Article

Robust Inference for Generalized Linear Models

CANTONI, Eva, RONCHETTI, Elvezio

Abstract

By starting from a natural class of robust estimators for generalized linear models based on the notion of quasi-likelihood, we define robust deviances that can be used for stepwise model selection as in the classical framework. We derive the asymptotic distribution of tests based on robust deviances, and we investigate the stability of their asymptotic level under contamination. The binomial and Poisson models are treated in detail. Two applications to real data and a sensitivity analysis show that the inference obtained by means of the new techniques is more reliable than that obtained by classical estimation and testing procedures.

Robust Inference for Generalized Linear Models

Eva Cantoni and **Elvezio Ronchetti**

Department of Econometrics

University of Geneva

CH - 1211 Geneva 4, Switzerland

May 1999

Revised January 2001

Abstract

By starting from a natural class of robust estimators for generalized linear models based on the notion of quasi-likelihood, we define robust deviances that can be used for stepwise model selection as in the classical framework. We derive the asymptotic distribution of tests based on robust deviances and we investigate the stability of their asymptotic level under contamination. The binomial and Poisson models are treated in detail. Two applications to real data and a sensitivity analysis show that the inference obtained by means of the new techniques is more reliable than that obtained by classical estimation and testing procedures.

1 Introduction

Generalized linear models (McCullagh and Nelder, 1989) are a powerful and popular technique for modeling a large variety of data. In particular, generalized linear models allow to model the relationship between the predictors and a function of the mean of the response for continuous and discrete response variables. The response variables Y_i, for $i = 1, \ldots, n$ are supposed to come from a distribution belonging to the exponential family, such that $\mathsf{E}[Y_i] = \mu_i$ and $\mathsf{V}[Y_i] = V(\mu_i)$ for $i = 1, \ldots, n$ and

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of parameters, $\mathbf{x}_i \in \mathbb{R}^p$, and $g(.)$ is the link function.

The non-robustness of the maximum likelihood estimator for $\boldsymbol{\beta}$ has been studied extensively in the literature: cf. for instance the early work of Pregibon (1982) on logistic regression, Stefanski, Carroll, and Ruppert (1986), Künsch, Stefanski, and Carroll (1989), Morgenthaler (1992), and Ruckstuhl and Welsh (1999). In more recent work, Preisser and Qaqish (1999) consider a class of robust estimators in the general framework of generalized estimating equations.

The quasi-likelihood estimator of the parameter of model (1) (see Wedderburn, 1974, McCullagh and Nelder, 1989, and Heyde, 1997) shares the same non-robustness properties. This estimator is the solution of the system of estimating equations

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} Q(y_i, \mu_i) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \mu_i' = 0, \tag{2}$$

where $\mu_i' = \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i$, and $Q(y_i, \mu_i)$ is the quasi-likelihood function. The solution of (2) is an M-estimator (see Huber, 1981, and Hampel, Ronchetti, Rousseeuw, and Stahel, 1986) defined by the score function $\tilde{\psi}(y_i, \mu_i) = \frac{(y_i - \mu_i)}{V(\mu_i)} \mu_i'$. Its influence function (Hampel, 1974 and Hampel et al., 1986) is proportional to $\tilde{\psi}$ and is unbounded.

Therefore, large deviations of the response from its mean or outlying points in the explanatory variables $\mathbf{x}_i$ can have a large influence on the estimator. Thus, the quasi-likelihood estimator – as well as the maximum likelihood estimator – is not robust. Several robust alternatives have been proposed in the literature; see the references given above.

However, in spite of the fair amount of existing literature, robust inference for generalized linear models seems to be very limited. Moreover, only the logistic regression situation is usually considered in detail, and the problem of developing robust alternatives to classical tests is not addressed globally for the whole class of generalized linear models.

In this paper we propose a robust approach to inference based on robust deviances which are natural generalizations of quasi-likelihood functions. Our robust deviances are based on the same class of robust estimators as that proposed by Preisser and Qaqish (1999) in the more general setup of generalized estimating equations. Although these estimators are not optimally robust, they form a class of M-estimators easy to deal with, and which admits handy inference not only for logistic regression but for the whole class of generalized linear models.

One could argue that two alternative approaches could be considered. A first possibility would be to view variable selection as a parametric hypothesis and to use Wald, score or likelihood ratio tests for which robust versions are available; see e.g. Heritier and Ronchetti (1994) and Markatou and He (1994). While this would in principle be feasible, Wald and score tests do not seem to be used much in the classical analysis of generalized linear models. Moreover, robust likelihood ratio tests cannot be proposed in this case, because the optimal robust score function does not admit an analytic primitive function and numerical integration in the space

of parameters for computing such a primitive is generally unfeasible. A second approach would be to rely on the robust model selection based on Akaike Criterion, Mallows' C_p or similar techniques; see e.g. Ronchetti and Staudte (1994), Sommer and Huggins (1996) and Ronchetti (1997) for a review. This approach has the advantage to perform a full model search. However, when the number of variables is moderate to large such a full search is impossible and a stepwise selection is the only feasible alternative.

For these reasons and in view of the importance of the notion of deviance for model building in generalized linear models, we propose robust deviances based on generalizations of quasi-likelihood functions. The general structure of the classical approach by quasi-likelihood is preserved, which offers the advantage of having robust tools playing the same role as deviances, ANOVA tables, stepwise procedures, and so on.

The paper is organized as follows. In the next section we discuss robust estimators of a generalized linear model based on quasi-likelihood. As an illustration, we focus in particular on the estimation of binomial and Poisson models. In Section 3, we discuss inference and propose a family of test statistics for model selection. We derive their asymptotic distribution through the development of an asymptotically equivalent quadratic form and we study their robustness properties through the influence function. Section 4 presents some computational aspects and Section 5 gives two applications. Finally, in Section 6 we discuss some potential research directions.

