weighted version $\widehat{\boldsymbol{\beta}}_{\rm GM}$ are comparable to that of the classical one at the center of the distribution of the covariates under all missing schemes. Figs. 1 and 2 make clear that the behaviour of the influence function depends on the value of the underlying regression parameters. This is especially evident for the robust unweighted estimator, i.e., those corresponding to $w_1\equiv 1$, in which large values of the EIF are attained when the score function is unable to bound the effect of the covariates. On the other hand, the plots show that using a weight function to downweight carriers with large Mahalanobis distances, the norm of the expected influence at points further away remains completely downweighted.

4.2. Asymptotic variances

As mentioned above, under regularity conditions, the influence function IF$(\mathbf{z},F)$ at point $\mathbf{z}$ of functional estimator $T(F)$ can be seen as a derivative in a stronger sense and, thus, the asymptotic variance of the estimators can be derived from the influence function using (14), (see also [19], pp. 71–75).

So, in terms of the mentioned approximation, the expressions obtained for influence functions of the simplified and propensity estimators in Theorems 4.1 and 4.2, respectively, enable us to derive heuristically the asymptotic covariance matrix of the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{\rm P}$. Hence, if we denote

$$\mathbf{B}=\mathbb{E}_F\left(\Psi^2\left(y,\mathbf{x}^{\rm T}\boldsymbol{\beta},\tau\right)w_1^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\rm T}\right)\quad\text{and}\quad\mathbf{B}_{\rm P}=\mathbb{E}_F\left(\Psi^2\left(y,\mathbf{x}^{\rm T}\boldsymbol{\beta},\tau\right)w_1^2(\mathbf{x})p(\mathbf{x})^{-1}\mathbf{x}\mathbf{x}^{\rm T}\right),\tag{16}$$

under regularity conditions and according to (14), the asymptotic covariance matrix of the simplified estimator $\widehat{\boldsymbol{\beta}}$ is $\boldsymbol{\Sigma}=\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, while that of the propensity estimators $\widehat{\boldsymbol{\beta}}_{\rm P}$ is $\boldsymbol{\Sigma}_{\rm P}=\mathbf{A}_{\rm P}^{-1}\mathbf{B}_{\rm P}\mathbf{A}_{\rm P}^{-1}$.

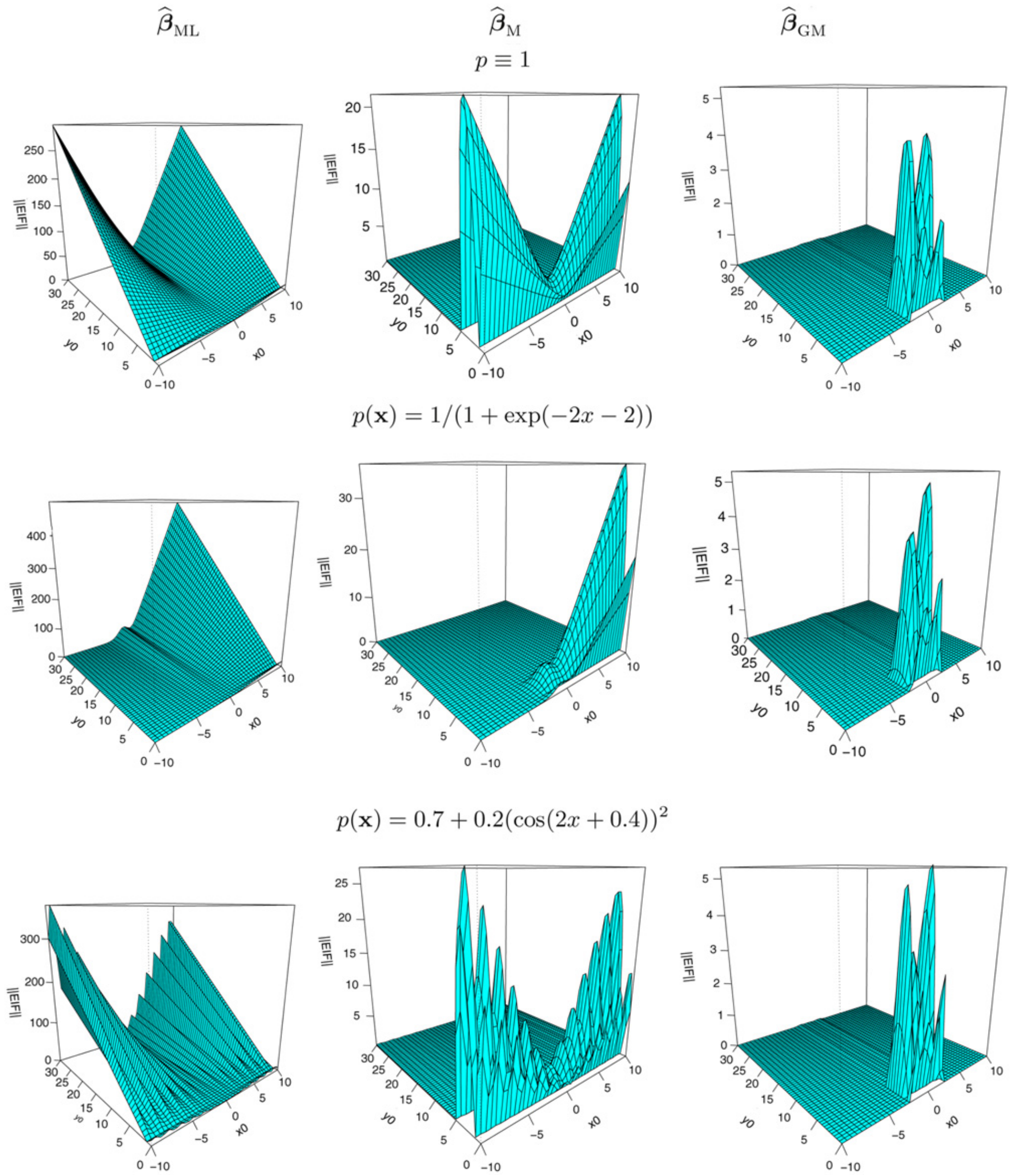

Fig. 1. Norm of the expected influence functions, $\|EIF\|$, under the Poisson model for the functionals based on the simplified estimators for $\beta_0 = 0$ and $\beta_1 = 0$.

Remark 4.2. When considering the log–Gamma regression model, the estimator $\widehat{\boldsymbol{\beta}}_n$ defined in (13) has an asymptotic covariance matrix which can be written as $\boldsymbol{\Sigma} = \gamma(\phi, \tau, c_0)\mathbf{C}$, for a properly constant $\gamma(\phi, \tau, c_0)$ that depends on the score function and the shape parameter. The matrix $\mathbf{C}$ is given by $\mathbf{C} = \mathbf{D}^{-1}\mathbb{E}\left(p(\mathbf{x})w_1^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)\mathbf{D}^{-1}$ with $\mathbf{D} = \mathbb{E}\left(p(\mathbf{x})w_1(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$. So, when a $0 - 1$ weight function is considered, the asymptotic matrix $\mathbf{C}$ reduces to $\mathbf{C} = \mathbf{D}^{-1} = [\mathbb{E}\left(p(\mathbf{x})w_1(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)]^{-1}$.

On the other hand, the asymptotic covariance matrix of the propensity score estimator for this model equals $\boldsymbol{\Sigma}_{\mathrm{P}} = \gamma(\phi, \tau, c_0)\mathbf{C}_{\mathrm{P}}$, where $\mathbf{C}_{\mathrm{P}} = \mathbf{D}_{\mathrm{P}}^{-1}\mathbb{E}\left(\left(w_1^2(\mathbf{x})/p(\mathbf{x})\right)\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)\mathbf{D}_{\mathrm{P}}^{-1}$ and $\mathbf{D}_{\mathrm{P}} = \mathbb{E}\left(w_1(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$.

In the particular case of univariate covariates x and $0 - 1$ weight function w_1, we have that $\mathbf{C} = \left(\mathbb{E}\left(p(x)w_1(x)x^2\right)\right)^{-1}$ and $\mathbf{C}_{\mathrm{P}} = \mathbb{E}\left(w_1(x)/p(x)x^2\right)\left(\mathbb{E}\left(w_1(x)x^2\right)\right)^{-2}$. The Cauchy–Schwartz inequality entails that $\left(\mathbb{E}\left(w_1(x)x^2\right)\right)^2 \leq \mathbb{E}\left(w_1(x)/p(x)x^2\right)$ $\mathbb{E}\left(w_1(x)p(x)x^2\right)$ hence, the simplified estimator is more efficient than its propensity version for the log–Gamma regression model.

