Journal of Statistical Planning and Inference 146 (2014) 31–48

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Robust estimators for generalized linear models

Marina Valdora [a], Víctor J. Yohai [b,c,*]

[a] Universidad de Buenos Aires, Argentina
[b] Universidad de Buenos Aires, Argentina
[c] CONICET, Argentina

ARTICLE INFO

Article history:
Received 5 February 2013
Received in revised form
25 September 2013
Accepted 27 September 2013
Available online 9 October 2013

Keywords:
M-estimators
Transformations
Breakdown point

ABSTRACT

In this paper we propose a family of robust estimators for generalized linear models. The basic idea is to use an M-estimator after applying a variance stabilizing transformation to the response. We show the consistency and asymptotic normality of these estimators. We also obtain a lower bound for their breakdown point. A Monte Carlo study shows that the proposed estimators compare favorably with respect to other robust estimators for generalized linear models with Poisson response and log link.

1. Introduction

Generalized linear models (GLMs) are widely used in data analysis. It is well known that the maximum likelihood estimator for these models is very sensitive to outliers. To overcome this problem, several robust estimators for GLM have been proposed. Künsch et al. (1989) derived optimal conditionally unbiased bounded influence (CUBIF) estimators. These estimators are highly robust for a small fraction of outlier contamination. However, Maronna et al. (1979) showed that in the case of a linear model, the breakdown point of these estimators tends to 0 when the number of regressors tends to infinity. Cantoni and Ronchetti (2001) defined robust estimators for GLM which can be considered a robustification of the quasi-likelihood estimators introduced by Wedderburn (1974). These estimators are defined only by the estimating equations and this forces the use of monotone score functions. As a consequence, as we will see in the Monte Carlo study in Section 8, their robustness is very limited. Morgenthaler (1992) also proposed a robustification of the quasi-likelihood estimators but using an l_1 measure of fit. Therefore the corresponding score function, the sign function, is monotone too. Bergesio and Yohai (2011) introduced projection estimators for GLM which are highly robust but their computation requires algorithms of high complexity. Since these estimators are not asymptotically normal, they propose one-step M-estimators starting at the P-estimator. These estimators keep most of the properties of the P-estimators and, in addition, they are asymptotically normal. Another class of estimators are the M-estimators proposed by Bianco and Yohai (1996) and further studied by Croux and Haesbroeck (2003). Bianco et al. (2013) proposed general M-estimators for GLM for data sets with missing values in the responses. For GLM models where F_λ is the Bernoulli family of distributions we can also cite works of Carroll and Pederson (1993), Christmann (1994), Rousseeuw and Christmann (2003), Bondell (2005) and Čížek (2008).

In this paper we introduce a rather simple and highly robust family of estimators for GLM. The proposed estimators are redescending M-estimators applied to transformed responses. The purpose of transforming the responses is to stabilize

* Corresponding author at: Universidad de Buenos Aires, Argentina. Tel.: +54 11 4576 3375
E-mail addresses: victoryohai@gmail.com, vyohai@dm.uba.ar (V.J. Yohai).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.09.016

their variances to an almost constant value and so allowing a correct scaling of the loss function used to define the M-estimator.

The proposed estimators are not applicable to the case that F_λ is the Bernoulli family of distributions. It is easy to see that in this case the proposed estimator coincides with an ordinary M-estimator where y is not transformed.

In Section 2 we introduce the M-estimators based on transformations (MT) and the weighted M-estimators based on transformations (WMT) and obtain the variance stabilizing transformations required for some families of distributions. In Sections 3 and 4 we study the consistency and asymptotic normality of WMT-estimators respectively. In Section 5 we obtain a lower bound for the asymptotic breakdown point of MT-estimators. In Section 6 we report the results of a Monte Carlo study to compare the performance of MT- and WMT-estimators to that of other existing estimators for Poisson regression. In Section 7 we consider a real data set and compare the fit given by the MT-estimator with other existing estimators. In Section 8 we present the conclusions. Finally, Appendix A is an appendix containing all the proofs.

2. Proposed estimators

2.1. Definition of M-estimators based on transformations

We consider a generalized linear model (GLM) where $y \in \mathbb{R}$ is the response and $\mathbf{x} = (x_1, \ldots, x_p)'$ is a vector of explanatory variables. It is assumed that

$$y|\mathbf{x} \sim F_\lambda, \tag{1}$$

where F_λ is a discrete or continuous exponential family of distributions in $\mathbb{R}$, $\lambda \in \Lambda \subset \mathbb{R}$ with the same support D and

$$\lambda = g(\boldsymbol{\beta}_0'\mathbf{x}), \tag{2}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is unknown and $g : \mathbb{R} \to \mathbb{R}$ is a known link function. We will assume that λ takes values in an interval $(\lambda^{(1)}, \lambda^{(2)})$ where $\lambda^{(1)}$ may be $-\infty$ and $\lambda^{(2)} = +\infty$. We will also assume that $g : \mathbb{R} \to (\lambda^{(1)}, \lambda^{(2)})$ is continuous and strictly increasing and

$$\lim_{u \to -\infty} g(u) = \lambda^{(1)}, \quad \lim_{u \to \infty} g(u) = \lambda^{(2)}. \tag{3}$$

Suppose that $t : \mathbb{R} \to \mathbb{R}$ is such that the variance of $t(y)$ is almost constant when y has distribution F_λ. Let $\rho : \mathbb{R} \to \mathbb{R}$ be a continuous and bounded function with a unique local minimum at 0 and define $m(\lambda)$ by

$$m(\lambda) = \arg\min_u E_\lambda(\rho(t(y) - u)).$$

Let us assume that $m(\lambda)$ is continuous and univocally defined for all λ. Then, given a random sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ of the model (1) and (2), we define the weighted M-estimator based on transformations (WMT-estimator) of $\boldsymbol{\beta}_0$ by

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} L_n(\boldsymbol{\beta}), \tag{4}$$

where

$$L_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho\big(t(y_i) - m(g(\boldsymbol{\beta}'\mathbf{x}_i))\big) w\Big(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n\Big), \tag{5}$$

where $w(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a function of the Mahalanobis distance, that is

$$w(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \omega(((\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))^{1/2}),$$

where $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ are the robust estimators of location and scatter matrix of $\mathbf{x}$ based on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and ω is a non-negative non-increasing function. The purpose of the weighting function $w(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ is to penalize high leverage observations. We will use consistent estimates $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$ so that $\widehat{\boldsymbol{\mu}}_n \to \boldsymbol{\mu}_0$ a.s. and $\widehat{\boldsymbol{\Sigma}}_n \to \boldsymbol{\Sigma}_0$ a.s., where $\boldsymbol{\Sigma}_0$ is positive definite. Note that

$$E_{\beta_0}(\rho(t(y_i) - m(g(\boldsymbol{\beta}'\mathbf{x}_i)))w(\mathbf{x}_n, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)) = E[E_{\beta_0}(\rho(t(y_i) - m(g(\boldsymbol{\beta}'\mathbf{x}_i)))|\mathbf{x})w(\mathbf{x}_n, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0))].$$

Since $E_{\beta_0}(\rho(t(y_i) - m(g(\boldsymbol{\beta}'\mathbf{x}_i)))|\mathbf{x})$ is minimized when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ for all $\mathbf{x}$, then $E_{\beta_0}(\rho(t(y_i) - m(g(\boldsymbol{\beta}'\mathbf{x}_i)))w(\mathbf{x}_i, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0))$ is also minimized when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Therefore WMT-estimators are Fisher consistent. Note that since the variance of $t(y)$ given $\mathbf{x}$ is almost constant, it is not necessary to use a scale in the definition of the M-estimator. Since ρ is bounded, the estimator defined by (4) is robust even if we do not use weights, that is, when $\omega \equiv 1$. However, in cases in which high leverage outliers are expected, the robustness of the estimator may be increased by using a weight function. In some cases the use of these weights may decrease the robustness of the estimator. This occurs when there are good high leverage observations, that is when there are observations where $(\mathbf{x} - \widehat{\boldsymbol{\mu}}_n)\widehat{\boldsymbol{\Sigma}}_n^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is large but the response y is generated according to the nominal GLM. In these cases the weight function would penalize good observations and therefore it may increase the influence of the outliers with low leverage that have larger weights. For this reason we should not discard to take $\omega(u) = 1$. In this case it is called the MT-estimator.

2.2. Some examples of transformations for variance stabilization

Denote by $\boldsymbol{\mu}(\lambda)$ and $\nu(\lambda)$ the mean and variance of F_λ respectively, then a first order Taylor expansion shows that taking

$$t(y) = \int_0^y \frac{du}{[\nu(\boldsymbol{\mu}^{-1}(u))]^{1/2}} \tag{6}$$

we obtain that $\mathrm{var}(t(y))$ is approximately constant. If

$$\nu(\lambda) = \boldsymbol{\mu}(\lambda)^q, \tag{7}$$

then (6) yields

$$t(y) = \begin{cases} y^{-(q/2)+1} & \text{if } q \neq 2, \\ \log(y) & \text{if } q = 2. \end{cases} \tag{8}$$

2.2.1. Poisson regression

In this case F_λ has as support the set $\mathbb{N}$ of non-negative integers and the probability function is $p(x,\lambda) = \exp(-\lambda)\lambda^x/x!$. We also have $\boldsymbol{\mu}(\lambda) = \lambda$ and $\nu(\lambda) = \lambda$ and therefore $t(y) = y^{1/2}$. Fig. 1 shows the plot of $\mathrm{var}(y^{1/2})$ as a function of λ and confirms that this function is quite constant except for small values of λ. We should mention the fact that Poisson regression is often used to model rare events, i.e. $E(X) = \mathrm{var}(X) = \lambda$ is small. However even in this case, the MT and WMT procedures for the Poisson regression studied in the simulations described in Section 5 may have a high degree of robustness. In fact, since these estimators use a function ρ in the family given in (13) with $c=2.4$, observations y with $|y^{1/2} - m(\exp(\boldsymbol{\beta}'\mathbf{x}))| > 2.4$ are completely rejected. Since when $\exp(\boldsymbol{\beta}'\mathbf{x})$ is close to 0, $m(\exp(\boldsymbol{\beta}'\mathbf{x}))$ is close to 0 too, this implies that in this case values of $y > 5$ are completely rejected. Moreover, values of $y=4$ or 5 are significantly downweighted.

2.2.2. Exponential regression

Consider now the case where F_λ has support in the set $\mathbb{R}^+$ of positive real numbers with density

$$p(y,\lambda) = \lambda \exp(-\lambda y)\, I\,(y > 0).$$

We also have $\boldsymbol{\mu}(\lambda) = 1/\lambda$ and $\nu(\lambda) = 1/\lambda^2$ and therefore, according to (7) and (8), $q=2$ and $t(y) = \log(y)$. In this case $\log(y) = w - \log(\lambda)$ where $\exp(w)$ has distribution F_1, and then $\mathrm{var}_\lambda(\log(y))$ is constant. This value is approximately 1.645.