2 Robust Estimation Based on Quasi-likelihood

2.1 General Definition

We consider a general class of M-estimators of Mallows's type, where the influence of deviations on y and on $\mathbf{x}$ are bounded separately. The estimator is the solution of the estimating equations:

$$\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mu_i) = \mathbf{0}, \tag{3}$$

where $\boldsymbol{\psi}(y, \mu) = \nu(y, \mu) w(\mathbf{x}) \mu' - a(\boldsymbol{\beta})$, $a(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}[\nu(y_i, \mu_i)] w(\mathbf{x}_i) \mu_i'$ with the expectation taken with respect to the conditional distribution of $y|\mathbf{x}$, $\nu(\cdot, \cdot)$, $w(\mathbf{x})$ are weight functions defined below, and $\mu_i = \mu_i(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$. The constant $a(\boldsymbol{\beta})$ ensures the Fisher consistency of the estimator. The estimating equation (3) for generalized linear models is a special case of equation (1) p. 575 for generalized estimating equations in Preisser and Qaqish (1999), where our function $\nu(y, \mu) w(\mathbf{x})$ is (in their notation) $V^{-1}(\mu) w(\mathbf{x}, y, \boldsymbol{\beta})(y - \mu)$ and $a(\boldsymbol{\beta}) = \mu' \; V^{-1}(\mu) \; c$.

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. The estimating equation (3) corresponds to the minimization of the quantity

$$Q_M(\mathbf{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} Q_M(y_i, \mu_i), \tag{4}$$

with respect to $\boldsymbol{\beta}$, where the functions $Q_M(y_i, \mu_i)$ can be written as

$$Q_M(y_i, \mu_i) = \int_{\tilde{s}}^{\mu_i} \nu(y_i, t) w(\mathbf{x}_i) dt - \frac{1}{n} \sum_{j=1}^{n} \int_{\tilde{t}}^{\mu_j} \mathsf{E}\big[\nu(y_j, t) w(\mathbf{x}_j)\big] dt, \tag{5}$$

with $\tilde{s}$ such that $\nu(y_i, \tilde{s}) = 0$, and $\tilde{t}$ such that $\mathsf{E}[\nu(y_i, \tilde{t})] = 0$. Note that differences of deviances, as the test statistic (8), are independent of $\tilde{s}$ and $\tilde{t}$.

The structure of (3) is suggested by the classical quasi-likelihood equations. The estimator defined by equation (3) is an M-estimator characterized by the score func-

tion $\boldsymbol{\psi}(y_i, \mu_i) = \nu(y_i, \mu_i) w(\mathbf{x}_i) \mu_i' - a(\boldsymbol{\beta})$. Its influence function is then $\mathsf{IF}(y; \boldsymbol{\psi}, F) = M(\boldsymbol{\psi}, F)^{-1} \boldsymbol{\psi}(y, \mu)$, where $M(\boldsymbol{\psi}, F) = -\mathsf{E}[\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\psi}(y, \mu)]$; cf. Hampel et al. (1986). Moreover, the estimator has an asymptotic normal distribution with asymptotic variance $\Omega = M(\boldsymbol{\psi}, F)^{-1} Q(\boldsymbol{\psi}, F) M(\boldsymbol{\psi}, F)^{-1}$, where $Q(\boldsymbol{\psi}, F) = \mathsf{E}[\boldsymbol{\psi}(y, \mu) \boldsymbol{\psi}(y, \mu)^T]$. It is then clear that the choice of a bounded function $\boldsymbol{\psi}$ ensures robustness by putting a bound on the influence function. Therefore, a bounded function $\nu(y, \mu)$ is introduced to control deviations in the y-space, and leverage points are down-weighted by the weights $w(\mathbf{x})$. Simple choices for $\nu(\cdot, \cdot)$ and $w(\cdot)$ suggested by robust estimators in linear models are $\nu(y_i, \mu_i) = \psi_c(r_i) \frac{1}{V^{1/2}(\mu_i)}$ (see (6) below) and $w(\mathbf{x}_i) = \sqrt{1 - h_i}$, where h_i is the i-th diagonal element of the hat matrix $H = X(X^T X)^{-1} X^T$. More sophisticated choices for $w(\cdot)$ are available (see Staudte and Sheather, 1990, p. 258, for a discussion in linear regression or Carroll and Welsh, 1988). Weights defined on H do not have high breakdown properties, and from this point of view, other choices of $w(\mathbf{x}_i)$ are more suitable. For example, $w(\mathbf{x}_i)$ can be chosen as the inverse of the Mahalanobis distance defined through a high breakdown estimate of the center and of the covariance matrix of the $\mathbf{x}_i$ (see, for example, the minimum volume ellipsoid estimator or the minimum covariance determinant estimator in Rousseeuw and Leroy, 1987, p. 258 ff.). Finally notice that the choice of $\nu(y_i, \mu_i) = \frac{y_i - \mu_i}{V(\mu_i)}$ and $w(\mathbf{x}_i) = 1$ for all i, recovers the classical quasi-likelihood estimator, so that for a judicious choice of $\nu(y_i, \mu_i)$ and of the weights $w(\mathbf{x}_i)$, the function $Q_M(\mathbf{y}, \boldsymbol{\mu})$ can be seen as the robust counterpart of the classical quasi-likelihood function.

The form of this estimator is attractive because the estimating equation (3) corresponds to the minimization of (4) and this leads to a natural definition of robust deviance; see Section 3.1.

2.2 Robust Estimation for Binomial and Poisson Models

We consider here the particular case of (3), defined by $\nu(y_i, \mu_i) = \psi_c(r_i)\frac{1}{V^{1/2}(\mu_i)}$, where $r_i = \frac{y_i - \mu_i}{V^{1/2}(\mu_i)}$ are the Pearson residuals and ψ_c is the Huber function defined by

$$\psi_c(r) = \begin{cases} r & \mid r \mid \leq c, \\ c\,\mathrm{sign}(r) & \mid r \mid > c. \end{cases} \tag{6}$$

We call the estimator defined in this way, the Mallows quasi-likelihood estimator. It solves the set of estimating equations

$$\sum_{i=1}^{n} \left[\psi_c(r_i) w(\mathbf{x}_i) \frac{1}{V^{1/2}(\mu_i)} \mu_i' - a(\boldsymbol{\beta})\right] = \mathbf{0}, \tag{7}$$

where $a(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[\psi_c(r_i)] w(\mathbf{x}_i) \frac{1}{V^{1/2}(\mu_i)} \mu_i'$. Using the same notation as in the linear regression case, when $w(\mathbf{x}_i) = 1$ we call this estimator Huber quasi-likelihood estimator.