4.3. Outlier detection

As mentioned above, in addition to its theoretical interest and its helpfulness in calibration purposes, measuring the influence of an observation on the classical estimates can be used as a diagnostic tool to detect influential observations. It is

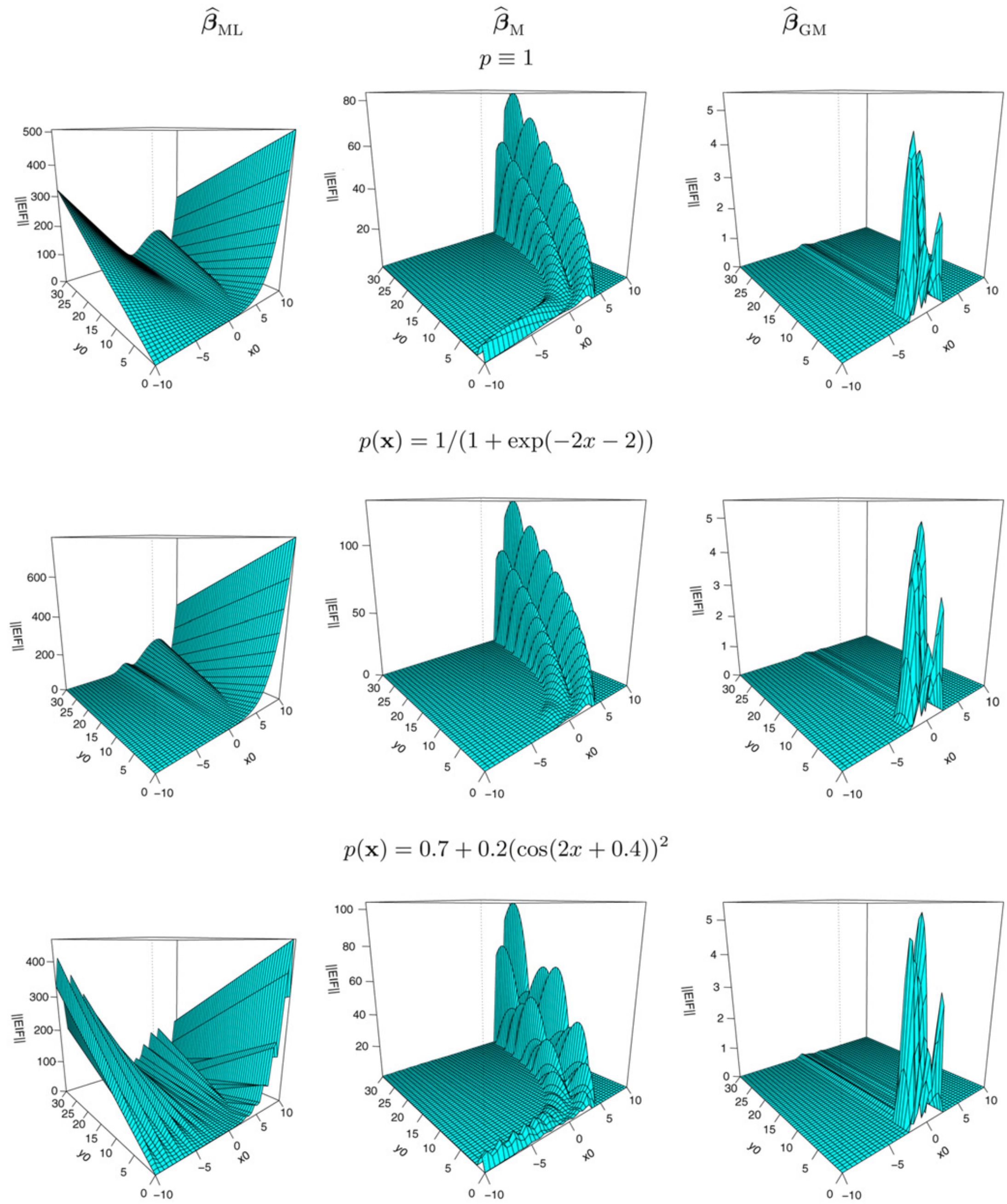

Fig. 2. Norm of the expected influence functions, $\|EIF\|$, under the Poisson model for the functionals based on the simplified estimators for $\beta_0 = 0$ and $\beta_1 = 0.4$.

well known that an outlier may not be an influential observation for the estimation of the parameter of interest, however an influential observation is usually an outlier. An influential observation can be described as an observation with high influence on something, usually an estimate of the parameters of interest. A well known procedure to detect outlying points for multivariate normal samples is the robustified Mahalanobis distance considered in [24], see also [23] for a review on robust statistical methods and outlier detection. García Ben and Yohai [11] extended the popular normal Q–Q plots to generalized linear regression models, where the distribution of deviance residuals may be far from normality, by means of an estimate of the distribution of the deviance residuals. Even if the aim of this paper is to adapt robust estimators for a GLM to the situation in which missing responses arise rather than to focus on outlier identification, it is important to know which points are influential for the regression parameter. The influence function serves this purpose. As described in other multivariate and regression settings, the influence function of the functional related to a robust estimator is usually bounded and so, when estimating the unknown parameters appearing in its expression by a robust procedure, it hardly changes when contaminated data points are included in the sample. On the other hand, if we consider the influence function of the functional related to the classical estimators, and we plug-in classical estimators of the parameters involved on it, a masking effect may appear and so outlying observations are not detected (see, for instance, [22]. An alternative approach is to consider the influence function for the classical estimators but plugging-in robust parameter estimates rather than classical ones, in order to avoid masking and to detect influential observations. So, when considering the GLM with missing responses the influence diagnostic measure (DM) related to the regression parameter can be computed, for the observed responses, that is, for a response y such that $\delta = 1$, as $\mathrm{DM}(y, \mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\tau})$, where $(\widehat{\boldsymbol{\beta}}, \widehat{\tau})$ are the robust simplified estimators. Therefore, in order to get an equivariant diagnostic measure, the $\mathrm{DM}(y, \mathbf{x}, \mathbf{b}, t)$ becomes

$$\mathrm{DM}(y, \mathbf{x}, \boldsymbol{\beta}, \tau) = \mathrm{IF}(\mathbf{z}, \boldsymbol{\beta}_{\mathrm{ML}}, F)^{\mathrm{T}}\mathbf{B}^{-1}\mathrm{IF}(\mathbf{z}, \boldsymbol{\beta}_{\mathrm{ML}}, F) = \delta\Psi^2_{\mathrm{ML}}\left(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tau\right)\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{x},$$

where $\boldsymbol{\beta}_{\mathrm{ML}}$ is the functional related to the classical estimators, Ψ_{ML} stands for their score function and $\mathbf{B} = \mathbf{B}(\boldsymbol{\beta}, \tau) = \mathbb{E}_F\left(\Psi^2_{\mathrm{ML}}(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tau)\, w_1^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$. Note that, in most cases, the term $\Psi^2_{\mathrm{ML}}(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t)$ equals the square of the Pearson residuals numerator, while measures based on squared deviance residuals are considered in [11] to define Q–Q plots in generalized linear models. The diagnostic measure defined takes also into account the effect of high leverage points on observations with intermediate residuals.

In Section 6, we illustrate the use of the diagnostic measure DM for a real data set, while in Sections 5.2 and 5.3 the performance of DM is tested numerically.

5. Monte Carlo study

5.1. Simulation study for the Poisson model

A simulation study for a Poisson model is conducted to compare the performance of the classical simplified and propensity estimators, $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{P,ML}}$, and two robust alternatives, a simplified or propensity M-estimator denoted $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{P,M}}$ respectively and its weighted counterpart denoted $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{P,GM}}$ as introduced in Section 2.1. The M-estimator bounds large values of the deviance, while the weighted estimator downweights also observations with large values of the covariates.

We consider the Poisson regression model with the canonical link function, that is $(y_i, \mathbf{x}_i^{\mathrm{T}})$, $1 \le i \le n$, satisfy model (15) with $\mathbf{x}_i = (1, x_{1i})^{\mathrm{T}}$, where $x_{1i} \sim N(0, 1)$, $\beta_0 = 0$ and $\beta_1 = 0.4$. The sample size is $n = 100$ and the number of Monte Carlo replications is $NR = 500$. We follow a scheme similar to that considered in [1].