2.2.3. Binomial regression

Assume now that F_λ is a $\mathrm{Bi}(k,\lambda)$ distribution, that is, the probability function is

$$p(y,\lambda) = \binom{k}{y} \lambda^y (1-\lambda)^{k-y}, \quad y = 0, 1, \ldots, k, \quad 0 \leq \lambda \leq 1.$$

Then $\boldsymbol{\mu}(\lambda) = k\lambda$, $\nu(\lambda) = k\lambda(1-\lambda)$, and (6) yields $t(y) = \arcsin(\sqrt{y/k})$. Fig. 2 shows the plot of $\mathrm{var}_\lambda(\arcsin\sqrt{y/k})$ for $k=5$ which is also quite constant.

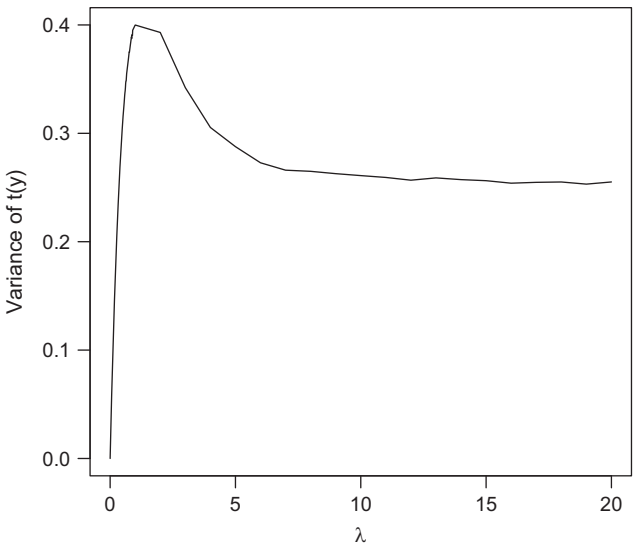

Fig. 1. Variance of $t(y) = \sqrt{y}$ for $y \sim P(\lambda)$.

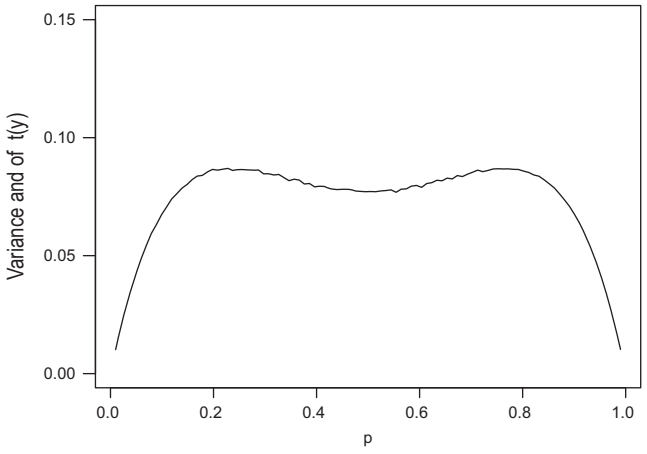

Fig. 2. Variance of $t(y)=\arcsin(\sqrt{y/5})$ for $y\sim Bi(5,\lambda)$.

3. Consistency

In this section we study the consistency of the estimator defined by (4). We need the following assumptions for the consistency of the MT-estimator:

A1 $\sup_\lambda \text{ var }_\lambda(t(y))=A<\infty$.
A2 $m(\lambda)$ is univocally defined for all λ and $\lambda_1<\lambda_2$ implies $m(\lambda_1)<m(\lambda_2)$.
A3 F_λ is continuous in λ.
A4 Suppose that $\lambda_1>\lambda_2$, $X_1\sim F_{\lambda_1}$ and $X_2\sim F_{\lambda_2}$, then X_1 is stochastically larger (or smaller) than X_2.
A5 The function t is strictly increasing and continuous.
B1 $\rho(u)\geq 0$, $\rho(0)=0$ and $\rho(u)=\rho(-u)$.
B2 $\lim_{u\to\infty}\rho(u)=a<\infty$. Without loss of generality we will assume $a=1$.
B3 $0\leq u<v$ implies $\rho(u)\leq\rho(v)$.
B4 $0\leq u<v$ and $\rho(u)<1$ implies $\rho(u)<\rho(v)$.
B5 ρ is continuous.
B6 Let A as in A1, then there exists η such that $\rho(A^{1/2}+\eta)<1$.
B7 There exist $\boldsymbol{\mu}_0\in\mathbb{R}^p$ and a positive definite matrix $\boldsymbol{\Sigma}_0$ such that $\widehat{\boldsymbol{\mu}}_n\to\boldsymbol{\mu}_0$ a.s. and $\widehat{\boldsymbol{\Sigma}}_n\to\boldsymbol{\Sigma}_0$ a.s.
B8 The weight function ω is continuous, bounded and non-increasing and $\sup\omega=1$.
B9 Let $S=\{\mathbf{t}\in\mathbb{R}^p:\|\mathbf{t}\|=1\}$. Then

$$\inf_{\mathbf{t}\in S} P(\{\mathbf{t}'\mathbf{x}\neq\mathbf{0}\}\cap\{\mathbf{w}(\mathbf{x},\boldsymbol{\mu}_0,\Sigma_0)>0\})>0$$

Let

$$m_1=\inf_\lambda m(\lambda)=\lim_{\lambda\to\lambda^{(1)}} m(\lambda), \tag{9}$$

$$m_2=\sup_\lambda m(\lambda)=\lim_{\lambda\to\lambda^{(2)}} m(\lambda), \tag{10}$$

where $\lambda^{(1)}$ and $\lambda^{(2)}$ are defined in (3), and

$$m_3=m(g(0)).$$

Call

$$\Phi_0(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})=w(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\rho(t(y)-m(g(\boldsymbol{\beta}'\mathbf{x}))), \tag{11}$$

then it is immediate that

$$\lim_{\gamma\to\infty}\Phi_0(y,\mathbf{x},\gamma\mathbf{t},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)=\Phi_0^*(y,\mathbf{x},\text{sign}(\mathbf{t}'\mathbf{x})),$$

where

$$\Phi_0^*(y,\mathbf{x},j)=\begin{cases} w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\rho(t(y)-m_1) & \text{if } j=-1,\\ w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\rho(t(y)-m_3) & \text{if } j=0,\\ w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\rho(t(y)-m_2) & \text{if } j=1.\end{cases}$$

We have the following consistency theorem.

Theorem 1. *Let* $(y_1,\mathbf{x}_1),\ldots,(y_n,\mathbf{x}_n)$, *be i.i.d. observations satisfying* (1) *and* (2). *Assume* A1–A5, B1–B9, *let* $\widehat{\boldsymbol{\beta}}_n$ *be the estimator defined by* (4) *and put*

$$\tau=\inf_{\mathbf{t}\in S} E_{\beta_0}(\Phi_0^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x})))-E_{\beta_0}(\Phi_0(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)). \tag{12}$$

Then (i) $\tau>0$ *and* (ii) *Assume also that* $P(\mathbf{t}'\mathbf{x}=0)<\tau$ *for all* $\mathbf{t}\in\mathbb{R}^p$, *then* $\widehat{\boldsymbol{\beta}}_n\to\boldsymbol{\beta}_0$ *a.s.*

Remark 1. Obviously for the MT-estimator B7 and B8 are not necessary and B9 is reduced to $\inf_{\mathbf{t}\in S}P(\mathbf{t}'\mathbf{x}\neq\mathbf{0})>\mathbf{0}$.

4. Asymptotic normality

The following additional assumptions are required to prove the asymptotic normality of the estimator defined by (4):

C1 F_λ has three continuous and bounded derivatives as a function of λ and the link function $g(\lambda)$ is twice continuously differentiable.

C2 ρ has three continuous and bounded derivatives. We write $\psi=\rho'$.

Let $\boldsymbol{\Psi}=(\Psi_1,\ldots\Psi_p):\mathbb{R}^p\times\mathbb{R}\times\mathbb{R}^p\to\mathbb{R}^p$ be defined by

$$\begin{aligned}\boldsymbol{\Psi}_j(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})&=w(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{\partial}{\partial\boldsymbol{\beta}_j}\rho(t(y)-m(g(\boldsymbol{\beta}'\mathbf{x})))\\&=w(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\psi(t(y)-m(g(\boldsymbol{\beta}'\mathbf{x})))m'(g(\boldsymbol{\beta}'\mathbf{x}))g'(\boldsymbol{\beta}'\mathbf{x})x_j.\end{aligned}$$

Denote by $\mathbf{J}_\psi(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})=(J_\psi^{j,k}(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}))_{1\le j,k\le p}$ the Jacobian matrix of $\boldsymbol{\Psi}$ with respect to $\boldsymbol{\beta}$, that is

$$J_\psi^{j,k}(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{\partial}{\partial\boldsymbol{\beta}_k}\boldsymbol{\Psi}_j(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}),\quad 1\le j,\ k\le p.$$

Note that assumptions C1, C2 and Lemma 5 proved in Appendix imply that $\boldsymbol{\Psi}$ and $\mathbf{J}_\psi$ are well defined.

Differentiating $L_n(\boldsymbol{\beta})$ we obtain the following estimating equations for the WMT-estimator of $\boldsymbol{\beta}$:

$$\sum_{i=1}^n\boldsymbol{\Psi}(y_i,\mathbf{x}_i,\boldsymbol{\beta},\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n)=0.$$

C3 $E_\lambda(\psi'(t(y)-m(\lambda)))\neq 0$ for all λ.

C4 There exists $\varepsilon>0$ such that $E_{\beta_0}(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le\varepsilon}|J_\psi^{j,k}(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)|)<\infty$, for all $1\le j,k\le p$, where $\|\ \|$ denotes the l_2 norm, and $E_{\beta_0}(\mathbf{J}_\psi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)))$ is non-singular.

A family of functions satisfying the conditions B1–B6 and C2 is given by

$$\rho_k(u)=\begin{cases}1-\left(1-\left(\dfrac{u}{k}\right)^2\right)^4 & \text{if } |u|\le k\\ 1 & \text{if } |u|\ge k\end{cases} \tag{13}$$

with $k>A^{1/2}$. Note that the functions in the popular bisquare family have a similar expression with the exponent 4 replaced by 3. However functions in the bisquare family have only two derivatives at $c=\pm k$ instead of three as is required by C2.

Observe that C4 is satisfied when the function ω is 0 outside a compact set or when $\mathbf{x}$ takes values in a compact set. Consider the case of Poisson regression, $\omega=1$ and $\psi=0$ outside a compact set. Then, it can be proved that a sufficient condition for C4 is that $E(y\|\mathbf{x}\|^2)=E(e^{\beta_0\mathbf{x}}\|\mathbf{x}\|^2)<\infty$.

The following theorem gives the asymptotic distribution of WMT-estimators.