The tuning constant c is typically chosen to ensure a given level of asymptotic efficiency. In Section 3.2 we propose an alternative procedure for the choice of the tuning constant. $a(\boldsymbol{\beta})$ is a correction term to ensure Fisher consistency; see Hampel et al. (1986) for general parametric models and He and Simpson (1993), Section 4.1, for power series distributions. Note that $a(\boldsymbol{\beta})$ can be computed explicitly for binomial and Poisson models and does not require numerical integration; cf. Appendix A. The matrices $M(\boldsymbol{\psi}_c, F)$ and $Q(\boldsymbol{\psi}_c, F)$ can also be easily computed for the Mallows quasi-likelihood estimator:

$$Q(\boldsymbol{\psi}_c, F) = \frac{1}{n} X^T A X - a(\boldsymbol{\beta}) a(\boldsymbol{\beta})^T,$$

where A is a diagonal matrix with elements $a_i = \mathsf{E}[\psi_c(r_i)^2] w^2(\mathbf{x}_i) \frac{1}{V(\mu_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2$, and

$$M(\boldsymbol{\psi}_c, F) = \frac{1}{n} X^T B X,$$

where B is a diagonal matrix with elements $b_i = \mathsf{E}[\psi_c(r_i)\frac{\partial}{\partial\mu_i}\log h(y_i|\mathbf{x}_i,\mu_i)]\frac{1}{V^{1/2}(\mu_i)}w(\mathbf{x}_i)(\frac{\partial\mu_i}{\partial\eta_i})^2$, and $h(\cdot)$ is the conditional density or probability of $y_i|\mathbf{x}_i$. We refer to Appendix B for further details and for the computation of these matrices for binomial and Poisson models.

3 Robust Inference

3.1 Model Selection Based on Robust Deviances

The function $Q_M(\mathbf{y},\boldsymbol{\mu})$ defined in (4) and (5) allows to develop robust tools for inference and model selection based on robust quasi-deviances.

Denote by $\mathbf{a} = (\mathbf{a}_{(1)}^T, \mathbf{a}_{(2)}^T)^T$ the partition of a vector $\mathbf{a}$ into $(p-q)$ and q components and the corresponding partition of a matrix A by

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where $A_{11} \in \mathbb{R}^{(p-q)\times(p-q)}$, $A_{12} \in \mathbb{R}^{(p-q)\times q}$, $A_{21} \in \mathbb{R}^{q\times(p-q)}$ and $A_{22} \in \mathbb{R}^{q\times q}$.

To evaluate the adequacy of a model, we define a robust goodness-of-fit measure — called robust quasi-deviance — based on the notion of robust quasi-likelihood function, i.e.

$$D_{QM}(\mathbf{y},\boldsymbol{\mu}) = -2Q_M(\mathbf{y},\boldsymbol{\mu}) = -2\sum_{i=1}^{n} Q_M(y_i,\mu_i),$$

where Q_M is defined by (4) and (5).

$D_{QM}(\mathbf{y},\boldsymbol{\mu})$ describes the quality of a fit and will be used to define a statistic for model selection. Let us consider the model M_p, with p parameters. Suppose that the corresponding set of parameters is $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^T = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T)^T$. We are interested in testing the null hypothesis $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$. This is equivalent to

consider a nested model $M_{p-q} \subset M_p$ with $(p-q)$ parameters, and testing whether the sub-model M_{p-q} holds.

We estimate the vector of parameters by solving (3) for the complete model, and we obtain an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Under the null hypothesis, the same procedure yields an estimator $\dot{\boldsymbol{\beta}}$ of $(\boldsymbol{\beta}_{(1)}, \mathbf{0})$. We write $\hat{\boldsymbol{\mu}}$ and $\dot{\boldsymbol{\mu}}$ for the estimated linear predictors associated to the estimate $\hat{\boldsymbol{\beta}}$ and $\dot{\boldsymbol{\beta}}$ respectively. Then, we define a robust measure of discrepancy between two nested models by

$$\begin{aligned} \Lambda_{QM} &= \Big[D_{QM}(\mathbf{y}, \dot{\boldsymbol{\mu}}) - D_{QM}(\mathbf{y}, \hat{\boldsymbol{\mu}}) \Big] \\ &= 2\Big[\sum_{i=1}^{n} Q_M(y_i, \hat{\mu}_i) - \sum_{i=1}^{n} Q_M(y_i, \dot{\mu}_i) \Big], \end{aligned} \tag{8}$$

where the function $Q_M(y_i, \mu_i)$ is defined by (5).

The statistic (8) is in fact a generalization of the quasi-deviance test for generalized linear models, which is recovered by taking $Q_M(y_i, \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt$. Moreover, when the link function is the identity (linear regression), (8) becomes the τ-test statistic defined in Hampel et al. (1986), Chapter 7.

The same forms for the functions $\nu(y_i, \mu_i)$ and $w(\mathbf{x}_i)$ as in the estimation problem can be considered here. In particular, a Mallows quasi-deviance statistic can be defined by taking $\nu(y_i, \mu_i) = \psi_c(r_i)/V^{1/2}(\mu_i)$.

The following Proposition establishes the asymptotic distribution of the test statistic (8). We assume the conditions for the existence, consistency, and asymptotic normality of M-estimators as given by (A.1)-(A.9) in Heritier and Ronchetti (1994), p. 902. These conditions have been studied by Huber (1967, 1981), Clarke (1986) and Bednarski (1993).