In order to compare the behaviour of the estimators, we consider samples without outliers, denoted C_0, and contaminated samples with 10% of outliers, labelled as C_1 in the plots. For the contaminated data, the outlying points, $(y_0, x_{1,0})$, are all equal to $(y_0, x_{1,0}) = (20, 2.5)$, which gives an expected value of marginal expectation of the response variable y equal to 2.718.

We choose the three missingness models described in Section 4.1 and also $p \equiv 0.8$. In this case, the logistic model leads to a probability of missing equal to 0.999 at x_0, while the cosine one to a missing probability equal to 0.781. The propensity estimators, $\widehat{\boldsymbol{\beta}}_{\mathrm{P,ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{P,M}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{P,GM}}$, were calculated only for the logistic and cosine missing probabilities.

As in Section 4.1, the weighted estimators used the Tukey's bisquare weight function with tuning constant $\chi^2_{1,0.975}$ and the weights were computed over the robust Mahalanobis distances based on the median and the MAD. Besides, the robust estimators bound the deviances using the loss function (9) with tuning constant $c = 4$, that yields, when $p \equiv 1$ and $\beta_1 = 0.4$, an asymptotic efficiency equal to 0.808 and 0.765 for $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$, respectively.

The boxplots of the resulting estimates are shown in Figs. 3 and 4. Table 1 gives summary measures of the estimators under the different missing schemes when no contamination is introduced. The summary measures computed are the mean, the standard deviation (SD) and the mean square error (MSE) over replications. As it can be seen, for all considered missing patterns, when there are no outliers, the robust estimates have a behaviour comparable with that of the classical estimator with a larger loss of efficiency for the weighted estimators of the slope parameter β_1. Besides, the simplified estimators are more efficient than their propensity relatives. However, in the presence of outlying points the robust estimates outperform the classical estimates in all cases, and this is more evident in the case of the slope estimators. In general, there are small differences between propensity and simplified robust estimators.

5.2. Outlier detection for the Poisson model

To assess the performance of the detection measure DM in the Poisson regression model, we follow a similar simulation scheme to that described in Section 5.1, but now we consider three types of contaminated samples. Each contaminated sample contains 10% of outliers of the form $(y_0, x_{1,0})$. We choose $x_{1,0} = 2.5$, while y_0 equals one of the following values 10, 15 and 20. The simplified robust estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$ was computed as in Section 5.1. Under a Poisson regression model with natural link, the matrix $\mathbf{B}(\boldsymbol{\beta}) = \mathbb{E}_F\left(\Psi^2_{\mathrm{ML}}(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\, w^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right) = \mathbb{E}_F\left(\exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})w^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$. So, when computing the diagnostic measure and the cut–off points, we estimate $\mathbf{B}(\boldsymbol{\beta})$ using a hard rejection weight function, that is, $w(\mathbf{x}_i) = 1$ when $(x_{1i} - \widehat{\mu})^2/\widehat{\sigma}^2 \le \chi^2_{1,0.975}$ and $w_1(\mathbf{x}_i) = 0$ otherwise, where $\widehat{\mu} = \mathrm{MEDIAN}(x_{1i})$ and $\widehat{\sigma} = \mathrm{MAD}(x_{1i})$. We numerically compute the cut-off values $c_{1-\alpha}$, so as to ensure that in non-contaminated samples the percentage of correct observations mislabelled as outliers is approximately equal to $100\alpha\%$, with $\alpha = 0.05$ and 0.01. The robust estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$ are computed and then, each observation is classified as an outlier if $\mathrm{DM}(y, \mathbf{x}, \widehat{\boldsymbol{\beta}}_{\mathrm{GM}}) > c_{1-\alpha}$. It is worth noting that when missing responses arise, we computed $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$ defined through (4) using the portion of the data at hand. However, when classifying an observation through $\mathrm{DM}(y, \mathbf{x}, \widehat{\boldsymbol{\beta}}_{\mathrm{GM}}) > c_{1-\alpha}$, we force δ_i to be equal to 1, when an outlier appears, to be have an idea of the misclassification

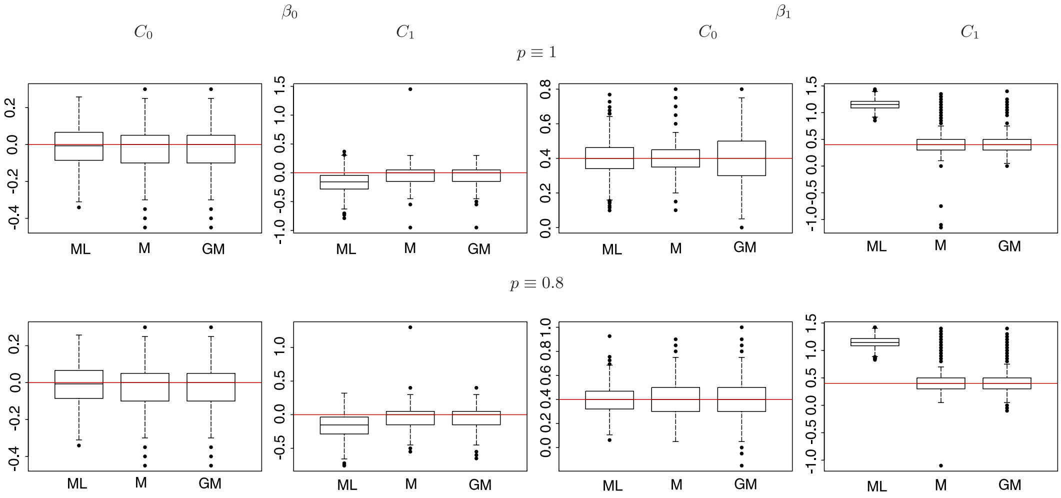

Fig. 3. Boxplots for the simplified estimators of β_0 and β_1 under the Poisson model when $p \equiv 1$ and $p \equiv 0.8$.

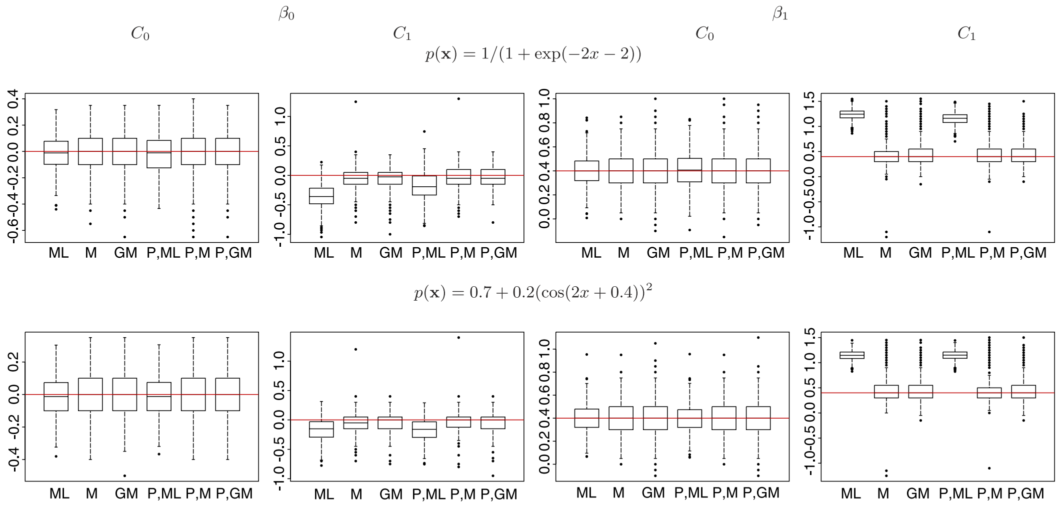

Fig. 4. Boxplots for the simplified and propensity estimators of β_0 and β_1 under the Poisson model when $p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$ and $p(\mathbf{x}) = 0.7 + 0.2(\cos(2x + 0.4))^2$.

Table 1
Summary measures for the regression estimators under a Poisson regression model with no contamination and $\beta_0 = 0$ and $\beta_1 = 0.4$.