Theorem 2. *Assume* A1–A5, B1–B9, C1–C4 . *Let* $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$ *be i.i.d. random vectors satisfying* (1) *and* (2) *and let* $\widehat{\boldsymbol{\beta}}_n$ *be defined by* (4). *Then*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\xrightarrow{\mathcal{D}}\mathcal{N}(0,\mathbf{B}^{-1}\mathbf{A}\mathbf{B}'^{-1}),$$

where $\mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *denotes the p-dimensional multivariate normal distribution with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$ *and*

$$\mathbf{A}=E_{\beta_0}(\boldsymbol{\Psi}(y,\mathbf{x},\beta_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\boldsymbol{\Psi}(y,\mathbf{x},\beta_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)'),\quad \mathbf{B}=E_{\beta_0}(\mathbf{J}_\psi(y,\mathbf{x},\beta_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)).$$

To use this result to make asymptotic inference we need to estimate the matrices A and B. To that end, let

$$\widehat{\mathbf{A}}_n=E_{H_n}(\boldsymbol{\Psi}(y,\mathbf{x},\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n)\boldsymbol{\Psi}(y,\mathbf{x},\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n)')$$

and

$$\widehat{\mathbf{B}}_n=E_{H_n}(\mathbf{J}_\psi(y,\mathbf{x},\widehat{\boldsymbol{\beta}}_n,\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n)),$$

where H_n is the empirical distribution of $(y,\mathbf{x})$. Under the assumptions of Theorem 2, by Lemma 6 $\widehat{\mathbf{A}}_n\to\mathbf{A}$ a.s. and $\widehat{\mathbf{B}}_n\to\mathbf{B}$ a.s. Then, the asymptotic covariance matrix $\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{'-1}$ can be consistently estimated by $\widehat{\mathbf{B}}_n^{-1}\widehat{\mathbf{A}}_n\widehat{\mathbf{B}}_n^{'-1}$.

5. Asymptotic breakdown point

The asymptotic breakdown point (ABP) is a measure of robustness of an estimator introduced by Hampel (1971). Roughly speaking, the breakdown point of an estimator is the smallest fraction of atypical points that can take the estimator beyond any limit. More formally, let $(y,\mathbf{x})$ be a random vector in $\mathbb{R}\times\mathbb{R}^p$ with distribution H_0, $\mathcal{D}$ the set of all the distributions on $\mathbb{R}\times\mathbb{R}^p$ and $\mathbf{T}$ a functional defined on $\mathcal{D}$ with values in $\mathbb{R}^p$. Given a sample $(y_1,\mathbf{x}_1),\ldots,(y_n,\mathbf{x}_n)$, consider the estimator defined by $\widehat{\boldsymbol{\beta}}_n=\mathbf{T}(H_n)$, where H_n is the empirical distribution of the sample. Then the asymptotic breakdown point of the functional $\mathbf{T}$ at $H_0\in\mathcal{D}$ is defined by

$$\varepsilon^*(\mathbf{T},H_0)=\sup_{\varepsilon}\left\{\varepsilon\in(0,1):\ \sup_{H^*\in\mathcal{D}}\{\|\mathbf{T}(1-\boldsymbol{\varepsilon})H_0+\boldsymbol{\varepsilon}H^*)\|\}<\infty\right\}.$$

The MT-estimator $\widehat{\boldsymbol{\beta}}_n$ defined in (4) can also be written as $\mathbf{T}(H_n)$ where

$$\mathbf{T}(H)=\arg\min_{\beta\in\mathbb{R}^p}E_H(\rho(t(y)-m(g(\boldsymbol{\beta}'\mathbf{x})))).\tag{14}$$

The following theorem gives a lower bound for the breakdown point of this functional.

Theorem 3. *Let* $(y,\mathbf{x})$ *be a random vector with distribution* H_0 *such that* $P_{H_0}(\boldsymbol{\alpha}'\mathbf{x}=0)=0$ *for all* $\boldsymbol{\alpha}\in\mathbb{R}^p$. *Suppose* $y|\mathbf{x}\sim F_{g(\boldsymbol{\beta}_0'\mathbf{x})}$. *Let*

$$\varepsilon_0=\frac{E_{H_0}(\min(\rho(t(y)-m_1),\rho(t(y)-m_2)))-E_{H_0}(\rho(t(y)-m(g(\boldsymbol{\beta}_0'\mathbf{x}))))}{1+E_{H_0}(\min(\rho(t(y)-m_1),(\rho(t(y)-m_2)))-E_{H_0}(\rho(t(y)-m(g(\boldsymbol{\beta}_0'\mathbf{x}))))},$$

where m_1 *and* m_2 *are defined in* (9) *and* (10) *respectively. Then the ABP of the functional* $\mathbf{T}$ *defined by* (14) *at* H_0 *satisfies*

$$\varepsilon^*(T,H_0)\geq\varepsilon_0.$$

When $m_1=0$ and $m_2=\infty$ as in the Poisson case, we have

$$\min(\rho(t(y)-m_1),\rho(t(y)-m_2))=\rho(t(y))$$

and then

$$\varepsilon_0=\frac{E_{H_0}(\rho(t(y)))-E_{H_0}(\rho(t(y)-m(g(\boldsymbol{\beta}_0'\mathbf{x}))))}{1+E_{H_0}(\rho(t(y)))-E_{H_0}(\rho(t(y)-m(g(\boldsymbol{\beta}_0'\mathbf{x}))))}.$$

In this case ε_0 is small only when the probability that $m(g(\boldsymbol{\beta}_0'\mathbf{x}))$ is close to zero is large. For the Poisson model this happens if $\boldsymbol{\beta}_0'\mathbf{x}$ is negative and has a large absolute value. Note that in this case $P(y=0)$ is large and a small fraction of inliers equal to 0 can make the fraction of observed zeros larger than 0.5. Therefore the good non-null observations may be mistaken as outliers.

If $m_1=-\infty$ and $m_2=\infty$ as in the exponential case then

$$\varepsilon_0=\frac{E_{H_0}(1-\rho(t(y)-m(g(\boldsymbol{\beta}_0'\mathbf{x}))))}{1+E_{H_0}(1-\rho(t(y)-m(g(\boldsymbol{\beta}_0'\mathbf{x}))))}.$$

In the case of exponential regression with $g(u)=\log(u)$, we have $m(\lambda)=m(1)-\log(\lambda)$ and therefore

$$\varepsilon_0=\frac{E(1-\rho(t(y)-m(1)))}{1+E(1-\rho(t(y)-m(1)))},$$

where y is a random variable with distribution $\mathcal{E}(1)$, and therefore ε_0 is independent of $\boldsymbol{\beta}_0$. For example if we use a MT estimator with ρ in the family given in (13) and $k=6$ we have $\varepsilon_0=0.463$.

We will study the breakdown point in a Poisson regression model where $\mathbf{x}=(1,\mathbf{x}^*)$ and $\mathbf{x}^*$ has distribution $\mathcal{N}_{p-1}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ and ρ is in the family (13) with $k=2.4$. Put $\boldsymbol{\beta}_0=(\beta_0,\boldsymbol{\beta}_0^*)$ then the distribution of $\boldsymbol{\beta}_0'\mathbf{x}$ is $\mathcal{N}_1(\mu,\sigma^2)$ where $\mu=\boldsymbol{\beta}_0^{*'}\boldsymbol{\mu}_0+\beta_0$ and

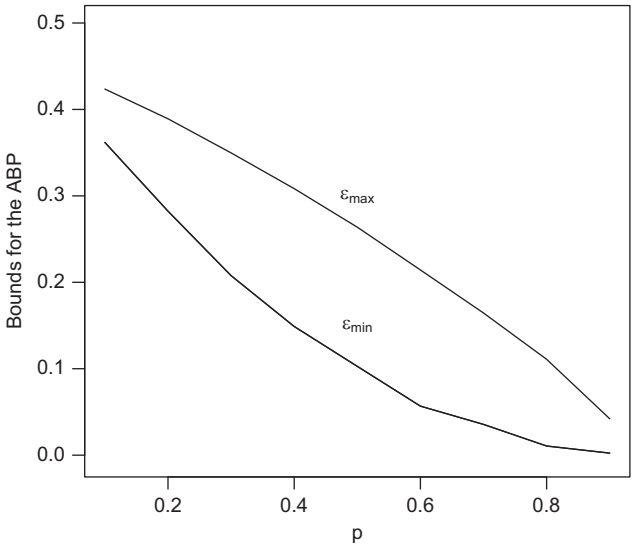

Fig. 3. ε_{min} and ε_{max} as functions of $p = P(y = 0)$.

$\sigma^2 = \boldsymbol{\beta}_0^{*'} \boldsymbol{\Sigma}_0 \boldsymbol{\beta}_0^*$ and therefore ε_0 depends only on $\boldsymbol{\mu}$ and σ^2. The lower bound ε_0 depends on $\boldsymbol{\mu}$ and σ^2 and is highly correlated with $P(y = 0)$. Let $\varepsilon_{\min}(p)$ and $\varepsilon_{\max}(p)$ be defined by

$$\varepsilon_{\min}(p) = \inf_{\boldsymbol{\mu},\sigma^2} \{\varepsilon_0(\boldsymbol{\mu}, \sigma^2) : P_{\boldsymbol{\mu},\sigma^2}(y = 0) = p\}$$

and

$$\varepsilon_{\max}(p) = \sup_{\boldsymbol{\mu},\sigma^2} \{\varepsilon_0(\boldsymbol{\mu}, \sigma^2) : P_{\boldsymbol{\mu},\sigma^2}(y = 0) = p\}.$$

Fig. 3 shows the curves $\varepsilon_{\min}(p)$ and $\varepsilon_{\max}(p)$. These two curves are lower and upper bounds of ε_0 as a function of $P(y = 0)$ when $\mathbf{x}^*$ has a multivariate normal distribution.

6. Monte Carlo study

We performed a Monte Carlo study to compare the behaviors of MT- and WMT-estimators to that of other existing estimators for Poisson regression and log link when the sample size is 100. Let $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $P(\lambda)$ the Poisson distribution with parameter λ. In the Monte Carlo study we took as covariates $\mathbf{x} = (1, \mathbf{x}^*)$ where $\mathbf{x}^*$ has distribution $\mathcal{N}_5(\mathbf{0}, \mathbf{I})$ and the distribution of $y|\mathbf{x}$ is $\mathcal{P}(\exp(\boldsymbol{\beta}_0' \mathbf{x}))$. We considered three different models, with values of $\boldsymbol{\beta}_0 : \boldsymbol{\beta}_{0,1} = (0, 1, 0, 0, 0, 0)$, $\boldsymbol{\beta}_{0,2} = (2, 1, 0, 0, 0, 0)$ and $\boldsymbol{\beta}_{0,3} = (2, 1.5, 0, 0, 0, 0)$. For each of these models we have simulated the case when the samples do not contain outliers and the case when the samples have 10% of identical outliers of the form $(\mathbf{x}_0, y_0)$. We took $\mathbf{x}_0 = (1, 3, 0, 0, 0, 0)$ and y_0 in a grid of values of the form $y_0 = \mu_0 - k\varepsilon_1,\ 1 \le k \le K_1$ $y_0 = \mu_0 + k\varepsilon_2,\ 1 \le k \le K_2$, where $\mu_0 = e^{\boldsymbol{\beta}_0 \mathbf{x}_0} = E_{\boldsymbol{\beta}_0}(y|\mathbf{x} = \mathbf{x}_0)$. The values $\varepsilon_1,\ \varepsilon_2,\ K_1$ and K_2 were chosen so that the grid covers values y_0 close to those yielding the maximum mean squared error. We simulated the following estimators: the maximum likelihood estimator (ML), the robust quasi likelihood estimators proposed by Cantoni and Ronchetti (2001) with no weights (QL) and with weights (WQL), the conditionally unbiased bounded influence estimator (CUBIF) proposed by Künsch et al. (1989), the one step M-estimator starting from the projection estimate (PM) proposed in Bergesio and Yohai (2011) and the MT- and WMT- estimators proposed here. The MT- and the WMT-estimators were computed using a function ρ in the family given in (13). In the case of the WMT-estimators the weight function ω that we use is

$$\omega(t) = \begin{cases} 1 & \text{if } t \le \chi_{0.965,5}, \\ \dfrac{\chi_{0.975,5} - t}{\chi_{0.975,5} - \chi_{0.965,5}} & \text{if } \chi_{0.965,5} < t \le \chi_{0.975,5}, \\ 0 & \text{if } t > \chi_{0.975,5}, \end{cases} \tag{15}$$

where $\chi_{\alpha,p}$ is such that $P(X \le \chi_{\alpha,p}) = \alpha$ where X has a chi-squared distribution with p degrees of freedom. The estimators of location and scatter used to compute the weights are S-estimators with asymptotic breakdown point equal to 0.5 with ρ function in the bisquare family. These S-estimators were computed with the function SestCov, method="bisquare" in the package rrcov of R. It is easy to check that using ρ in the family (13) and ω given by (15), all the assumptions of Theorem 2 are satisfied.