Proposition 1 *Under conditions (A.1)-(A.9) in Heritier and Ronchetti (1994), [C1], [C2] of Appendix C, and under* $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$, *the test statistic* Λ_{QM} *defined*

by (8) *equals*

$$n\mathbf{L}_n^T C(\boldsymbol{\psi}, F)\mathbf{L}_n + o_P(1) = n\mathbf{R}_{n(2)}^T M(\boldsymbol{\psi}, F)_{22.1}\mathbf{R}_{n(2)} + o_P(1), \tag{9}$$

where $C(\boldsymbol{\psi}, F) = M^{-1}(\boldsymbol{\psi}, F) - \tilde{M}^+(\boldsymbol{\psi}, F)$, $\sqrt{n}\mathbf{L}_n$ *is normally distributed* $\mathcal{N}\big(\mathbf{0}, Q(\psi, F)\big)$, $M(\boldsymbol{\psi}, F)_{22.1} = M(\boldsymbol{\psi}, F)_{22} - M(\boldsymbol{\psi}, F)_{12}^T M(\boldsymbol{\psi}, F)_{11}^{-1} M(\boldsymbol{\psi}, F)_{12}$, *and* $\sqrt{n}\mathbf{R}_n$ *is normally distributed* $\mathcal{N}\big(\mathbf{0}, M^{-1}(\psi, F)Q(\psi, F)M^{-1}(\psi, F)\big)$.

Moreover, Λ_{QM} *is asymptotically distributed as*

$$\sum_{i=1}^{q} d_i N_i^2,$$

where $N_1, \ldots, N_q$ *are independent standard normal variables,* $d_1, \ldots, d_q$ *are the* q *positive eigenvalues of the matrix* $Q(\boldsymbol{\psi}, F)\big(M^{-1}(\boldsymbol{\psi}, F) - \tilde{M}^+(\boldsymbol{\psi}, F)\big)$, *and* $\tilde{M}^+(\boldsymbol{\psi}, F)$ *is such that* $\tilde{M}^+(\boldsymbol{\psi}, F)_{11} = M(\boldsymbol{\psi}, F)_{11}^{-1}$ *and* $\tilde{M}^+(\boldsymbol{\psi}, F)_{12} = 0$, $\tilde{M}^+(\boldsymbol{\psi}, F)_{21} = 0$, $\tilde{M}^+(\boldsymbol{\psi}, F)_{22} = 0$.

The proof is given in Appendix D. A similar result can be obtained for the distribution of Λ_{QM} under contiguous alternatives $\boldsymbol{\beta}_{(2)} = n^{-1/2}\Delta$. In such a case Λ_{QM} is asymptotically distributed as $\sum_{i=1}^{q}(d_i^{1/2}N_i + S^T\Delta)^2$, where S is such that $SS^T = M_{22.1}$ and $S^T(M^{-1}(\boldsymbol{\psi}, F_{\boldsymbol{\beta}_0})Q(\boldsymbol{\psi}, F_{\boldsymbol{\beta}_0})M^{-1}(\boldsymbol{\psi}, F_{\boldsymbol{\beta}_0}))_{22}S = D$ and D is the diagonal matrix with elements $d_1, \ldots, d_q$.

3.2 Robustness Properties and Choice of the Tuning Constant

The robustness properties of the test based on (8) can be investigated by showing that a small amount of contamination at a point $\mathbf{z}$ has bounded influence on the asymptotic level and power of the test. This ensures the local stability of the test.

The global reliability (or robustness against large deviations) could be measured by the breakdown point as defined in He, Simpson, and Portnoy (1990). However, we focus here on small deviations which are probably the main concern at the inference stage of a statistical analysis.

We consider the sequence of ϵ-contaminations

$$F_{\epsilon,n} = \big(1 - \frac{\epsilon}{\sqrt{n}}\big) F_{\boldsymbol{\beta}_0} + \frac{\epsilon}{\sqrt{n}} G, \tag{10}$$

where G is an arbitrary distribution (see Heritier and Ronchetti, 1994) and investigate the asymptotic level of the test under (10).

Proposition 2 *Consider a parametric model $F_{\boldsymbol{\beta}_0}$ and the null hypothesis H_0 : $\boldsymbol{\beta}_{(2)} = 0$. Denote by $F^{(n)}$ the empirical distribution and by $\mathbf{U}_n$ the functional $\mathbf{U}(F^{(n)})$ such that $\mathbf{U}(F_{\boldsymbol{\beta}_0}) = 0$, $\mathsf{IF}(\mathbf{z}; \mathbf{U}, F_{\boldsymbol{\beta}_0})$ is bounded and*

$$\sqrt{n}(\mathbf{U}_n - \mathbf{U}(F_{\epsilon,n})) \sim \mathcal{N}(0, \Sigma) \tag{11}$$

uniformly over the ϵ-contamination $F_{\epsilon,n}$. Let $\alpha(F)$ be the level of the test based on the quadratic form $n\mathbf{U}_n^T A \mathbf{U}_n$ when the underlying distribution is F. The nominal level is $\alpha(F_{\boldsymbol{\beta}_0}) = \alpha_0$.

Then, under the ϵ-contamination $F_{\epsilon,n}$, we have

$$\lim_{n\to\infty} \alpha(F_{\epsilon,n}) = \alpha_0 + \epsilon^2 \boldsymbol{\kappa}^T \cdot \operatorname{diag}\left(P \left(\int \mathsf{IF}(\mathbf{z};\mathbf{U}, F_{\boldsymbol{\beta}_0}) dG(\mathbf{z})\right) \left(\int \mathsf{IF}(\mathbf{z};\mathbf{U}, F_{\boldsymbol{\beta}_0}) dG(\mathbf{z})\right)^T P^T\right) + o(\epsilon^2),$$

where $\boldsymbol{\kappa} = -\frac{\partial}{\partial \boldsymbol{\lambda}} H_{d_1,\ldots,d_q}(\eta_{1-\alpha_0}; \boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\mathbf{0}}$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^T = (\xi_1^2, \ldots, \xi_q^2)^T$, $H_{d_1,\ldots,d_q}(.; \boldsymbol{\lambda})$ is the c.d.f. of the random variable $\sum_{i=1}^q d_i \chi_1^2(\xi_i^2)$, $\eta_{1-\alpha_0}$ is the $(1-\alpha_0)$-quantile of $\sum_{i=1}^q d_i \chi_1^2(0)$, P is an orthogonal matrix such that $P^T D P = \Sigma A$, and D is the diagonal matrix with elements $d_1, \ldots, d_q$, the eigenvalues of ΣA. Moreover, $\operatorname{diag}(R)$ *indicates the vector with components the diagonal elements of the matrix R.*

If the influence function of the functional $\mathbf{U}$ is bounded, then the asymptotic level under contamination is also bounded. The proof of this proposition is presented in Appendix E. A similar result can be obtained for the power, showing that the asymptotic power is stable under contamination.