	β_0						β_1					
	$\widehat{\beta}_{ML}$	$\widehat{\beta}_{M}$	$\widehat{\beta}_{GM}$	$\widehat{\beta}_{P,ML}$	$\widehat{\beta}_{P,M}$	$\widehat{\beta}_{P,GM}$	$\widehat{\beta}_{ML}$	$\widehat{\beta}_{M}$	$\widehat{\beta}_{GM}$	$\widehat{\beta}_{P,ML}$	$\widehat{\beta}_{P,M}$	$\widehat{\beta}_{P,GM}$
	$p \equiv 1$											
Mean	−0.0148	−0.0179	−0.0165				0.3980	0.3991	0.3987			
SD	0.1109	0.1203	0.1237				0.1025	0.1154	0.1451			
MSE	0.0125	0.0148	0.0150				0.0105	0.0133	0.0211			
	$p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$											
Mean	−0.0151	−0.0145	−0.0151	−0.0211	−0.0225	−0.0167	0.4018	0.4026	0.4014	0.4068	0.4039	0.4050
SD	0.1300	0.1438	0.1448	0.1447	0.1600	0.1505	0.1267	0.1405	0.1709	0.1501	0.1684	0.1797
MSE	0.0167	0.0205	0.0207	0.0205	0.0251	0.0224	0.0161	0.0197	0.0292	0.0225	0.0283	0.0322
	$p(\mathbf{x}) = 0.7 + 0.2(\cos(2x + 0.4))^2$											
Mean	−0.0146	−0.0144	−0.0142	−0.0143	−0.0149	−0.0136	0.3984	0.3975	0.3980	0.3981	0.3986	0.3972
SD	0.1167	0.1287	0.1354	0.1173	0.1275	0.1352	0.1188	0.1320	0.1627	0.1192	0.1327	0.1655
MSE	0.0134	0.0164	0.0181	0.0136	0.0160	0.0181	0.0141	0.0174	0.0265	0.0142	0.0176	0.0274

error obtained for the outliers. Table 2 reports the mean over replications of the proportion of outliers detected for each contamination and missing scheme considered, while the corresponding medians are all equal to 1. The obtained means show that on average DM successfully identifies at least 90% of the introduced outliers under all different missingness probabilities. The reported results reveal the proposed measure DM as a promising tool for the practitioner in the diagnosis of outlying observations when dealing with missing responses in a Poisson regression model.

Table 2
Mean over replications of the proportion of outliers detected.

y_0	$\alpha = 0.05$	$\alpha = 0.01$
$p \equiv 1$		
10	0.968	0.942
15	0.999	0.999
20	1	1
$p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$		
10	0.956	0.915
15	0.999	0.999
20	1	1
$p(\mathbf{x}) = 0.7+0.2(\cos(2x+0.4))^2$		
10	0.964	0.933
15	0.998	0.997
20	1	1

5.3. *Outlier detection for the Gamma model*

The gamma regression model considered was

$$y_i|\mathbf{x} \sim \Gamma(\tau, \mu(\mathbf{x})) \quad \text{with } \mu(\mathbf{x}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \; i = 1, \ldots, n, \tag{17}$$

with $\tau = 3$ and $\mathbf{x} = (\mathbf{z}^{\mathrm{T}}, 1)^{\mathrm{T}}$ where $\mathbf{z} \sim N(0, \mathbf{I}_{k-1})$. The sample size was $n = 100$ and the number of Monte Carlo replications was $NR = 1000$.

To study the behaviour of the detection measure, we consider samples contaminated with 5% of outliers all equal, say $(y_0, \mathbf{x}_0)$ with $\mathbf{x}_0 = (\mathbf{z}_0, 1)$. Since the magnitude of the effect of these outliers depends on $\mathbf{z}_0$ only throughout $\|\mathbf{z}_0\|^2$, without loss of generality they were chosen as $\mathbf{z}_{0,1} = x_0$, $\mathbf{z}_{0,j} = 0$ for $j > 1$ and $y_0 = \exp(\boldsymbol{\beta}\mathbf{x}_0 + m_0\, x_0)$. The value m_0 represents the slope of the outliers residuals. We chose three values of x_0 corresponding to low leverage outliers with $x_0 = 1$, moderate leverage outliers with $x_0 = 3$ and high leverage outliers with $x_0 = 10$. As values for m_0 we considered $m_0 = 0.5$ and 2.5.

The robust estimators were computed as described in Section 2.4, where, due to the fact that the responses have a density, the correction term $G(s)$ appearing in (8) is not needed. For the weighted estimators, we used the Tukey's bisquare weight function with tuning constant $c = \chi^2_{k-1,0.95}$. The weights were computed over the robust Mahalanobis distances $d(\mathbf{z}, \widehat{\boldsymbol{\mu}}_{\mathbf{z}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}) = \left\{(\mathbf{z} - \widehat{\boldsymbol{\mu}}_{\mathbf{z}})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}^{-1}(\mathbf{z} - \widehat{\boldsymbol{\mu}}_{\mathbf{z}})\right\}^{1/2}$, where $(\widehat{\boldsymbol{\mu}}_{\mathbf{z}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})$ stands for an S-estimator with breakdown point 0.5 using 1000 sub-samples. Form now on, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\tau}$ will refer to the robust weighted estimators.

The results concerning the behaviour of the estimators when $k = 3$ can be found in [2] and lead to conclusions analogous to those reported for the Poisson regression model. For that reason, we only report here results regarding the performance of the diagnostic measure, for $k = 3$, 5 and 10, which are not reported there. The values of the regression parameter equal, in each situation, $\boldsymbol{\beta} = 0$ when $k = 3$, while for $k = 5$ and 10 we choose $\boldsymbol{\beta} = \mathbf{1}_k/\sqrt{k}$ with $\mathbf{1}_k$ the vector with all its components equal to one. We considered three models for the missing probability, $p \equiv 1$, that is, no missing responses are introduced, a logistic model for missingness $p(\mathbf{x}) = 1/(1+\exp(-\mathbf{z}^{\mathrm{T}}\boldsymbol{\lambda}-2))$ and $p(\mathbf{x}) = 0.4+0.5(\cos(\mathbf{z}^{\mathrm{T}}\boldsymbol{\lambda}+0.4))^2$, both with $\boldsymbol{\lambda} = 2 \times \mathbf{1}_{k-1}$.

It is worth noting that for the log–gamma regression model the matrix $\mathbf{B}(\boldsymbol{\beta}, \tau) = \mathbb{E}_F\left(\Psi^2_{\mathrm{ML}}(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tau)\, w^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$ equals $\mathbf{B}(\boldsymbol{\beta}, \tau) = (1/\tau)\mathbb{E}_F\left(w^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right) = (1/\tau)\mathbf{D}$, so that when computing the diagnostic measure and the cut-off points, we estimate $\mathbf{B}(\boldsymbol{\beta}, \tau)$ as $\widehat{\mathbf{B}} = (1/\widehat{\tau})\widehat{\mathbf{D}}$, where the estimator $\widehat{\mathbf{D}}$ of the matrix $\mathbb{E}_F\left(w^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$ is evaluated using a hard rejection weight function, that is, $w(\mathbf{x}) = 1$ when $d^2(\mathbf{z}, \widehat{\boldsymbol{\mu}}_{\mathbf{z}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}) \le \chi^2_{k-1,0.95}$ and $w_1(\mathbf{x}) = 0$ elsewhere, with $d(\mathbf{z}, \widehat{\boldsymbol{\mu}}_{\mathbf{z}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})$ the robust Mahalanobis distance defined above.

Once the robust estimators $(\widehat{\boldsymbol{\beta}}, \widehat{\tau})$ are computed, an observation is classified as an outlier if $\mathrm{DM}(y, \mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\tau}) > c_{k,1-\alpha}$. The values of $c_{k,1-\alpha}$ have been obtained from simulation experiments to ensure that, in the absence of outliers, the proportion of correct observations mislabelled as outliers is approximately equal to 100α%. We select two values for α, $\alpha = 0.05$ and $\alpha = 0.01$. To assess the quality of the detection measure, Tables 3 and 4 report the mean and median over replications of the proportion of outliers detected for each dimension and each missing scheme, respectively. It is worth noting that when missing responses arise, we compute the estimators using the simplified robust weighted estimators, which means that some of the outliers may be missing by chance. However, as in Section 5.2, when computing $\mathrm{DM}(y, \mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\tau}) > c_{k,1-\alpha}$, we force δ_i to be equal to 1, when an outlier appears, to have an idea of the misclassification error obtained when classifying the outliers. From these Tables, we observe that as the dimension increases low and moderate outliers became more difficult to detect, in particular, when missing responses arise. This behaviour can be explained by the fact that, since we have preserved the same sample size, the effective number of observations may not be enough to estimate such a large number of parameters. It should be taken into account that, for both the logistic and the cosine missingness probabilities, the mean number of missing responses is 25% and 35%, respectively, when $k = 3$, while it reaches 36% in dimension $k = 10$ in both cases.