The tuning constants of the estimators were chosen to have an efficiency between 75% and 90%. This was not always possible for the PM-estimators where for some models the efficiency remained low even for large values of the tuning constant. The values of the tuning constants that we use are 0.9 for the QL-estimators, 2.8 for the WQL-estimators, 2.4 for the PM-estimators, 1.6 for the CUBIF-estimator, 2.3 for the MT-estimator and 2.9 for the WMT-estimator.

Given an estimator $\widehat{\boldsymbol{\beta}}$, we denote by MSE, the mean squared error defined by $E_{\boldsymbol{\beta}_0}(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2)$, where $\|\ \|$ denotes the l_2 norm. We estimate the MSE by

$$\widehat{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0\|^2,$$

where $\widehat{\boldsymbol{\beta}}_i$ is the value of the estimator at the i-th replication and N is the number of replications which was chosen equal to 1000. Table 1 gives the efficiencies with respect to the ML-estimator when there are no outliers for the three models. Tables 2–4 and Figs. 4–6 give the values of $\widehat{\mathrm{MSE}}$ for the contaminated samples for the three models. In these figures we do not show the MSE of WQL and WMT which have similar or worse behaviors than QL and MT respectively.

Table 1
Efficiencies without outliers with respect to the ML estimator.

	QL	WQL	CUBIF	PM	MT	WMT
$\beta_{0,1}$	0.88	0.77	0.81	0.83	0.74	0.70
$\beta_{0,2}$	0.88	0.78	0.82	0.80	0.87	0.79
$\beta_{0,3}$	0.88	0.78	0.71	0.45	0.86	0.78

Table 2
$\widehat{MSE}$ under contamination for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,1}$. Between $15 \le y_0 \le 26$, $\widehat{MSE} \le 0.09$, $\boldsymbol{\mu}_0 = e^{\boldsymbol{\beta}_0'\mathbf{x}_0} = E_{\boldsymbol{\beta}_0}(y|\mathbf{x}=\mathbf{x}_0) = 20.06$.

y_0	0	1	2	3	4	5	10	30	35	40	45
ML	0.58	0.47	0.39	0.33	0.28	0.23	0.11	0.09	0.13	0.17	0.22
QL	0.65	0.51	0.41	0.34	0.28	0.24	0.11	0.09	0.13	0.18	0.24
WQL	1.20	0.83	0.61	0.48	0.39	0.32	0.15	0.12	0.16	0.22	0.28
CUBIF	0.81	0.60	0.46	0.37	0.30	0.25	0.11	0.09	0.12	0.15	0.19
PM	0.37	0.35	0.34	0.29	0.26	0.23	0.11	0.10	0.12	0.15	0.18
MT	0.13	0.46	0.43	0.35	0.29	0.24	0.12	0.11	0.15	0.21	0.27
WMT	0.26	0.60	0.45	0.36	0.30	0.25	0.11	0.11	0.16	0.21	0.27

Table 3
$\widehat{MSE}$ under contamination for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,2}$. Between $80 \le y_0 \le 240$, $\widehat{MSE} \le 0.09$, $\boldsymbol{\mu}_0 = e^{\boldsymbol{\beta}_0'\mathbf{x}_0} = E_{\boldsymbol{\beta}_0}(y|\mathbf{x}=\mathbf{x}_0) = 148.41$.

y_0	0	10	20	30	40	50	60	70	280	320	360	400
ML	0.53	0.40	0.31	0.24	0.18	0.14	0.11	0.08	0.10	0.15	0.21	0.27
QL	0.55	0.47	0.34	0.25	0.19	0.14	0.11	0.08	0.10	0.15	0.20	0.25
WQL	0.80	0.70	0.46	0.32	0.23	0.17	0.13	0.10	0.10	0.15	0.19	0.24
CUBIF	0.26	0.30	0.30	0.25	0.18	0.14	0.10	0.07	0.05	0.05	0.05	0.05
PM	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.03	0.03	0.03	0.03
MT	0.01	0.01	0.01	0.02	0.04	0.08	0.09	0.08	0.09	0.11	0.12	0.10
WMT	0.01	0.010	0.01	0.05	0.10	0.12	0.11	0.09	0.09	0.13	0.16	0.17

Table 4
$\widehat{MSE}$ under contamination for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,3}$. Between $400 \le y_0 \le 1000$, $\widehat{MSE} \le 0.13$, $\boldsymbol{\mu}_0 = e_0^{\boldsymbol{\beta}'}\mathbf{x_0} = E_{\boldsymbol{\beta}_0}(y|\mathbf{x}=\mathbf{x}_0) = 665.134$.

y_0	0	50	100	150	200	1200	1400	1600	1800
ML	1.34	0.90	0.64	0.47	0.34	0.17	0.29	0.42	0.57
QL	1.30	0.83	0.55	0.40	0.29	0.15	0.24	0.33	0.43
WQL	1.90	0.90	0.55	0.38	0.27	0.13	0.20	0.27	0.35
CUBIF	0.05	0.05	0.05	0.05	0.03	0.04	0.05	0.05	0.05
PM	0.15	0.14	0.15	0.16	0.11	0.11	0.11	0.11	0.10
MT	0.01	0.01	0.01	0.01	0.06	0.11	0.13	0.12	0.10
WMT	0.01	0.01	0.01	0.04	0.06	0.11	0.16	0.18	0.18

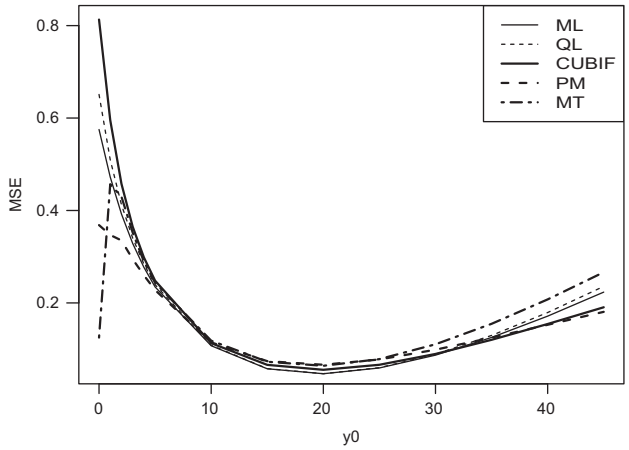

Fig. 4. $\widehat{MSE}$ under contamination for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,1}$, $\mu_0 = 20.06$.

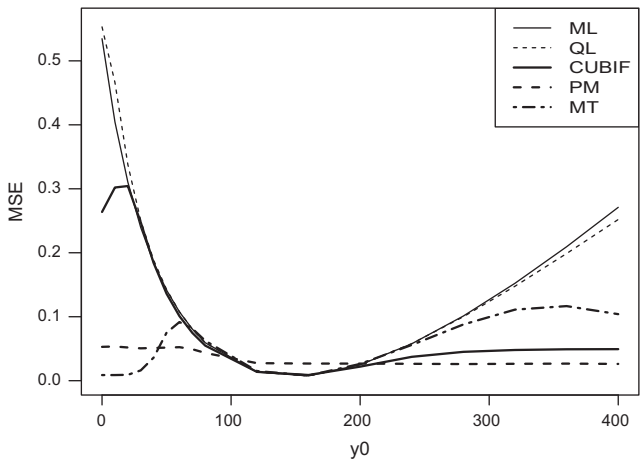

Fig. 5. $\widehat{MSE}$ under contamination for $\boldsymbol{\beta} = \boldsymbol{\beta}_{0,2}$, $\mu_0 = 148.41$.

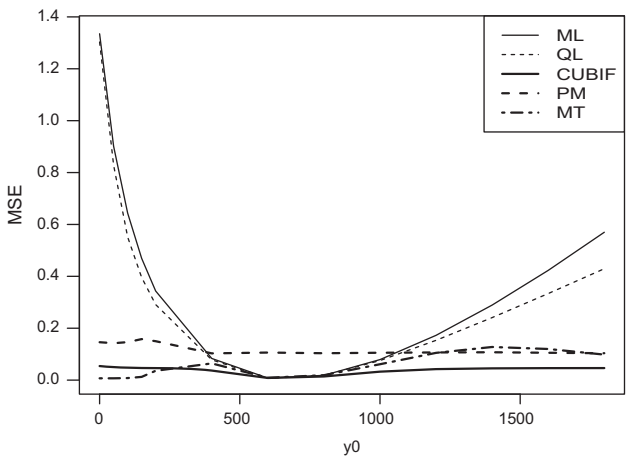

Fig. 6. $\widehat{MSE}$ under contamination for $\boldsymbol{\beta} = \boldsymbol{\beta}_{0,3}$, $\mu_0 = 665.134$.

We note that except for the PM, all the simulated estimators have a similar efficiency for the three considered models. The PM has a lower efficiency, especially when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,3}$. Table 2 and Fig. 4 show that when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,1}$ the most robust estimator is PM followed by MT. Instead when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,2}$, according to Table 3 and Fig. 5 the most robust estimators are PM

Table 5
Computing times (t) and maximum supported outlier fraction (ε) for different values of N and p.

p	$N=500$		$N=1000$		$N=1500$	
	t	ε	t	ε	t	ε
10	0:51	0.37	1:29	0.42	2:06	0.44
20	1:21	0.21	2:02	0.24	2:35	0.25
30	1:42	0.15	2:29	0.16	3:10	0.17
40	2:14	0.11	2:59	0.13	3:55	0.14
50	2:42	0.09	3:46	0.10	4:32	0.11

and MT, which have a similar behavior. Finally, when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,3}$, from Table 4 and Fig. 6 we infer that PM, MT and CUBIF are the most robust estimators. For the three considered values of $\boldsymbol{\beta}_0$, the MT estimator has a good behavior without and with outliers. The PM-estimator for $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{0,1}$ has a slightly better performance under outlier contamination, but its efficiency for clean data may be much lower.