Note that this proposition generalizes the result of Proposition 4 in Heritier and Ronchetti (1994), which can be recovered by taking $\boldsymbol{\lambda} = \lambda_1 = \delta(\epsilon)$ and $A = I_q$.

The general result of Proposition 2 can be applied to the robust quasi-likelihood test statistic (8) and in the special case of a point mass contamination $G(\mathbf{z}) = \Delta_{\mathbf{z}}$. This gives the following Corollary.

Corollary 1 *Under conditions (A.1)-(A.9) in Heritier and Ronchetti (1994), and for any M-estimator $\hat{\boldsymbol{\beta}}_{(2)}$ with bounded influence function, the asymptotic level of the robust quasi-likelihood test statistic* (8) *under a point mass contamination is given by*

$$\lim_{n\to\infty} \alpha(F_{\epsilon,n}) = \alpha_0 + \epsilon^2 \boldsymbol{\kappa}^T \cdot \mathrm{diag}\left(P \; \mathsf{IF}(\mathbf{z}; \hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0}) \mathsf{IF}(\mathbf{z}; \hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0})^T \; P^T \right) + o(\epsilon^2),$$

where P is an orthogonal matrix such that $P^T D P = \Omega_{22} M_{22.1}$, Ω is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ defined in Section 2.1, and D is the diagonal matrix with elements $d_1, \ldots, d_q$ defined in Proposition 1.

The result is obtained by applying Proposition 2 with $G(\mathbf{z}) = \Delta_{\mathbf{z}}$, $\mathbf{U} = \hat{\boldsymbol{\beta}}_{(2)}$, $\Sigma = \Omega_{22}$, $A = M_{22.1}$, and by using the Fréchet differentiability of $\hat{\boldsymbol{\beta}}_{(2)}$; see Heritier and Ronchetti (1994).

Hence, a bounded influence M-estimator $\hat{\boldsymbol{\beta}}_{(2)}$ ensures a bound on the asymptotic level of the robust quasi-likelihood test under contamination.

We can now undertake a complete robust analysis of a generalized linear model: the estimation of parameters can be performed via M-estimation according to (3), and the test statistic (8) allows us to make inference and model choice.

The function $\nu(y_i, \mu_i)$ which appears in the definition of $Q_M(y_i, \mu_i)$, is often tuned by a constant; cf. for instance (6). As suggested in Ronchetti and Trojani (2001), we can consider the problem from the point of view of inference and choose the constant that controls the maximal bias on the asymptotic level of the test in a neighborhood of the model. To serve this last purpose, one can use the Corollary above. The maximal level α of the robust quasi-likelihood test statistic in a neighborhood of the model of radius ϵ is given by

$$\alpha = \alpha_0 + \epsilon^2 \gamma(\hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0})^2 \boldsymbol{\kappa}^T \operatorname{diag}\left(P \mathbf{1}\mathbf{1}^T P^T\right), \tag{12}$$

where $\gamma(\hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0}) = \sup_{\mathbf{z}} ||\mathsf{IF}(\mathbf{z}; \hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0})||$ and $\mathbf{1} = (1, \ldots, 1)^T$.

By (12), we can write

$$b = \frac{1}{\epsilon} \sqrt{\frac{\alpha - \alpha_0}{\boldsymbol{\kappa}^T \operatorname{diag}(P \mathbf{1}\mathbf{1}^T P^T)}}, \tag{13}$$

where b is the bound on the influence function of the estimator $\hat{\boldsymbol{\beta}}_{(2)}$. Then, for a fixed amount of contamination ϵ and by imposing a maximal error on the level of the test $\alpha - \alpha_0$, one can determine the bound b on the influence function of the estimator, and hence the tuning constant by solving $b = \gamma(\hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0}) = \gamma_c$ with respect to c. For example, if $q = 1$ we have $P = 1$, $\operatorname{diag}(P \mathbf{1}\mathbf{1}^T P^T) = 1$, and $\kappa = 0.1145$, see Ronchetti and Trojani (2001). In practice, the supremum on $\mathbf{z} = (y, \mathbf{x})$ is taken as the maximum over the sample of the supremum on $y|\mathbf{x}$. Note also that the solution depends on the unknown parameter $\boldsymbol{\beta}_0$; our experience shows that it does not vary much for different values of $\boldsymbol{\beta}$, so that one can safely plug-in a reasonable (robust) estimate. This is valid for a single test. However, in a stepwise procedure (as in

Section 5) several tests are performed, and one would have to choose a different value of c for each test. Since this is unreasonable from a practical point of view, we suggest to choose a global value of c by solving $b = \sup_{\mathbf{z}} ||\mathsf{IF}(\mathbf{z}; \hat{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0})||$, based on the fact that $\gamma(\hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0}) = \sup_{\mathbf{z}} ||\mathsf{IF}(\mathbf{z}; \hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0})|| \leq ||\sup_{\mathbf{z}} \mathsf{IF}(\mathbf{z}; \hat{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0})||$.

4 Computational Aspects

The solution of equation (3) can be obtained numerically by a Newton-Raphson procedure or by a Fisher scoring procedure. In the latter case, the algorithm is also known as the influence algorithm; cf. for instance Hampel et al. (1986), p. 263. However, there is a potential problem with multiple roots of equation (3). In this case, we recommend to use a bootstrap root search as proposed in Markatou, Basu, and Lindsay (1998), p. 743-744, based on the objective function Q_M defined in (4) as a selection rule; see also Hanfelt and Liang (1995).