Table 3
Mean over replications of the proportion of outliers detected.

k	$\alpha = 0.05$						$\alpha = 0.01$					
	$m_0 = 0.5$			$m_0 = 2.5$			$m_0 = 0.5$			$m_0 = 2.5$		
	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$
	$p \equiv 1$											
3	0	1	1	1	1	1	0	1	1	1	1	1
5	0	0.909	1	1	1	1	0	0.621	0.999	1	1	1
10	0	0.147	1	0.999	0.999	1	0	0.147	0.999	0.999	0.999	1
	$p(\mathbf{x}) = 1/(1 + \exp(-\mathbf{z}^{\mathrm{T}}\boldsymbol{\lambda} - 2))$											
3	0	0.991	1	1	1	1	0	0.926	1	1	1	1
5	0	0.433	0.998	1	1	1	0	0.329	0.998	1	1	1
10	0	0.161	0.998	0.994	1	1	0	0.116	0.994	0.982	1	1
	$p(\mathbf{x}) = 0.4 + 0.5(\cos(\mathbf{z}^{\mathrm{T}}\boldsymbol{\lambda} + 0.4))^2$											
3	0	0.981	1	1	1	1	0	0.898	1	1	1	1
5	0	0.478	0.999	1	1	1	0	0.350	0.999	1	1	1
10	0	0.198	1	1	1	1	0	0.148	1	0.983	1	1

Table 4
Median over replications of the proportion of outliers detected.

k	$\alpha = 0.05$						$\alpha = 0.01$					
	$m_0 = 0.5$			$m_0 = 2.5$			$m_0 = 0.5$			$m_0 = 2.5$		
	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$	$x_0 = 1$	$x_0 = 3$	$x_0 = 10$
	$p \equiv 1$											
3	0	1	1	1	1	1	0	1	1	1	1	1
5	0	1	1	1	1	1	0	1	1	1	1	1
10	0	0	1	1	1	1	0	0	1	1	1	1
	$p(\mathbf{x}) = 1/(1 + \exp(-\mathbf{z}^{\mathrm{T}}\boldsymbol{\lambda} - 2))$											
3	0	1	1	1	1	1	0	1	1	1	1	1
5	0	0	1	1	1	1	0	0	1	1	1	1
10	0	0	1	1	1	1	0	0	1	1	1	1
	$p(\mathbf{x}) = 0.4 + 0.5(\cos(\mathbf{z}^{\mathrm{T}}\boldsymbol{\lambda} + 0.4))^2$											
3	0	1	1	1	1	1	0	1	1	1	1	1
5	0	0	1	1	1	1	0	0	1	1	1	1
10	0	0	1	1	1	1	0	0	1	1	1	1

6. Example: epilepsy data

Thall and Vail [27] reported data from a clinical trial of 59 patients with epilepsy, see also Breslow [5]. The patients were randomized to receive either the anti-epileptic drug progabide or a placebo. This data set has been considered in the robust literature such as [19,15] in the setting of Poisson regression taking as response the total number of epileptic attacks patients have during the four follow-up periods. In our study, we chose as response variable Y_4, which records the number of epilepsy attacks patients have during the fourth follow-up period. We treat this response variable as a Poisson one. The explanatory variables $\mathbf{x}$ are

$x_1 = \mathbf{Age10}$: the age of the patients in years divided by 10
$x_2 = \mathbf{Base4}$: number of epileptic attacks recorded during an 8 week period prior to randomization divided by 4
$x_3 = \mathbf{Trt}$: the binary indicators for the progabide group.

A term $x_4 = \mathbf{Base4} * \mathbf{Trt}$ was introduced in order to take into account the interaction between these two factors.

The CUBIF estimator introduced by Künsch et al. [18] applied to the 59 observations identifies three possible outliers, labelled as observations 15, 18 and 49. It is worth noticing that the Pearson's fit statistic, the usual measure of overdispersion, is 2.543 when computed on the whole sample. It drops to 1.982 when the fit is computed without these three outliers, revealing only a mild overdispersion.

We compute the weighted general M-estimator, $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$, with weights based on the Tukey's bisquare function with tuning constant $\chi^2_{2,0.975}$. The weights were computed over the robust Mahalanobis distances of the continuous covariates (x_1, x_2), calculated with an S-estimator with breakdown point 0.5 using 50 subsamples and 4 iterations in each one. The same loss function as in Section 4.1 is considered to bound the deviances. Fig. 5 gives the Q–Q plots of the deviances corresponding to the CUBIF estimator, $\widehat{\boldsymbol{\beta}}_{\mathrm{CUBIF}}$, and to $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$. The left panel of Fig. 6 corresponds to the logarithm of influence diagnostic measure, $\log(\mathrm{DM}(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\mathrm{GM}}))$, against the indexed number of the observed responses for all the data. The logarithm is a suitable scale due to the large values obtained. Note that in this case, the term $\Psi^2_{\mathrm{ML}}(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ in $\mathrm{DM}(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}_{\mathrm{GM}})$ is related to the square of the Pearson residuals, so that the matrix $\mathbf{B}(\boldsymbol{\beta}) = \mathbb{E}_F\left(\Psi^2_{\mathrm{ML}}(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\, w^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$ equals $\mathbf{B}(\boldsymbol{\beta}) = \mathbb{E}_F\left(H(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})w^2(\mathbf{x})p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$.

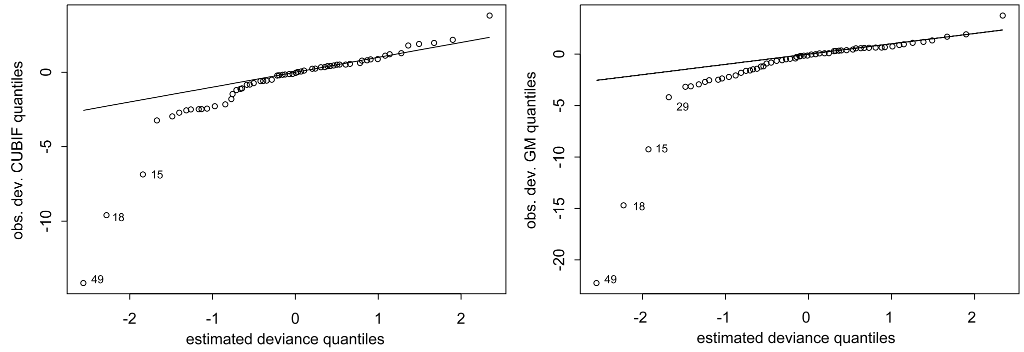

Fig. 5. Epilepsy data: Q–Q plots of the deviances corresponding to $\widehat{\boldsymbol{\beta}}_{\text{CUBIF}}$ on the left and to $\widehat{\boldsymbol{\beta}}_{\text{GM}}$ on the right.

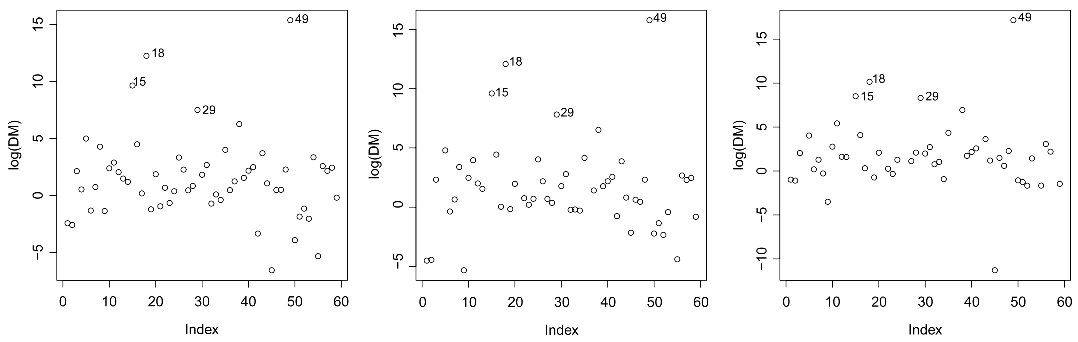

Fig. 6. Epilepsy data: Logarithm of the Diagnostic Measure, log(DM) versus the index of the observation, for $p \equiv 1$ on the left, for $p_{0.3}(\mathbf{x})$ in the middle panel and for $p_{0.5}(\mathbf{x})$ on the right.