6.1. Computational method

To obtain the function $m(\lambda)$ we note that if y has distribution $P(\lambda)$ then $y^{1/2} - \lambda^{1/2} \to_d \mathrm{N}(0, 1/4)$ when $\lambda \to \infty$. This implies that for large λ we can approximate $m(\lambda)$ by $\lambda^{1/2}$. Then we proceed as follows: for $0 < \lambda \leq 3$ we fit a cubic spline using a grid with step 0.1. The value of $m(\lambda)$ for each element of the grid was computed using the function "optimize" in R. For $\lambda > 3$, since $m(\lambda)$ is very close to $\lambda^{1/2}$, a good approximation to $m(\lambda)$ is obtained using one step Newton–Raphson starting at $\lambda^{1/2}$. For the computation of the MT-estimator defined in (4) we used the quasi-Newton optimization method BFGS provided in the function `optim` of R. As our objective function may have multiple local minima, a crucial part of the computation is the choice of initial values for the parameters. This initial value was obtained using a subsampling procedure plus a concentration step similar to the one proposed by Rousseeuw and Van Zomeren (1990). More precisely, we choose at random N subsamples of size p. For each subsample a candidate for the initial estimator is obtained by computing the maximum likelihood estimator. This candidate is improved by computing the maximum likelihood estimator of the $[n/2]$ observations with the smallest deviances. Finally we compute the loss function given in (5) to all the improved candidates and choose as an initial estimator the one that attains the minimum value. The number N of subsamples may be determined as in the case of linear regression, see for example Maronna et al. (2006, Chapter 5). This value depends on the number of regressors p, on the expected fraction of outliers ε and on the degree of the desired probability of obtaining at least one sample free of outliers γ.

Table 5 gives computing mean times expressed in minutes and seconds for MT estimators in a PC with an AMD Athlon II X3 450 Processor with a speed of 3.20 GHz and 8 GB of RAM memory for N equal 500,1000 and 1500 and different values of p. In the same table we show the corresponding values of ε when $\gamma = 0.99$. These values were obtained with a program written in R.

We observe that for $p=50$, if the fraction of expected outliers is not larger than 0.1, the estimator may be computed in a reasonable amount of time. Note that these times may be substantially decreased with a more powerful machine and/or with a code partially written in a lower level access to memory language, as for example C.

However the values of N required when p and ε increase may be very large and the algorithm becomes unfeasible. For these cases the minimum of L_n may be obtained using heuristic optimization methods, but this is a matter of further research.

7. Example: Epilepsy data

Breslow (1996) used a GLM with Poisson response and log link to study the effect of drugs in epilepsy patients. He considered data from a clinical trial of 59 patients with epilepsy, 31 of whom were randomized to receive the anti-epilepsy drug Progabide and 28 of whom received a placebo. The response variable is *SumY*: the number of attacks during four weeks in a given time interval. The explanatory variables are *Age* 10: patient age divided by ten, *Base4*: number of attacks in the four weeks prior to the study, *Trt*: a dummy variable that takes the values 1 or 0 if the patient received the drug or a placebo respectively and *Base4*∗*Trt*: to take into account the interaction between these two variables.

We fit the Poisson GLM with log link using the same estimators as in the simulation study except for WQL and WMT. Fig. 7 shows boxplots of the absolute values of the deviance residuals. In the left plot we consider all the residuals. To make the boxes and whiskers more clearly visible, in the right plot we eliminated the outliers detected by the boxplots for each fit. It is clear from the boxplots that the MT-estimator gives the best fit for the bulk of the data.

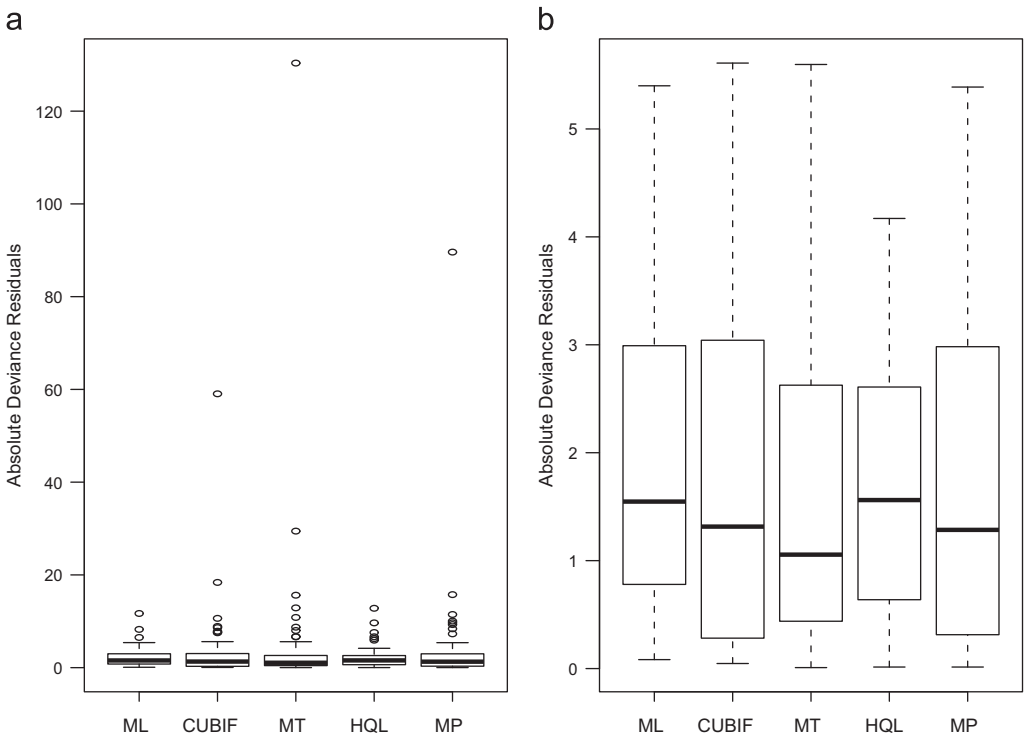

Fig. 7. Boxplots of the absolute values of the deviance residuals: (a) all the observations and (b) without the outliers.

8. Conclusions

We have presented two new families of estimators for GLM which combine M-estimation with a variance stabilizing response transformation: the WMT- and MT-estimators. We performed a Monte Carlo study to compare the proposed estimators with other robust estimators in the case of a GLM with Poisson response and log link. This Monte Carlo study shows that the MT-estimator compares favorably with the other robust estimators when efficiency and robustness are both considered. This study also shows that for the three simulated models, the WMT-estimator has a similar or worse behavior than that of the MT-estimator. Hence, penalizing high leverage observations does not improve the robustness of the estimator. We have also obtained the asymptotic normal distribution of the WMT- and MT-estimators. This distribution can be used for testing hypotheses about the vector of regression coefficients, $\boldsymbol{\beta}_0$, as well as to obtain confidence regions for $\boldsymbol{\beta}_0$ or some of its components.

Acknowledgments

This research was partially supported by Grants W276 from Universidad of Buenos Aires, PIP 112-2008-01-00216 and 112-2011-01-00339 from CONICET and PICT 2011-0397 from ANPCYT, Argentina. We also thank two anonymous referees for their comments and suggestions which contributed to a substantial improvement of the paper.

Appendix A

A.1. A general consistency theorem

We will prove Theorem 1 as a particular case of a more general consistency result.

Let $(y, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be a sample with $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$ and let $\Phi : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$. Consider the estimator $\widehat{\boldsymbol{\beta}}_n$ defined by

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \Phi(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}_n), \tag{16}$$

where $\boldsymbol{\theta}$ is a nuisance parameter with values in $\mathbb{R}^q$ and $\widehat{\boldsymbol{\theta}}_n$ is a sequence of estimators of $\boldsymbol{\theta}$.

We consider the following assumptions:

P0 There exists $\boldsymbol{\theta}_0$ such that $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ a.s.

P1 The function Φ is continuous and bounded and there exists a function $\vartheta(\mathbf{x},\boldsymbol{\theta}) : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ and a constant C such that

$$|\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\theta_2) - \Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\theta_1)| \le C|\vartheta(\mathbf{x},\boldsymbol{\theta}_2) - \vartheta(\mathbf{x},\boldsymbol{\theta}_1)|$$

for all $y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Besides, if $\boldsymbol{\theta}_0$ is as in P0, then $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$ implies $\sup_{\mathbf{x}}|\vartheta(\mathbf{x},\boldsymbol{\theta}_n) - \vartheta(\mathbf{x},\boldsymbol{\theta}_0)| \to 0$.

P2 Let $\boldsymbol{\theta}_0$ be as in P0. Then, there exists $\boldsymbol{\beta}_0$ such that

$$E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)) < E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta},\theta_0)) \tag{17}$$

for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

P3 Let S be as in B9, $\boldsymbol{\theta}_0$ as in P0 and $\boldsymbol{\beta}_0$ as in P2. Then there exists a function $\Phi^*(y,\mathbf{x},j)$, $j = -1,0,1$ such that for all $\mathbf{t} \in S$ we have

$$\lim_{\gamma\to\infty} \Phi(y,\mathbf{x},\gamma\mathbf{t},\boldsymbol{\theta}_0) = \Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x}))$$

and if $\mathbf{t}'\mathbf{x} \neq 0$ there exists a neighborhood of $\mathbf{t}$ where this convergence is uniform. Besides

$$\tau = \inf_{\mathbf{t}\in\mathbf{S}}[E_{\beta_0}(\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x})) - E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)] > \mathbf{0}. \tag{18}$$

Theorem 4. *Let $(y_i,\mathbf{x}_i), i \in N$, be a sequence of i.i.d. random vectors and assume P0–P3 and $P(\mathbf{t}'\mathbf{x}=0) < \tau/M$ for all $\mathbf{t} \in S$, where $M = \sup_{y,\mathbf{x},\,\boldsymbol{\beta}}\Phi(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\theta}_0)$. Let $\widehat{\boldsymbol{\beta}}_n$ be defined by (4), then $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ a.s.*

Given $\mathbf{t} \in \mathbb{R}^p$ and $\varepsilon > 0$, let $B(\mathbf{t},\varepsilon) = \{\mathbf{s} \in \mathbb{R}^p : \|\mathbf{s}-\mathbf{t}\| \le \varepsilon\}$. The following lemma is required to prove Theorem 4.