The test statistic Λ_{QM} of equation (8), can be computed directly. It involves n one-dimensional integrations, which are performed numerically. Our experience shows that it works well for binomial and Poisson models. To avoid these numerical integrations – especially in the case when n is large – one can consider using the asymptotic quadratic forms of Proposition 1 given by (9) which are asymptotically equivalent to the test statistic Λ_{QM}. A systematic study on the comparison of (8) with the asymptotic equivalent quadratic forms (9) is left for further work. Moreover, critical regions or p-values for the test statistic Λ_{QM} are easy to obtain. In fact, linear combinations of χ^2_1 variables have been well studied in the literature. Algorithms for the computation of these p-values have been proposed among others by Davies (1980) and by Farebrother (1990). Analytical approximations of these distributions

were studied by Pearson (1959) and Imhof (1961).

S-PLUS (MathSoft, Seattle) routines for estimation and inference based on robust quasi-likelihood are collected in a library and are available from the authors.

5 Applications

5.1 Binomial models

In this section, we analyze the damaged carrots dataset. It is taken from Phelps (1982) and is discussed by Williams (1987) and used in McCullagh and Nelder (1989) to illustrate techniques for checking for isolated departures from the model, because of the presence of an outlier in the y-space. The data are issued from a soil experiment and give the proportion of carrots showing insect damage in a trial with three blocks and eight dose levels of insecticide. The logarithm of the dose ranges from 1.52 to 2.36 in an equally spaced grid. The sample size is 24.

We assume a binomial model with logit link

$$\log\left(\frac{\mu}{m-\mu}\right) = \beta_0 + \beta_1 \log(\texttt{dose}) + \beta_2 \texttt{block2} + \beta_3 \texttt{block1},$$

where $\mu = \mathsf{E}[Y] = \mathsf{E}[\text{number of damaged carrots}]$, $\texttt{block}i$, $i = 1, 2$ are indicators variables taking the value of 1 if measures are taken in block i and 0 otherwise.

Different techniques — plot of deviance residuals, plot of Pearson residuals and Cook's distance — show that there is a single large outlier, namely observation 14 (`dose level` 6 and `block2`). On the other hand, this observation does not appear as a leverage point because its h_i value is small.

In the following we compare the classical and the robust analysis. The classical estimates are obtained by maximum likelihood. The robust estimates are based on

the Huber quasi-likelihood estimator defined by (7) with $w(\mathbf{x}_i) = 1$ for all i. The tuning constant of the Huber function is chosen to be 1.2, which is obtained by the procedure described at the end of Section 3.2 with $\alpha - \alpha_0 = 0.02$, $\epsilon = 0.04$ and $\kappa = 0.1145$.

[Table 1 about here.]

Table 1 shows the effect of observation 14: it seems to increase the value of β_2 corresponding to the variable `block2`. The robust technique automatically takes into account the particularity of observation 14: in the estimation procedure, most of the observations receive a weight equal to 1, or at least greater than 0.70, whereas observation 14 receives a weight equal to 0.26.

Also, the effect of observation 14 is clear on the value of the deviance. This seems dangerous because the deviance is used for assessing the significance of the variables used for modeling the response. This is confirmed by Table 2, where the results of a classical and robust stepwise procedure are compared.

[Table 2 about here.]

The classical analysis shows that all the variables, added sequentially, are highly significant on the basis of their deviance value. Model selection via a robust stepwise procedure based on the Huber quasi-deviance defined by equation (8) with $\nu(y_i, \mu_i) = \psi_c(r_i)/V^{1/2}(\mu_i)$ and $c = 1.2$ shows that the variable `block1` is not significant.

5.2 Poisson models

We use a dataset issued from a study of the diversity of arboreal marsupials in the Montane ash forest (Australia). This dataset was collected in view of the man-

agement of hardwood forest to take conservation and recreation values, as well as wood production, into account. The study is fully described in Lindenmayer et al. (1990, 1991). The number of different species of arboreal marsupials (possum) was observed on 151 different 3ha sites with uniform vegetation. For each site the following measures were recorded: number of shrubs, number of cut stumps from past logging operations, number of stags (hollow-bearing trees), a bark index reflecting the quantity of decorticating bark, a habitat score indicating the suitability of nesting and foraging habitat for Leadbeater's possum, the basal area of acacia species, the species of eucalypt with the greatest stand basal area (*Eucalyptus regnans*, *Eucalyptus delegatensis*, *Eucalyptus nitens*), and the aspect of the site. The problem is to model the relationship between diversity and these other variables.

Weisberg and Welsh (1993) used these data to investigate by nonparametric techniques the shape of the link function. Their conclusion was that the canonical link fits this dataset well. Therefore, we consider a Poisson generalized linear models with log-link to describe diversity as a function of

`shrubs` + `stumps` + `stags` + `bark` + `habitat` + `acacia` + `eucalyptus` + `aspect`,

where `eucalyptus` is a factor with three levels and `aspect` is a factor with four levels. Hence, the model involves the estimation of a parameter of dimension 12.

The robust estimation of parameters via a Mallows quasi-likelihood estimator defined by (7) with tuning constant $c = 1.6$ and weights $w(\mathbf{x}_i) = \sqrt{1 - h_i}$ gives the result of Table 3. In the same table, we report within parentheses the results obtained by means of classical quasi-likelihood. It has to be noticed that 4 observations, namely observations 59, 110, 133, 139, receive a weight with respect to their residual between 0.68 and 0.88. This shows that these 4 observations are potentially influential not only for the estimation procedure, but also for inference and model

selection. As one can see from Table 3, based on the asymptotic confidence intervals, many explanatory variables do not enter significantly in the model, and a reduction of the number of variables in the model is necessary.

[Table 3 about here.]

We applied a forward stepwise procedure based on quasi-likelihood and on the robust version of it. Starting from the null model where only the constant term is fitted, we tested whether it is appropriate to add the next explanatory variable. We chose to retain a variable if the p-value was smaller than 5%. Table 4 shows the p-value obtained at each step of the procedure. Bold p-values indicate the variables which have been retained in the model.