Table 5
Analysis of Epilepsy data. Original data set.

	$\widehat{\boldsymbol{\beta}}_{\text{ML}}$	$\widehat{\boldsymbol{\beta}}_{\text{ML}}^{-\text{OUT}}$	$\widehat{\boldsymbol{\beta}}_{\text{CUBIF}}$	$\widehat{\boldsymbol{\beta}}_{\text{GM}}$
Age10	0.156	0.147	0.063	0.052
Base4	0.083	0.152	0.142	0.163
Trt	−0.367	−0.212	−0.168	−0.217
Base4 $*$ **Trt**	0.007	−0.012	−0.011	−0.015
Intercept	0.791	0.308	0.606	0.580

It is evident that $\widehat{\boldsymbol{\beta}}_{\text{GM}}$ provides a better fit and from both the Q–Q plot and the diagnostic measure defined, we identify observations labelled 15, 18 and 49 as possible severe outliers and observation 29 as a possible mild outlier. The atypicity of observation 18 is more evident from the plot of the log(DM). Effectively, the deviance residual shown in the Q–Q plot for observation 18 is much smaller than that of observation 49, while even if $\log(\text{DM}(y_{18}, \mathbf{x}_{18}, \widehat{\boldsymbol{\beta}}_{\text{GM}}))$ is smaller than $\log(\text{DM}(y_{49}, \mathbf{x}_{49}, \widehat{\boldsymbol{\beta}}_{\text{GM}}))$, it still provides a large value for $\text{DM}(y_{18}, \mathbf{x}_{18}, \widehat{\boldsymbol{\beta}}_{\text{GM}})$, showing that the effect of observation 18 is enlarged due to its leverage. Table 5 gives the values of the estimated regression parameters computed with the whole sample. The maximum likelihood estimator, $\widehat{\boldsymbol{\beta}}_{\text{ML}}$, the CUBIF estimator, $\widehat{\boldsymbol{\beta}}_{\text{CUBIF}}$, and the weighted general M-estimator, $\widehat{\boldsymbol{\beta}}_{\text{GM}}$, are reported. Besides, we compute the maximum likelihood estimator without the four outlying observations. It is worth noticing that the usual measure of overdispersion computed without the four outliers drops to 1.897.

To evaluate the performance of the proposed estimators when missing data arise in the responses, we consider several missing probability schemes to introduce artificially missing responses in the data, as in [17]. Our goal for this example is to obtain estimates of the parameters with two missing data fractions and compare the results with the complete data estimates. We introduce missing responses among the non outlying points following a logistic missing probability model given by $p_\lambda(\mathbf{x}) = 1/(1 + (\exp(\lambda * \textbf{Age10} - 2.8)))$, with $\lambda = 0.5$ and 0.3, which results in 10 and 5 missing responses, respectively (approximately 20% and 10% of missing responses, respectively). The results corresponding to the maximum likelihood, $\widehat{\boldsymbol{\beta}}_{\text{ML}}$, and weighted general M-estimates, $\widehat{\boldsymbol{\beta}}_{\text{GM}}$, applied to the sample with no missing responses and to the incomplete data sets are shown in Table 6. The robust estimators allow one to compute the diagnostic measure for the two missing responses schemes considered. In Fig. 6, we plot the indexed number of the observed responses against the computed value of log(DM). As we can see, the measure DM can detect the more severe outlying points and also the mild outlier labelled 29, independently of the missing probability pattern. Besides, these four observations are detected as outlying points in the boxplots of log(DM) in the three analysed situations. This confirms the usefulness of DM as an empirical tool for the purpose of identifying potential outliers.

Table 6
Epilepsy data: Missing responses introduced according to $p(\mathbf{x}) = 1$ and $p_\lambda(\mathbf{x}) = 1/(1 + (\exp(\lambda * \mathbf{Age10} - 2.8)))$.

	$p \equiv 1$		$p_{0.3}(\mathbf{x})$		$p_{0.5}(\mathbf{x})$	
	$\widehat{\boldsymbol{\beta}}_{\text{ML}}$	$\widehat{\boldsymbol{\beta}}_{\text{GM}}$	$\widehat{\boldsymbol{\beta}}_{\text{ML}}$	$\widehat{\boldsymbol{\beta}}_{\text{GM}}$	$\widehat{\boldsymbol{\beta}}_{\text{ML}}$	$\widehat{\boldsymbol{\beta}}_{\text{GM}}$
Age10	0.156	0.052	0.187	−0.015	0.271	0.043
Base4	0.083	0.163	0.083	0.159	0.079	0.119
Trt	−0.367	−0.217	−0.300	−0.234	−0.239	−0.643
Base4 ∗ **Trt**	0.007	−0.015	0.006	0.007	0.010	0.054
Intercept	0.791	0.580	0.681	0.737	0.438	0.794

7. Concluding remarks

We have introduced resistant estimators for the regression parameter under a generalized regression model, when there are missing responses and it can be suspected that anomalous observations are present in the sample. The estimators considered are Fisher-consistent and, thus, lead to strongly consistent estimators.

The simulation study confirms the expected inadequate behaviour of the classical estimators and the possible sensitivity of the unweighted robust estimators in the presence of mild outliers according to the parameter values. The proposed robust weighted procedures for the regression parameter perform quite similarly under the central model or under the contaminations studied.

As is well known, the influence function allows one both to study the influence of a given observation on estimators of the regression parameter and to compute heuristically the asymptotic variances. For the proposed estimators, the asymptotic variances computed show that in some situations the simplified estimators are more efficient than the propensity ones. Besides, the influence function shows that, even when M-estimators may have a good performance in some settings, they may be sensitive to high leverage points for some missing probabilities such as the logistic one. In this sense, weighted M-estimators turn out to be an alternative to be considered even if some loss of efficiency is expected.

Based on the influence function, a diagnostic measure is defined. Through simulated examples and a real data set its usefulness to detect atypical points is illustrated.

Acknowledgments

This work began while Isabel Rodrigues was visiting the Instituto de Cálculo of the Universidad de Buenos Aires and received financial support from Portuguese National Funds through FCT (Fundação para a Ciência e a Tecnologia) under the scope of project PEst-OE/MAT/UI0822/2011 and from *Fundação Calouste Gulbenkian* (FCG). This research was also partially supported by Grants PIP 112-200801-00216 from CONICET, PICT 0821 from ANPCYT and W276 from the Universidad de Buenos Aires at Buenos Aires, Argentina and by the joint cooperation program ANPCYT–FCT PO/09/05–Proc. 44100.

Appendix

Proof of Lemma 2.1. We begin by proving the Fisher consistency of $\widetilde{\boldsymbol{\beta}}(F)$. Using (12) and the independence between δ and u, we have that

$$\frac{\mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{d^*(y,\mathbf{x},\mathbf{b})}}{\sigma(\tau,\mathbf{b})}\right)\right)}{\mathbb{E}(\delta)} = \frac{\mathbb{E}_F\left(p(\mathbf{x})\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\widetilde{\boldsymbol{\beta}}-\mathbf{b})}}{\sigma(\tau,\mathbf{b})}\right)\right)}{\mathbb{E}(p(\mathbf{x}))}.$$

Note that $\sigma(\tau, \mathbf{b})$ is a function of $\widetilde{\boldsymbol{\beta}} - \mathbf{b}$. Using Lemma 1 in [3], we get that for any fixed c

$$\begin{aligned}\mathbb{E}_F\left(p(\mathbf{x})\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\widetilde{\boldsymbol{\beta}}-\mathbf{b})}}{c}\right)\Bigg|\mathbf{x}\right) &= p(\mathbf{x})\mathbb{E}_F\left(\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\widetilde{\boldsymbol{\beta}}-\mathbf{b})}}{c}\right)\Bigg|\mathbf{x}\right)\\ &\geq p(\mathbf{x})\mathbb{E}_F\left(\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\mathbf{0})}}{c}\right)\Bigg|\mathbf{x}\right).\end{aligned} \tag{18}$$