Lemma 1. *Let $(y_1,\mathbf{x}_1),\ldots,(y_n,\mathbf{x}_n)$ be i.i.d. random vectors and $\widehat{\boldsymbol{\beta}}_n$ be defined by (16) and assume P0, P1, P3 and $P(\mathbf{t}'x=0) < \tau/M$ for all $\mathbf{t} \in S$. Then there exists a compact set $C \subset \mathbb{R}^p$ and $n_0 \in N$ such that, if $n \ge n_0$ then $\widehat{\boldsymbol{\beta}}_n \in C$ almost surely.*

Proof. It is enough to prove that there exists K_0 such that

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\widehat{\boldsymbol{\theta}}_n\right) < \overline{\lim_{n\to\infty}}\inf_{\|\boldsymbol{\beta}\|>K_0}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta},\widehat{\boldsymbol{\theta}}_n\right) \text{ a.s.} \tag{19}$$

By P0 and P1

$$\left|\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\widehat{\boldsymbol{\theta}}_n\right) - \frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0\right)\right| \le C\sup_{\mathbf{x}}|\vartheta\left(\mathbf{x},\widehat{\boldsymbol{\theta}}_n\right) - \vartheta(\mathbf{x},\boldsymbol{\theta}_0)| \to 0 \text{ a.s.} \tag{20}$$

Then it is enough to show that there exists K_0 such that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0\right) < \overline{\lim_{n\to\infty}}\inf_{\|\boldsymbol{\beta}\|>K_0}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_0\right) \text{ a.s.} \tag{21}$$

To prove (21) it suffices to show that there exists $K_0 > 0$ such that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0\right) < \overline{\lim_{n\to\infty}}\inf_{\gamma>K_0}\inf_{\mathbf{s}\in S}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\gamma\mathbf{s},\boldsymbol{\theta}_0\right) \text{ a.s.} \tag{22}$$

We start proving that for all $\mathbf{t} \in S$ there exists $\varepsilon > 0$ such that

$$E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)) < E_{\beta_0}\left(\overline{\lim_{\gamma\to\infty}}\inf_{\mathbf{s}\in B(\mathbf{t},\varepsilon)}\Phi(y,\mathbf{x},\gamma\mathbf{s},\boldsymbol{\theta}_0)\right). \tag{23}$$

Since $P(\mathbf{X}'\mathbf{s}=0) < \tau/M$ for all $\mathbf{s} \in \mathbb{R}^p$, given $\mathbf{t} \in S$ it is easy to show that there exist positive numbers ς and K such that if

$$C_{\mathbf{t}} = \{\mathbf{x} : |\mathbf{t}'\mathbf{x}| > \varsigma, \|\mathbf{x}\| < K\}$$

then

$$P(C_{\mathbf{t}}) > 1 - \tau/M$$

Let $\varepsilon = \varsigma/(2K)$ and $\mathbf{x} \in C_{\mathbf{t}}$. Then for all $\mathbf{s} \in B(t,\varepsilon)$ we have $|\mathbf{s}'\mathbf{x} - \mathbf{t}'\mathbf{x}| \le \varsigma/2$ and therefore $|\mathbf{s}'\mathbf{x}| \ge \varsigma/2$ and $\operatorname{sign}(\mathbf{s}'\mathbf{x}) = \operatorname{sign}(\mathbf{t}'\mathbf{x})$. Given any $\mathbf{x}$ and y, by P3 we have

$$\overline{\lim_{\gamma\to\infty}}\inf_{\mathbf{s}\in B(\mathbf{t},\varepsilon)}\Phi(y,\mathbf{x},\gamma\mathbf{s},\boldsymbol{\theta}_0) \le \lim_{\gamma\to\infty}\Phi(y,\mathbf{x},\gamma\mathbf{t},\boldsymbol{\theta}_0) = \Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x})).$$

Let us suppose that the strict inequality holds for some point $\mathbf{x} \in C_{\mathbf{t}}$ and $y \in \mathbb{R}$, that is

$$\overline{\lim_{\gamma\to\infty}}\inf_{\mathbf{s}\in B(\mathbf{t},\varepsilon)}\Phi(y,\mathbf{x},\gamma\mathbf{s},\boldsymbol{\theta}_0) < \Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x})),$$

then, there exist $\zeta>0$, a sequence of positive numbers $\gamma_n\to\infty$ and $\mathbf{s}_n\in B(\mathbf{t},\varepsilon)$ such that

$$\Phi(y,\mathbf{x},\gamma_n\mathbf{s}_n,\boldsymbol{\theta}_0)<\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x}))-\zeta \tag{24}$$

and then

$$\lim_{n\to\infty}\Phi(y,\mathbf{x},\gamma_n\mathbf{s}_n,\theta_0)\le\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x}))-\zeta. \tag{25}$$

We can assume that $\mathbf{s}_n\to\mathbf{s}_0$, where $\mathbf{s}_0\in B(\mathbf{t},\varepsilon)$ and $\mathbf{s}_0'\mathbf{x}\neq 0$. Moreover since for n sufficiently large the sign of $\mathbf{s}_n'\mathbf{x}$ is the same as the sign of $\mathbf{s}_0'\mathbf{x}$ and of the $\mathbf{t}'\mathbf{x}$, by P3 we have

$$\lim_{n\to\infty}\Phi(y,\mathbf{x},\gamma_n\mathbf{s}_n,\boldsymbol{\theta}_0)=\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{s}_0'\mathbf{x}))=\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x})),$$

contradicting (25). Then

$$\varliminf_{\gamma\to\infty}\inf_{\mathbf{s}\in B(\mathbf{t},\varepsilon)}\Phi(y,\mathbf{x},\gamma s,\boldsymbol{\theta}_0)=\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x})).$$

Given a set A, we denote by A^c its complement. Then using that $P(C_{\mathbf{t}}^c)<\tau/M$, $\sup\Phi^*\le M$ and (18) we get

$$\begin{aligned}E_{\beta_0}\left(\varliminf_{\gamma\to\infty}\inf_{\mathbf{s}\in B(\mathbf{t},\varepsilon)}\Phi(y,\mathbf{x},\gamma\mathbf{s},\boldsymbol{\theta}_0)\right)&\ge E_{\beta_0}\left(\varliminf_{\gamma\to\infty}\inf_{\mathbf{s}\in B(\mathbf{t},\varepsilon)}\Phi(y,\mathbf{x},\gamma\mathbf{s},\boldsymbol{\theta}_0)I_{C_{\mathbf{t}}}(\mathbf{x})\right)\\&=E_{\beta_0}(\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x}))I_{C_{\mathbf{t}}}(\mathbf{x}))\\&\ge E_{\beta_0}(\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x}))))\\&\quad-E_{\beta_0}(\Phi^*(y,\mathbf{x},\operatorname{sign}(\mathbf{t}'\mathbf{x})))I_{C_{\mathbf{t}}^c}(\mathbf{x}))\\&>E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0))\end{aligned}$$

proving (23).

Finally we show how to derive (22) from (23) using the Law of Large Numbers and the compactness of S. By (23) and the Dominated Convergence Theorem, for all $\mathbf{t}\in S$ there exist $\zeta_{\mathbf{t}}$, $\varepsilon_{\mathbf{t}}$ and $K_{\mathbf{t}}$ such that

$$E_{\beta_0}\left(\inf_{\gamma>K_{\mathbf{t}}}\inf_{\mathbf{s}\in B(\mathbf{t},\varepsilon_{\mathbf{t}})}\Phi(y,\mathbf{x},\gamma\mathbf{s},\boldsymbol{\theta}_0)\right)>E_{\beta_0}\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)+\zeta_{\mathbf{t}}. \tag{26}$$

Since S is compact, there exists a finite set $\mathbf{t}_1,\mathbf{t}_2,\ldots,\mathbf{t}_j$ of elements of S such that $S\subset\bigcup_{h=1}^{j}B(\mathbf{t}_h,\varepsilon_{\mathbf{t}_h})$. Let $K_0=\max\{K_{\mathbf{t}_1},\ldots,K_{\mathbf{t}_j}\}$ and $\zeta_0=\min\{\zeta_{\mathbf{t}_1},\ldots,\zeta_{\mathbf{t}_j}\}$. Then

$$\inf_{\gamma>K_0}\inf_{\mathbf{s}\in S}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\gamma\mathbf{s},\boldsymbol{\theta}_0)\ge\inf_{1\le h\le j}\inf_{\gamma>K_h}\inf_{\mathbf{s}\in B(\mathbf{t}_h,\varepsilon_{\mathbf{t}_h})}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\gamma\mathbf{s},\boldsymbol{\theta}_0),$$

and by the Law of Large Numbers and (26) we get

$$\varliminf_{n\to\infty}\inf_{\gamma>K_0}\inf_{\mathbf{s}\in S}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\gamma\mathbf{s},\boldsymbol{\theta}_0)\ge\inf_{1\le h\le j}\varliminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\inf_{\gamma>K_h}\inf_{\mathbf{s}\in B(\mathbf{t}_h,\varepsilon_{\mathbf{t}_h})}\Phi(y_i,\mathbf{x}_i,\gamma\mathbf{s},\boldsymbol{\theta}_0) \tag{27}$$

$$=\inf_{1\le h\le j}E\left(\inf_{\gamma>K_h}\inf_{\mathbf{s}\in B(\mathbf{t}_h,\varepsilon_{\mathbf{t}_h})}\Phi(y_i,\mathbf{x}_i,\gamma\mathbf{s},\boldsymbol{\theta}_0)\right)\ge E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0))+\zeta_0\ \text{a.s.} \tag{28}$$

Since

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)=E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\theta_0)), \tag{29}$$

(22) follows from (28). □

Proof of Theorem 4. Let C be the compact set given by Lemma 1. Then, according to this lemma, it is enough to prove that for all U open neighborhood of $\boldsymbol{\beta}_0$ we have

$$\varliminf_{n\to\infty}\inf_{\beta\in C-U}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta},\widehat{\boldsymbol{\theta}}_n\right)>\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Phi\left(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\widehat{\boldsymbol{\theta}}_n\right)\ \text{a.s.}, \tag{30}$$

and by (20) it is enough to show

$$\varliminf_{n\to\infty}\inf_{\beta\in C-U}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_0)>\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)\ \text{a.s.} \tag{31}$$

Using the Dominated Convergence Theorem and a standard compactness argument we can find $\mathbf{t}_j\in\mathbb{R}^p$, $\varepsilon_j>0$, $\zeta_j>0$, $1\le j\le h$ such that

$$\bigcup_{j=1}^{h}B(\mathbf{t}_j,\varepsilon_j)\supset C-U$$

and

$$E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)) \le E_{\beta_0}\left(\inf_{\boldsymbol{\beta}\in B(\mathbf{t}_j,\varepsilon_j)} \Phi(y,\mathbf{x},\boldsymbol{\beta},\theta_0)\right)-\zeta_j, \quad 1\le j\le h. \tag{32}$$

We also have

$$\inf_{\boldsymbol{\beta}\in C-U}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_0) \ge \inf_{1\le j\le h}\inf_{\boldsymbol{\beta}\in B(\mathbf{t}_j,\varepsilon_j)}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_0).$$

Then, putting $\zeta_0=\min_{1\le j\le h}\zeta_j$, by the Law of Large Numbers and (32) we get

$$\begin{aligned}\varliminf_{n\to\infty}\inf_{\boldsymbol{\beta}\in C-U}\frac{1}{n}\sum_{i=1}^{n}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_0) &\ge \inf_{1\le j\le h}\varliminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\inf_{\boldsymbol{\beta}\in B(\mathbf{t}_j,\varepsilon_j)}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_0)\\ &\ge \inf_{1\le j\le h}E_{\beta_0}\left(\inf_{\boldsymbol{\beta}\in B(\mathbf{t}_j,\varepsilon_j)}\Phi(y_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\theta}_0)\right)\\ &\ge E_{\beta_0}(\Phi(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\theta}_0))+\zeta_0.\end{aligned}$$

and then from (29) we get (31). □

A.2. Proof of the consistency of WMT-estimators

In this section we prove Theorem 1. We need the following lemmas.

Lemma 2. *Assume A1–A2 and B1–B6. Then, there exists ε_0 such that*

$$E_\lambda(\rho(t(y)-m(\lambda)))<1-\varepsilon_0 \tag{33}$$

for all λ.