[Table 4 about here.]

As one can see from the table, the models chosen by the classical and the robust analysis are essentially the same, even if the p-values involved are sometimes quite different. The variable `habitat` is at the border of the decision rule and external consideration may be used to judge if it has to be kept in the model. It has to be noticed that the correlation between `habitat` and `acacia` is high (0.54) and one of these variables can be dropped.

[Table 5 about here.]

In the robust final fit, observations 59, 110, 133, 139 receive a weights with respect to their residuals between 0.68 and 0.86, as it was already the case in the full model. On the other hand, with respect to the influence of position, the only observations receiving a weight less than 0.9 is the first one. There were three other

observations which seemed to be potentially dangerous in the model containing the whole set of variables. Probably, this outlyingness was due to some explanatory variables, which were not retained in the final model.

For the final model as presented in Table 5, we investigate the sensitivity of Mallows quasi-likelihood tests compared to classical tests by considering the following procedure: we let the response of the observation receiving the lowest weight in the estimation of the final model, namely observation 110, span the range of values from 0 to 6. These values cover the range of the response in the sample. In each situation, we test the null hypothesis that the coefficient corresponding to the variable `habitat` is equal to 0. The p-values of these tests are represented in Figure 1.

[Figure 1 about here.]

The p-value of the robust test ($c = 1.6$) is stable, irrespective to the response value taken by observation 110. This p-value ranges from 2.6 to 3.3%. On the other hand, the p-value of the classical test ($c = \infty$), varies much more: from 2.3 to 6.5%, giving rise to a different model choice, if the decision rule is set at 5%. Moreover, by letting observation 110 take arbitrarily large values, the p-value of the robust test is bounded, whereas the p-value of the classical test continues to increase.

6 Conclusion

In this paper we proposed a natural class of robust testing procedures for generalized linear models. They are a valuable complement to classical techniques and are more reliable in the presence of outlying points and other deviations from the assumed model. Further research includes the extension of these procedures to generalized estimating equations and to nonparametric models like generalized additive models.

A Fisher consistency correction

We derive the constant

$$a(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\big[\psi_c\big(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\big)\big]\frac{1}{V^{1/2}(\mu_i)}\mu_i',$$

for binomial and Poisson models, which reduces to the computation of $\mathsf{E}\big[\psi_c\big(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\big)\big]$.

Let us define $j_1 = \lfloor \mu_i - cV^{1/2}(\mu_i)\rfloor$, and $j_2 = \lfloor \mu_i + cV^{1/2}(\mu_i)\rfloor$.

The binomial model states that $Y_i \sim \mathcal{B}(m_i, p_i)$, so that $\mathsf{E}[Y_i] = \mu_i = m_i p_i$ and $\mathsf{V}[Y_i] = \mu_i\frac{m_i-\mu_i}{m_i}$. Then we have

$$\begin{aligned}
\mathsf{E}\big[\psi_c\big(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\big)\big] &= \sum_{j=-\infty}^{\infty} \psi_c\big(\frac{j-\mu_i}{V^{1/2}(\mu_i)}\big)\mathsf{P}(Y_i=j)\mathbf{I}_{\{j\in[0,m_i]\}} \\
&= c\big(\mathsf{P}(Y_i \geq j_2+1) - \mathsf{P}(Y_i \leq j_1)\big) \\
&+ \frac{\mu_i}{V^{1/2}(\mu_i)}\big[\mathsf{P}(j_1 \leq \tilde{Y}_i \leq j_2-1) - \mathsf{P}(j_1+1 \leq Y_i \leq j_2)\big],
\end{aligned}$$

with $\tilde{Y}_i \sim \mathcal{B}(m_i-1, p_i)$.

The Poisson model states that $Y_i \sim \mathcal{P}(\mu_i)$, and hence $\mathsf{E}[Y_i] = V(\mu_i) = \mu_i$. Then,

$$\begin{aligned}
\mathsf{E}\big[\psi_c\big(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\big)\big] &= \sum_{j=-\infty}^{\infty} \psi_c\big(\frac{j-\mu_i}{V^{1/2}(\mu_i)}\big)\mathsf{P}(Y_i=j)\mathbf{I}_{\{j\geq 0\}} \\
&= c\big(\mathsf{P}(Y_i \geq j_2+1) - \mathsf{P}(Y_i \leq j_1)\big) + \frac{\mu_i}{V^{1/2}(\mu_i)}\big[\mathsf{P}(Y_i=j_1) - \mathsf{P}(Y_i=j_2)\big].
\end{aligned}$$

B Asymptotic variance

We first determine the matrix $Q(\boldsymbol{\psi}_c, F)$ in the particular situation of Mallows quasi-likelihood estimator. Using its definition, we have

$$\begin{aligned}
Q(\boldsymbol{\psi}_c, F) &= \mathsf{E}\big[\big(\psi_c(r)w(\mathbf{x})\frac{1}{V^{1/2}(\mu)}\mu' - a(\boldsymbol{\beta})\big)\big(\psi_c(r)w(\mathbf{x})\frac{1}{V^{1/2}(\mu)}\mu' - a(\boldsymbol{\beta})\big)^T\big] \\
&= \frac{1}{n}X^T A X - a(\boldsymbol{\beta})a(\boldsymbol{\beta})^T,
\end{aligned}$$

where A is the diagonal matrix with elements $a_i = \mathsf{E}[\psi_c(r_i)^2] w^2(\mathbf{x}_i) \frac{1}{V(\mu_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2$, since $\mu_i' = \big(\frac{\partial \mu_i}{\partial \eta_i}\big)\mathbf{x}_i$. In the same manner, writing $s(y, \mathbf{x}, \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log h(y_i|\mathbf{x}_i, \mu_i)$, we derive the expression of $M(\boldsymbol{\psi}_c, F)$,

$$\begin{aligned} M(\boldsymbol{\psi}_c, F) &= \mathsf{E}\big[\big(\psi_c(r) w(\mathbf{x}) \frac{1}{V^{1/2}(\mu)} \mu' - a(\boldsymbol{\beta})\big) s(y, \mathbf{x}, \boldsymbol{\beta})^T\big] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}\big[\psi_c(r_i) \frac{\partial}{\partial \mu_i} \log h(y_i|\mathbf{x}_i, \mu_i)\big] \frac{1}{V^{1/2}(\mu_i)} w(\mathbf{x}_i) \mu_i' \mu_i'^T \\ &= \frac{1}{n} X^T B X, \end{aligned}$$

where B is the diagonal matrix with elements $b_i = \mathsf{E}[\psi_c(r_i) \frac{\partial}{\partial \mu_i} \log h(y_i|\mathbf{x}_i, \mu_i)] \frac{1}{V^{1/2}(\mu_i)} w(\mathbf{x}_i) (\frac{\partial \mu_i}{\partial \eta_i})^2$.