From (10), (18) and the fact that ϕ is strictly increasing in the set where it is not constant, we get that for any $\mathbf{b} \neq \widetilde{\boldsymbol{\beta}}$

$$\mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\mathbf{0})}}{\sigma(\tau,\mathbf{b})}\right)\Bigg|\mathbf{x}\right) < \mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\widetilde{\boldsymbol{\beta}}-\mathbf{b})}}{\sigma(\tau,\mathbf{b})}\right)\Bigg|\mathbf{x}\right) = b\,\mathbb{E}(p(\mathbf{x})).$$

Using that $\mathbb{E}_F\left(\delta\phi\left(\sqrt{\widetilde{d}(u,\mathbf{x},\mathbf{0})}/\sigma\right)\right)$ is decreasing in σ and that $\mathbb{E}_F\left(\delta\phi\left(\sqrt{\widetilde{d}(u,\mathbf{x},0)}/\sigma(\tau,\widetilde{\boldsymbol{\beta}})\right)\right) = b\,\mathbb{E}(p(\mathbf{x}))$, we get that $\sigma(\tau,\widetilde{\boldsymbol{\beta}}) < \sigma(\tau,\mathbf{b})$, which implies the Fisher-consistency of the functional. Using analogous arguments the Fisher-consistency of $\boldsymbol{\beta}(F)$ can be derived. □

Proof of Proposition 3.1. We will show that $S_n(\widehat{\boldsymbol{\beta}},\widehat{\tau}) - S_n(\widehat{\boldsymbol{\beta}},\tau) \xrightarrow{a.s.} 0$. Note that $\mathbb{E}\left(S_n(\mathbf{b},t)\right) = S(\mathbf{b},t)$. Using standard empirical process arguments, from **A3**, we have that

$$V_n = \sup_{\mathbf{b},t}\left|\frac{1}{n}\sum_{i=1}^n \delta_i\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)w_1(\mathbf{x}_i) - \mathbb{E}\left(\delta_i\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)w_1(\mathbf{x}_i)\right)\right| \xrightarrow{a.s.} 0.$$

Therefore, we get

$$\begin{aligned}\sup_{\mathbf{b}}\left|S_n(\mathbf{b},\widehat{\tau}) - S_n(\mathbf{b},\tau)\right| &\le \sup_{\mathbf{b}}\left|S_n(\mathbf{b},\widehat{\tau}) - S(\mathbf{b},\widehat{\tau})\right| + \sup_{\mathbf{b}}\left|S(\mathbf{b},\widehat{\tau}) - S(\mathbf{b},\tau)\right| \\ &\quad + \sup_{\mathbf{b}}\left|S_n(\mathbf{b},\tau) - S(\mathbf{b},\tau)\right| \le 2V_n + \sup_{\mathbf{b}}\left|S(\mathbf{b},\widehat{\tau}) - S(\mathbf{b},\tau)\right|.\end{aligned}$$

The equicontinuity of $S(\mathbf{b},\tau)$ and the consistency of $\widehat{\tau}$, entail that, $\sup_{\mathbf{b}}|S_n(\mathbf{b},\widehat{\tau}) - S_n(\mathbf{b},\tau)| \xrightarrow{a.s.} 0$. Thus, $S_n(\widehat{\boldsymbol{\beta}},\widehat{\tau}) - S_n(\widehat{\boldsymbol{\beta}},\tau) \xrightarrow{a.s.} 0$ so, the sequence of estimators $\widehat{\boldsymbol{\beta}}$ satisfies that $\inf_{\mathbf{b}} S_n(\mathbf{b},\tau) - S_n(\widehat{\boldsymbol{\beta}},\tau) \xrightarrow{a.s.} 0$, which allows us to apply the results from [16]. □

Proof of Proposition 3.2. As in the proof of Proposition 3.1, let us show that $S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\tau,p) \xrightarrow{a.s.} 0$.

First assume that **A5**(b) holds, then, using standard empirical process arguments, from **A3**, we have that

$$V_n = \sup_{\mathbf{b},t,\boldsymbol{\lambda}}\left|\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{G_{\mathrm{P}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\lambda})}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)w_1(\mathbf{x}_i) - \mathbb{E}\left(\frac{\delta_i}{G_{\mathrm{P}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\lambda})}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)w_1(\mathbf{x}_i)\right)\right| \xrightarrow{a.s.} 0.$$

Therefore, since $\widehat{p}(\mathbf{x}) = p_{\boldsymbol{\lambda}}(\mathbf{x}) = G_{\mathrm{P}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda})$, we get

$$\begin{aligned}\sup_{\mathbf{b}}\left|S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)\right| &\le \sup_{\mathbf{b}}\left|S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},\widehat{p})\right| + \sup_{\mathbf{b}}\left|S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P}}(\mathbf{b},\tau,p)\right| \\ &\quad + \sup_{\mathbf{b}}\left|S_{\mathrm{P},n}(\mathbf{b},\tau,p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)\right| \le 2V_n + \sup_{\mathbf{b}}\left|S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P}}(\mathbf{b},\tau,p)\right|.\end{aligned}$$

Using the equicontinuity of $S_{\mathrm{P}}(\mathbf{b},\tau,p)$ and the consistency of $\widehat{\tau}$ and $\widehat{\boldsymbol{\lambda}}$, we get that, when **A5**(b) holds, $\sup_{\mathbf{b}}|S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)| \xrightarrow{a.s.} 0$.

Under **A5**(a), we obtain easily from **A1**, **A2** and **A4** that $S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},p) \xrightarrow{a.s.} 0$. Again, using standard empirical process arguments, from **A3**, we have that

$$V_n = \sup_{\mathbf{b},t}\left|\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{p(\mathbf{x}_i)}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)w_1(\mathbf{x}_i) - \mathbb{E}\left(\frac{\delta_i}{p(\mathbf{x}_i)}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)\right)w_1(\mathbf{x}_i)\right| \xrightarrow{a.s.} 0,$$

which implies that

$$\begin{aligned}\sup_{\mathbf{b}}\left|S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)\right| &\le \sup_{\mathbf{b}}\left|S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},p)\right| + \sup_{\mathbf{b}}\left|S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)\right| \\ &\quad + \sup_{\mathbf{b}}\left|S_{\mathrm{P},n}(\mathbf{b},\tau,p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)\right| \le 2V_n + \sup_{\mathbf{b}}\left|S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)\right|\end{aligned}$$

and so, using the consistency of $\widehat{\tau}$ and the equicontinuity of $S_{\mathrm{P}}(\mathbf{b},\tau,p)$, we obtain that, when **A5**(a) holds, $\sup_{\mathbf{b}}|S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)| \xrightarrow{a.s.} 0$.

Hence, $S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\tau,p) \xrightarrow{a.s.} 0$ which implies that the sequence of estimators $\widehat{\boldsymbol{\beta}}$ satisfies that $\inf_{\mathbf{b}} S_{\mathrm{P},n}(\mathbf{b},\tau,p) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\tau,p) \xrightarrow{a.s.} 0$ and the results from [16] can be applied. □

Proof of Theorem 4.1. To get the influence function of $\boldsymbol{\beta}(F)$, note that $\mathbb{E}_{F_{\mathbf{z}_0,\epsilon}}\left(\delta\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(F_{\mathbf{z}_0,\epsilon}),\tau(F_{\mathbf{z}_0,\epsilon})\right)w_1(\mathbf{x})\mathbf{x}\right) = \mathbf{0}_k$ implies

$$\mathbf{0}_k = (1-\epsilon)\mathbb{E}_F\left(\delta\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(F_{\mathbf{z}_0,\epsilon}),\tau(F_{\mathbf{z}_0,\epsilon})\right)w_1(\mathbf{x})\mathbf{x}\right) + \epsilon\,\delta_0\Psi\left(y_0,\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}(F_{\mathbf{z}_0,\epsilon}),\tau(F_{\mathbf{z}_0,\epsilon})\right)w_1(\mathbf{x}_0)\mathbf{x}_0. \tag{19}$$

Therefore, differentiating (19) with respect to ϵ and evaluating at $\epsilon = 0$, we obtain that

$$\begin{aligned}\mathbf{0}_k &= \mathbb{E}_F\left(\delta\chi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(F),\tau(F)\right)w_1(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)\mathrm{IF}(\mathbf{z}_0,\boldsymbol{\beta},F) + \frac{\partial}{\partial t}\mathbb{E}_F\left(\delta\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(F),t\right)w_1(\mathbf{x})\mathbf{x}\right)\Big|_{t=\tau(F)}\mathrm{IF}(\mathbf{z}_0,\tau,F) \\ &\quad + \delta_0\Psi\left(y_0,\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}(F),\tau(F)\right)w_1(\mathbf{x}_0)\mathbf{x}_0 - \mathbb{E}_F\left(\delta\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(F),\tau(F)\right)w_1(\mathbf{x})\mathbf{x}\right).\end{aligned}$$