Proof. By Chebyshev's inequality we have that for any $\delta>0$

$$\begin{aligned}E_\lambda(\rho(t(y)-m(\lambda))) &\le E_\lambda(\rho(t(y)-E_\lambda(t(y))))\\ &\le \rho(\delta)P(|t(y)-E_\lambda(t(y))|<\delta)+P(|t(y)-E_\lambda(t(y))|\ge\delta)\\ &= 1-(1-\rho(\delta))P(|t(y)-E_\lambda(t(y))|<\delta)\\ &\le 1-(1-\rho(\delta))\left(1-\frac{A}{\delta^2}\right).\end{aligned}$$

Then taking $\delta=A^{1/2}+\eta$ we get

$$E_\lambda(\rho(t(y)-m(\lambda)))\le 1-\varepsilon_0$$

with

$$\varepsilon_0=\left(1-\rho\left(A^{1/2}+\eta\right)\right)\left(1-\frac{A}{(A^{1/2}+\eta)^2}\right).$$

This proves the lemma. □

Lemma 3. *Assume B1–B6. Then A1–A5 imply that $m(\lambda)$ is continuous.*

Proof. Take a sequence $\lambda_i\to\lambda^{(0)}$, we have to prove that $m(\lambda_i)\to m(\lambda^{(0)})$. Suppose this is not true, then passing to a subsequence if necessary we can assume that $m(\lambda_i)$ converges to a value m_0 possibly $\mp\infty$. Suppose that $m_0=+\infty$ and take $\lambda^{(1)}<\lambda^{(0)}$. Then by the Dominated Convergence Theorem we have $E_{\lambda^{(1)}}(\rho(t(y)-m(\lambda_i))\to 1$. Then using A4 and A5 we get that $E_{\lambda_i}(\rho(t(y)-m(\lambda_i))\to 1$ contradicting the fact that by Lemma 2, $E_\lambda(\rho(t(y)-m(\lambda))<1-\varepsilon_0$ for all λ. Similarly we can prove that m_0 cannot be $-\infty$.

We consider now the case of finite m_0. By the definition of $m(\lambda)$ we have

$$E_{\lambda_i}(\rho(t(y)-m(\lambda_i)))\le E_{\lambda_i}(\rho(t(y)-m(\lambda^{(0)})))$$

for all $i\in N$. Taking limits we get

$$E_{\lambda^{(0)}}(\rho(t(y)-m_0)))\le E_{\lambda^{(0)}}(\rho(t(y)-m(\lambda^{(0)}))).$$

Therefore by the uniqueness of $m(\lambda^{(0)})$ we obtain that $m(\lambda^{(0)})=m_0$ proving the continuity of $m(\lambda)$. □

Lemma 4. *Let $\boldsymbol{\mu}_0\in\mathbb{R}^p$ and $\boldsymbol{\Sigma}_0$ a $p\times p$ positive definite matrix. Assume B8, then $\boldsymbol{\mu}_n\to\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_n\to\boldsymbol{\Sigma}_0$ imply*

$$\sup_{\mathbf{x}}|w(\mathbf{x},\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)-w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)|\to 0. \tag{34}$$

Proof. Suppose that (34) is not true. Then there exists a sequence $\mathbf{x}_n$ and $\varepsilon>0$ such that for all n

$$|w(\mathbf{x}_n,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)-w(\mathbf{x}_n,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)|>\varepsilon. \tag{35}$$

We can assume without loss of generality that $\mathbf{x}_n \to \mathbf{x}_0$ or $\|\mathbf{x}_n\| \to \infty$. In the first case

$$\lim_{n\to\infty}(\mathbf{x}_n-\boldsymbol{\mu}_n)'\boldsymbol{\Sigma}_n^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_n)=\lim_{n\to\infty}(\mathbf{x}_n-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_0)$$
$$=(\mathbf{x}_0-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}_0-\boldsymbol{\mu}_0).$$

Therefore, since by B8 ω is continuous we get

$$\lim_{n\to\infty} w(\mathbf{x}_n,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)=\lim_{n\to\infty} w(\mathbf{x}_n,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)=w(\mathbf{x}_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0),$$

contradicting (35). In the second case we have that $\lim_{n\to\infty}(\mathbf{x}_n-\boldsymbol{\mu}_n)'\boldsymbol{\Sigma}_n^{-1}(x_n-\boldsymbol{\mu}_n)=\lim_{n\to\infty}(\mathbf{x}_n-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_0)=\infty$. B8 implies that there exists a such that $\lim_{d\to\infty}\omega(d)=a$ and therefore $\lim_{n\to\infty}w(\mathbf{x}_n,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)=\lim_{n\to\infty}w(\mathbf{x}_n,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)=a$. This contradicts (35) too. □

Proof of Theorem 1. Let Φ_0 be defined by (11). According to Theorem 4, it is enough to show that P0–P3 hold when $\boldsymbol{\theta}=(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\widehat{\boldsymbol{\theta}}_n=(\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n)$, $\Phi=\Phi_0$ and $\vartheta(\mathbf{x},\boldsymbol{\theta})=w(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})$.

P1 follows from Lemma 4. Put

$$D(\lambda,u)=E_\lambda(\rho(t(y)-u))$$

Take $\boldsymbol{\beta}\neq\boldsymbol{\beta}_0$. Then by A2 we have

$$D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}'\mathbf{x})))-D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}_0'\mathbf{x})))\geq\mathbf{0} \tag{36}$$

and

$$D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}'\mathbf{x})))-D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}_0'\mathbf{x})))>\mathbf{0}\quad\text{if }\boldsymbol{\beta}_0'\mathbf{x}\neq\boldsymbol{\beta}'\mathbf{x}. \tag{37}$$

Moreover

$$E_{\beta_0}(\Phi_0(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)|\mathbf{x})=D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}'\mathbf{x}))w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0). \tag{38}$$

Let $V=\{\mathbf{x}:(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{x}\neq\mathbf{0}\}\cap\{\mathbf{x}:w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)>0\}$, then by B9 $P(V)>0$. From (36) and (38) we get that

$$E_{\beta_0}(\Phi_0(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0))-E_{\beta_0}(\Phi_0(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0))$$
$$\geq E\{[D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}'\mathbf{x})))-D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g\boldsymbol{\beta}_0'\mathbf{x})))]w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)I(V)\}.$$

Since for $\mathbf{x}\in V$ we have

$$[D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}'\mathbf{x})))-D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}_0'\mathbf{x})))]w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)>0$$

we obtain that $E_{\beta_0}(\Phi_0(y,\mathbf{x},\boldsymbol{\beta},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0))-E_{\beta_0}(\Phi_0(y,\mathbf{x},\boldsymbol{\beta}_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0))>0$ and therefore P2 is satisfied.

To prove P3 it is enough to show that $\tau>0$, where τ is defined in (12). By B9

$$\delta=\inf_{\mathbf{t}\in\mathbf{S}}P(\{\mathbf{t}'\mathbf{x}\neq\mathbf{0}\}\cap\{\mathbf{w}(\mathbf{x},\boldsymbol{\mu}_0,\Sigma_0)>0\})>0. \tag{39}$$

We are going to show that there exists $\zeta>0$ such that

$$\inf_{\mathbf{t}\in\mathbf{S}}P(\{\mathbf{t}'\mathbf{x}\neq\mathbf{0}\}\cap\{\omega(\mathbf{x},\boldsymbol{\mu}_0,\Sigma_0)>\zeta\})\geq\delta/2. \tag{40}$$

Suppose that this is not true. Then there exist sequences $\mathbf{t}_n\in S$ and $\zeta_n\to0$ such that

$$P(\{\mathbf{t}_n'\mathbf{x}\neq\mathbf{0}\}\cap\{\omega(\mathbf{x},\boldsymbol{\mu}_0,\Sigma_0)>\zeta_n\})<\delta/2,$$

and without loss of generality we can assume that $\mathbf{t}_n\to\mathbf{t}_0\in S$. Then we get that

$$P(\{\mathbf{t}_0'\mathbf{x}\neq\mathbf{0}\}\cap\{\omega(\mathbf{x},\boldsymbol{\mu}_0,\Sigma_0)>0\})\leq\delta/2,$$

contradicting (39). We can also find K_1 and K_2 such that

$$P(\boldsymbol{\beta}_0'\mathbf{x}\in[K_1,K_2])>1-\delta/4. \tag{41}$$

Then, if we put

$$V_{\mathbf{t}}=\{\mathbf{t}'\mathbf{x}\neq\mathbf{0}\}\cap\{\omega(\mathbf{x},\boldsymbol{\mu}_0,\Sigma_0)>\zeta\}\cap\{\boldsymbol{\beta}_0'\mathbf{x}\in[K_1,K_2]\}], \tag{42}$$

by (40) and (41) we have

$$P(V_t)>\delta/4 \tag{43}$$

for all $\mathbf{t}\in S$. By A2

$$D(g(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}_0'\mathbf{x})))<D(g(\boldsymbol{\beta}_0'\mathbf{x}),m_i),\quad i=-1,1 \tag{44}$$

and

$$D(\mathbf{g}(\boldsymbol{\beta}_0'\mathbf{x}),m(g(\boldsymbol{\beta}_0'\mathbf{x})))\leq D(\mathbf{g}(\boldsymbol{\beta}_0'\mathbf{x}),m_3) \tag{45}$$

Eqs. (44) and (45) imply that for all $\mathbf{x}\in\mathbb{R}^p$ and $\mathbf{t}\in S$

$$E_{\beta_0}((\Phi_0^*(y,\mathbf{x},\mathrm{sign}(\mathbf{t}'\mathbf{x}))-\Phi_0(y,\mathbf{x},\boldsymbol{\beta}_0'\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0))|\mathbf{x})\geq\mathbf{0}. \tag{46}$$

Let us define $C_i(\lambda)=D(\lambda,m_i)-D(\lambda,m(\lambda))$, $i=-1,1$. Then for all $\lambda\in[g(K_1),g(K_2)]$ and $i=-1,1$, we have that C_i is positive and continuous. Then

$$c_i=\min_{\lambda\in[g(K_1),g(K_2)]}C_i(\lambda)>0,\quad i=-1,1.$$

Put $c_0=\min(c_{-1},c_1)$, then by (46), (42) and (43) we get that

$$\begin{aligned}
&E_{\beta_0}(\Phi_0^*(y,\mathbf{x},\mathrm{sign}(\mathbf{t}'\mathbf{x}))-\Phi_0(y,\mathbf{x},\boldsymbol{\beta}_0'\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0))\\
&\quad\geq E(E_{\beta_0}((\Phi_0^*(y,\mathbf{x},\mathrm{sign}(\mathbf{t}'\mathbf{x}))-\Phi_0(y,\mathbf{x},\boldsymbol{\beta}'\mathbf{x}))|\mathbf{x})\mathbf{I}(\mathbf{x}\in V_{\mathbf{t}}))\\
&\quad\geq E\left(\min_{\lambda\in[g(K_1),g(K_2)]}\min_{i\in\{-1,1\}}C_i(\lambda)w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\mathbf{I}(\mathbf{x}\in V_{\mathbf{t}})\right)\\
&\quad\geq c_0E(w(\mathbf{x},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\mathbf{I}(\mathbf{x}\in V_{\mathbf{t}}))\\
&\quad\geq c_0\zeta\delta/4.
\end{aligned}$$

This implies that $\tau\geq c_0\zeta\delta/4$ and therefore P3 holds.

P0 follows from B7. □

A.3. Proof of the asymptotic normality of WMT-estimators

Let $f(y,\lambda)$ be the density function of F_λ in the continuous case and the probability function in the discrete case and call $\psi_0(y,\lambda)=\partial f(y,\lambda)/\partial\lambda$. Then we have the following lemma.

Lemma 5. *Assume A1–A5 , B1–B7 and C1–C4. Then m is twice differentiable.*

Proof. For each $\lambda\in(\lambda^{(1)},\lambda^{(2)})$ $m(\lambda)$ is defined as the minimum in m of $E_\lambda(\rho(t(y)-m))$. Therefore $E_\lambda(\psi(t(y)-m(\lambda)))=0$ for all λ. By C3 and the Implicit Function Theorem m is differentiable and

$$m'(\lambda)=\frac{\int\psi(t(y)-m(\lambda))\psi_0(y,\lambda)\,dy}{E_\lambda[\psi'(t(y)-m(\lambda))]}$$

in the continuous case and

$$m'(\lambda)=\frac{\sum_y\psi(t(y)-m(\lambda))\psi_0(y,\lambda)}{E_\lambda[\psi'(t(y)-m(\lambda))]}$$

in the discrete case. Assumptions C1–C3 imply that m' is also differentiable. □

The following lemma is proved in Yohai (1985).

Lemma 6. *Let $z_1,\ldots,z_n$ be i.i.d. random vectors with distribution F and let $\phi:\mathbb{R}^p\times\mathbb{R}^h\longrightarrow\mathbb{R}$ be a continuous function satisfying*

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\varepsilon}|\phi(\mathbf{z},\boldsymbol{\theta})|$$

has finite expectation under F for some $\varepsilon>0$. Let $\widehat{\xi}_n\to\theta_0$ a.s. Then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\phi\left(\mathbf{z}_i,\widehat{\xi}_n\right)=E(\phi(\mathbf{z},\boldsymbol{\theta}_0))\ \text{a.s.}$$

The following lemma is proved in the Supplemental Material of Bergesio and Yohai (2011).

Lemma 7. *Let $c(\mathbf{z},\boldsymbol{\beta}):\mathbb{R}^p\times\mathbb{R}^k\to\mathbb{R}^k$ be a continuously differentiable function in $\boldsymbol{\beta}$ and let $z_1,\ldots,z_n$ be i.i.d. random vectors of dimension p. Consider a sequence of estimators $\widehat{\boldsymbol{\beta}}_n$ such* that $n^{1/2}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)=\mathrm{O_p}(1)$. *Suppose also that there exists $\zeta>0$ such that*

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq\zeta}|c(\mathbf{z},\boldsymbol{\beta})|$$

has finite expectation and that

$$E\left[\frac{\partial c(\mathbf{z},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]=\mathbf{0}.$$

Then

$$\frac{1}{n^{1/2}}\sum_{i=1}^n c\left(\mathbf{z}_i,\widehat{\boldsymbol{\beta}}_n\right)-\frac{1}{n^{1/2}}\sum_{i=1}^n c(\mathbf{z}_i,\boldsymbol{\beta}_0)=\mathrm{o_p}(1).$$

Proof of Theorem 3. The estimator $\widehat{\beta}_n$ satisfies

$$\sum_{i=1}^{n} \Psi(y_i, \mathbf{x}_i, \widehat{\beta}_n, \widehat{\mu}_n, \widehat{\Sigma}_n) = 0.$$

Using a Taylor expansion we get

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{J}_\psi\left(y_i, \mathbf{x}_i, \xi_n, \widehat{\mu}_n, \widehat{\Sigma}_n\right)\left(\widehat{\beta}_n - \beta_0\right) = -\frac{1}{n}\sum_{i=1}^{n} \Psi\left(y_i, \mathbf{x}_i, \beta_0, \widehat{\mu}_n, \widehat{\Sigma}_n\right), \tag{47}$$

where ξ_n is an intermediate point between $\widehat{\beta}_n$ and β_0. Since by Theorem 1 $\widehat{\beta}_n$ is strongly consistent to β_0, then ξ_n is strongly consistent to $\widehat{\beta}_0$ too. Then, by C4 and Lemma 6

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{J}_\psi\left(y_i, \mathbf{x}_i, \xi_n, \widehat{\mu}_n, \widehat{\Sigma}_n\right) = \mathbf{B} \text{ a.s.} \tag{48}$$

Since by C3 the matrix $\mathbf{B}$ is non-singular, for n large enough

$$(1/n)\sum_{i=1}^{n} \mathbf{J}_\psi(y_i, \mathbf{x}_i, \xi_n, \widehat{\mu}_n, \widehat{\Sigma}_n)$$

is non-singular too. Then from (47) we get

$$\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) = -\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{J}_\psi\left(y_i, \mathbf{x}_i, \xi_n, \widehat{\mu}_n, \widehat{\Sigma}_n\right)\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi\left(y_i, \mathbf{x}_i, \beta_0, \widehat{\mu}_n, \widehat{\Sigma}_n\right). \tag{49}$$

By the Central Limit Theorem we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi(y_i, \mathbf{x}_i, \beta_0, \mu_0, \Sigma_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{A}) \tag{50}$$

and by Lemma 7

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left[\Psi\left(y_i, \mathbf{x}_i, \beta_0, \widehat{\mu}_n, \widehat{\Sigma}_n\right) - \Psi(y_i, \mathbf{x}_i, \beta_0, \mu_0, \Sigma_0)\right] = o_p(1). \tag{51}$$

Consequently, the theorem follows from (48)–(51) and Slutzky's lemma. □

A.4. Proof of Theorem 3

Suppose that ε is a positive real number such that there exists a sequence of distribution functions H_k such that $T((1-\varepsilon)H_0 + \varepsilon H_k) \to \infty$ as $k \to \infty$. We will show that $\varepsilon \geq \varepsilon_0$. We write $\beta_k = T((1-\varepsilon)H_0 + \varepsilon H_k)$ for brevity. Then we have

$$\begin{aligned}(1-\varepsilon)E_{H_0}(\rho(t(y) - m(g(\beta_k'\mathbf{x})))) &\leq (1-\varepsilon)E_{H_0}(\rho(t(y) - m(g(\beta_k'\mathbf{x})))) + \varepsilon E_{H_\mathbf{k}}(\rho(t(y) - m(g(\beta_k'\mathbf{x}))))\\ &\leq (1-\varepsilon)E_{H_0}(\rho(t(y) - m(g(\beta_0'\mathbf{x})))) + \varepsilon E_{H_\mathbf{k}}(\rho(t(y) - m(g(\beta_0'\mathbf{x}))))\\ &\leq (1-\varepsilon)E_{H_0}(\rho(t(y) - m(g(\beta_0'\mathbf{x})))) + \varepsilon.\end{aligned} \tag{52}$$

Let $\alpha_k = \beta_k / \|\beta_k\|$, then we may assume without loss of generality that $\alpha_k \to \alpha$. Then

$$\lim_{k\to\infty} (1-\varepsilon)E_{H_0}(\rho(t(y) - m(g(\beta_k'\mathbf{x})))) = (1-\varepsilon)E_{H_0}[\rho(t(y) - m_1)I(\alpha'\mathbf{x} < 0) + \rho(t(y) - m_2)I(\alpha'\mathbf{x} > 0)] \tag{53}$$

$$\lim_{k\to\infty} (1-\varepsilon)E_{H_0}(\rho(t(y) - m(g(\beta_k'\mathbf{x})))) \geq (1-\varepsilon)E_{H_0}(\min(\rho(t(y) - m_1), \rho(t(y) - m_2))). \tag{54}$$

Combining inequalities (52) and (54) we obtain

$$(1-\varepsilon)E_{H_0}(\min(\rho(t(y) - m_1), \rho(t(y) - m_2))) \leq (1-\varepsilon)E_{H_0}(\rho(t(y) - m(g(\beta_0'\mathbf{x})))) + \varepsilon,$$

and then

$$\varepsilon \geq \frac{E_{H_0}(\min(\rho(t(y) - m_1), \rho(t(y) - m_2))) - E_{H_0}(\rho(t(y) - m(g(\beta_0'\mathbf{x}))))}{1 + E_{H_0}(\min(\rho(t(y) - m_1), \rho(t(y) - m_2))) - E_{H_0}(\rho(t(y) - m(g(\beta_0'\mathbf{x}))))}.$$

This proves the theorem.

References

Bergesio, A., Yohai, V.J., 2011. Projection estimators for generalized linear models. Journal of the American Statistical Association 106, 661–671.
Bianco, A., Yohai, V.J., 1996. Robust estimation in the logistic regression model. In: Rieder, H. (Ed.), Robust Statistics, Data Analysis and Computer Intensive Methods, Proceedings of the Workshop in Honor of Peter Huber, Lecture Notes in Statistics, vol. 109. Springer-Verlag, New York, pp. 7–34.
Bianco, A.M., Boente, G., Rodrigues, I.M., 2013. Resistant estimators in Poisson and Gamma models with missing responses and an application to outlier detection. Journal of Multivariate Analysis 114, 209–226.
Bondell, H.D., 2005. Minimum distance estimation for the logistic regression model. Biometrika 92, 724–731.

Breslow, N.E., 1996. Generalized linear models: checking assumptions and strengthening conclusions. Statistica Applicata 8, 23–41.
Cantoni, E., Ronchetti, E., 2001. Robust inference for generalized linear models. Journal of the American Statistical Association 96, 1022–1030.
Carroll, R.J., Pederson, S., 1993. On robustness in the logistic regression model. Journal of the Royal Statistical Society Series B 55, 693–706.
Christmann, A., 1994. Least median of weighted squares in logistic regression with large strata. Biometrika 81, 413–417.
Čížek, P., 2008. Robust and efficient adaptive estimation of binary-choice regression models. Journal of the American Statistical Association 103, 687–696.
Croux, C., Haesbroeck, G., 2003. Implementing the Bianco and Yohai estimator for logistic regression. Computational Statistics and Data Analysis 44, 273–295.
Hampel, F.R., 1971. A general qualitative definition of robustness. Annals of Mathematical Statistics 42, 1887–1996.
Künsch, H., Stefanski, L., Carroll, R., 1989. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. Journal of the American Statistical Association 84, 460–466.
Maronna, R.A., Bustos, O., Yohai, V.J., 1979. Bias and efficiency robustness of general M-estimators for regression with random carriers. In: Gasser, T., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation. Lectures Notes in Mathematics, vol. 757. . Springer Verlag, New York, pp. 91–111.
Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Chichester, Wiley.
Morgenthaler, S., 1992. Least-absolute-deviations for generalized linear models. Biometrika 79, 747–754.
Rousseeuw, P.J., Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85, 633–651.
Rousseeuw, P.J., Christmann, A., 2003. Robustness against separation and outliers in logistic regression. Computational Statistics and Data Analysis 43, 315–332.
Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61, 439–447.
Yohai, V.J., 1985. High Breakdown Point and High Efficiency Robust Estimates for Regression. Technical Report No. 66. Department of Statistics, University of Washington. Available at http://www.stat.washington.edu/research/reports/1985/tr066.pdf.