Using the MAR condition (1) and **H3**, we conclude the proof. □

Proof Theorem 4.2. Following the same steps as in the proof of Theorem 4.1, we need to compute the influence function of $\boldsymbol{\beta}_{\mathrm{P}}(F)$. From $\mathbb{E}_{F_{\mathbf{z}_0,\epsilon}}\left(\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F_{\mathbf{z}_0,\epsilon}),\tau_{\mathrm{P}}(F_{\mathbf{z}_0,\epsilon})\right)w_1^*(\mathbf{x},F_{\mathbf{z}_0,\epsilon})\delta\mathbf{x}\right)=\mathbf{0}_k$ we have that

$$\mathbf{0}_k=(1-\epsilon)\mathbb{E}_F\left(\delta\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F_{\mathbf{z}_0,\epsilon}),\tau_{\mathrm{P}}\big(F_{\mathbf{z}_0,\epsilon}\big)\right)w_1^*(\mathbf{x},F_{\mathbf{z}_0,\epsilon})\mathbf{x}\right)+\epsilon\,\delta_0\Psi\left(y_0,\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F_{\mathbf{z}_0,\epsilon}),\tau_{\mathrm{P}}\big(F_{\mathbf{z}_0,\epsilon}\big)\right)w_1^*(\mathbf{x}_0,F_{\mathbf{z}_0,\epsilon})\mathbf{x}_0. \quad (20)$$

Differentiating (20) with respect to ϵ and evaluating at $\epsilon=0$, we obtain

$$\begin{aligned}\mathbf{0}_k&=\mathbb{E}_F\left(\chi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F),\tau_{\mathrm{P}}(F)\right)w_1^*(\mathbf{x},F)p(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)\mathrm{IF}(\mathbf{z}_0,\boldsymbol{\beta}_{\mathrm{P}},F)\\&\quad+\mathbb{E}_F\left(\frac{\partial}{\partial t}\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F),t\right)\Big|_{t=\tau_{\mathrm{P}}(F)}w_1^*(\mathbf{x},F)p(\mathbf{x})\mathbf{x}\right)\mathrm{IF}(\mathbf{z}_0,\tau_{\mathrm{P}},F)\\&\quad+\delta_0\Psi\left(y_0,\mathbf{x}_0^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F),\tau_{\mathrm{P}}(F)\right)w_1^*(\mathbf{x}_0,F)\mathbf{x}_0-\mathbb{E}_F\left(\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F),\tau_{\mathrm{P}}(F)\right)w_1^*(\mathbf{x},F)p(\mathbf{x})\mathbf{x}\right)\\&\quad-\mathbb{E}_F\left(\frac{1}{G(\mathbf{x},\boldsymbol{\lambda})}\Psi\left(y,\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{P}}(F),\tau_{\mathrm{P}}(F)\right)w_1^*(\mathbf{x},F)p(\mathbf{x})\mathbf{x}\,G_2(\mathbf{x},\boldsymbol{\lambda})^{\mathrm{T}}\right)\mathrm{IF}(\mathbf{z}_0,\boldsymbol{\lambda},F)\end{aligned}$$

where $G_2(\mathbf{x},\boldsymbol{\lambda})=\partial G(\mathbf{x},\mathbf{u})/\partial\mathbf{u}|_{\mathbf{u}=\boldsymbol{\lambda}}$ and the proof follows from **H5** and **H6**. □

References

[1] A. Bergesio, V. Yohai, Projection estimates for generalized linear models, J. Amer. Statist. Assoc. 106 (2011) 661–671.
[2] A. Bianco, G. Boente, I.M. Rodrigues, 2011 Resistant estimators in Poisson and Gamma models with applications to outlier detection. Available on www.ic.fcen.uba.ar/preprints/biancoboenterodrigues2011.pdf.
[3] A. Bianco, M. García Ben, V. Yohai, Robust estimation for linear regression with asymmetric errors, Canad. J. Statist. 33 (2005) 511–528.
[4] A. Bianco, V. Yohai, Robust estimation in the logistic regression model, in: Lecture Notes in Statistics, vol. 109, Springer-Verlag, New York, 1996, pp. 17–34.
[5] N.E. Breslow, Generalized linear models: Checking assumptions and strengthening conclusions, Statist. Appl. 8 (1996) 23–41.
[6] E. Cantoni, E. Ronchetti, Robust inference for generalized linear models, J. Amer. Statist. Assoc. 96 (2001) 1022–1030.
[7] E. Cantoni, E. Ronchetti, A robust approach for skewed and heavy–tailed outcomes in the analysis of health care expenditures, J. Health Economics 25 (2006) 198–213.
[8] C. Croux, G. Haesbroeck, Implementing the Bianco and Yohai estimator for logistic regression, Comput. Statist. Data Anal. 44 (2003) 273–295.
[9] D.L. Donoho, P.J. Huber, The notion of breakdown point, in: Peter J. Bickel, Kjell A. Doksum, J.L. Hodges (Eds.), A Festschrift for Erich L. Lehmann, Wadsworth, Belmont, CA, 1983.
[10] L.T. Fernholz, Von Mises calculus for statistical functionals, in: Lecture Notes in Statistics, vol. 19, Springer Verlag, New York, 1983.
[11] M. García Ben, V.J. Yohai, Quantile-Quantile plot for deviance residuals in the generalized linear model, J. Comp. Graph. Statist. 13 (2004) 36–47.
[12] F.R. Hampel, A general qualitative definition of robustness, Ann. Math. Statist. 42 (1971) 1887–1896.
[13] S. Heritier, E. Ronchetti, Robust bounded-influence tests in general parametric models, J. Amer. Statist. Assoc. 89 (1994) 897–904.
[14] S. Heritier, E. Cantoni, S. Copt, M.P. Victoria-Feser, Robust methods in biostatistics, in: Wiley Series in Probability and Statistics, Wiley, 2009.
[15] S. Hosseinian, 2009 Robust inference for generalized linear models: binary and Poisson regression, PhD. Thesis, Ecole Polytechnique Fédérale de Lausanne.
[16] P.J. Huber, The behavior of maximum likelihood estimates under nonstandard conditions, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 221–233.
[17] J. Ibrahim, Incomplete data in generalized linear models, J. Amer. Statist. Assoc. 85 (1990) 765–769.
[18] H. Künsch, L. Stefanski, R. Carroll, Conditionally unbiased bounded influence estimation in general regression models with applications to generalized linear models, J. Amer. Statist. Assoc. 84 (1989) 460–466.
[19] R. Maronna, D. Martin, V. Yohai, Robust Statistics: Theory and Methods, Wiley, New York, 2006.
[20] P. McCullagh, J. Nelder, Generalized Linear Models, 2nd edn., Champman and Hall, London, 1989.
[21] J. Nelder, R. Wedderburn, Generalized linear models, J. Royal Statist. Soc., Series A 135 (1972) 370–384.
[22] G. Pison, P.J. Rousseeuw, P. Filzmoser, C. Croux, A robust version of principal factor analysis, in: J. Bethlehem, P. van der Heijden (Eds.), Compstat: Proceedings in Computational Statistics, Physica–Verlag, Heidelberg, 2000, pp. 385–390.
[23] P.J. Rousseeuw, M. Debruyne, S. Engelen, M. Hubert, Robustness and outlier detection in chemometrics, Critical Rev. Anal. Chem. 36 (2006) 221–242.
[24] P.J. Rousseeuw, B.C. van Zomeren, Unmasking multivariate outliers and leverage points, J. Amer. Statist. Assoc. 85 (1990) 633–639.
[25] L. Stefanski, R. Carroll, D. Ruppert, Bounded score functions for generalized linear models, Biometrika 73 (1986) 413–424.
[26] M. Sued, J. Yohai, 2011 Robust location estimates with missing data. Unpublished manuscript. Available at http://mate.dm.uba.ar/vyohai/Pub.htm.
[27] P.F. Thall, S.C. Vail, Some covariance models for longitudinal count data with overdispersion, Biometrics 46 (1990) 657–671.