So, the determination of the asymptotic variance of a Mallows quasi-likelihood estimator involves the computation of the diagonal terms of the matrices A and B.

We determine the three terms: $\frac{\partial}{\partial \eta_i} g^{-1}(\eta_i)$, $\mathsf{E}[\psi_c(r_i)^2]$, and $\mathsf{E}[\psi_c(r_i) \frac{\partial}{\partial \mu_i} \log h(y_i|\mathbf{x}_i, \mu_i)]$ for binomial and Poisson models.

For the binomial model with logit link

$$\frac{\partial}{\partial \eta_i} g^{-1}(\eta_i) = m_i \frac{\exp(\eta_i)}{(1 + \exp(\eta_i))^2},$$

and

$$\begin{aligned} \mathsf{E}\big[\psi_c^2\big(\frac{Y_i - \mu_i}{V^{1/2}(\mu_i)}\big)\big] = c^2\big(\mathsf{P}(Y \le j_1) + \mathsf{P}(Y \ge j_2 + 1)\big) + \\ + \frac{1}{V(\mu_i)}\Big[\pi_i^2 m_i (m_i - 1) \mathsf{P}(j_1 - 1 \le \tilde{\tilde{Y}} \le j_2 - 2) + \\ + (\mu_i - 2\mu_i^2) \mathsf{P}(j_1 \le \tilde{Y} \le j_2 - 1) + \\ + \mu_i^2 \mathsf{P}(j_1 + 1 \le Y \le j_2)\Big], \end{aligned}$$

with $Y \sim \mathcal{B}(m_i, \pi_i)$, $\tilde{Y} \sim \mathcal{B}(m_i - 1, \pi_i)$ and $\tilde{\tilde{Y}} \sim \mathcal{B}(m_i - 2, \pi_i)$.

$\frac{\partial}{\partial \mu_i} \log h(y_i|\mathbf{x}_i, \mu_i)$ being equal to $\frac{Y_i - \mu_i}{V(\mu_i)}$, we have

$$\begin{aligned}
\mathsf{E}\big[\psi_c(r_i)\frac{\partial}{\partial \mu_i}\log h(y_i|\mathbf{x}_i,\mu_i)\big] &= \mathsf{E}\big[\psi_c\big(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\big)\frac{Y_i-\mu_i}{V(\mu_i)}\big] = \\
&= \frac{c\mu_i}{V(\mu_i)}\Big[\mathsf{P}(Y_i \le j_1) - \mathsf{P}(\tilde{Y}_i \le j_1 - 1) + \mathsf{P}(\tilde{Y}_i \ge j_2) - \mathsf{P}(Y_i \ge j_2 + 1)\Big] + \\
&\quad + \frac{1}{V^{3/2}(\mu_i)}\Big[\pi_i^2 m_i(m_i-1)\mathsf{P}(j_1 - 1 \le \tilde{\tilde{Y}}_i \le j_2 - 2) \\
&\quad + (\mu_i - 2\mu_i^2)\mathsf{P}(j_1 \le \tilde{Y}_i \le j_2 - 1) + \mu_i^2 \mathsf{P}(j_1 + 1 \le Y_i \le j_2)\Big].
\end{aligned}$$

For the Poisson model, we use the log-link $\eta_i = g(\mu_i) = \log(\mu_i)$ which leads to $\frac{\partial}{\partial \eta_i} g^{-1}(\eta_i) = \exp(\eta_i)$. We also have

$$\begin{aligned}
\mathsf{E}\big[\psi_c^2\big(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\big)\big] &= c^2\big(\mathsf{P}(Y_i \le j_1) + \mathsf{P}(Y_i \ge j_2 + 1)\big) + \\
&+ \frac{1}{V(\mu)}\Big[\mu^2 \mathsf{P}(j_1 - 1 \le Y_i \le j_2 - 2) + (\mu - 2\mu^2)\mathsf{P}(j_1 \le Y_i \le j_2 - 1) \\
&+ \mu^2 \mathsf{P}(j_1 + 1 \le Y_i \le j_2)\Big].
\end{aligned}$$

The score function equals $\frac{\partial}{\partial \mu_i}\log h(y_i|\mathbf{x}_i,\mu_i) = \frac{Y_i-\mu_i}{\mu_i} = \frac{Y_i-\mu_i}{V(\mu_i)}$, so that

$$\begin{aligned}
\mathsf{E}\big[\psi_c(r_i)\frac{\partial}{\partial \mu_i}\log h(y_i|\mathbf{x}_i,\mu_i)\big] &= \mathsf{E}\big[\psi_c\big(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\big)\frac{Y_i-\mu_i}{V(\mu_i)}\big] = \\
&= c\big(\mathsf{P}(Y_i = j_1) + \mathsf{P}(Y_i = j_2)\big) + \\
&+ \frac{1}{V^{3/2}(\mu_i)}\mu_i^2\big(\mathsf{P}(Y_i = j_1 - 1) - \mathsf{P}(Y_i = j_1) - \mathsf{P}(Y_i = j_2 - 1) + \mathsf{P}(Y_i = j_2)\big) + \\
&+ \mu_i \mathsf{P}(j_1 \le Y_i \le j_2 - 1).
\end{aligned}$$

C Conditions for Robust Quasi-deviance Tests

[C1]: Denote by D_n the set of all sample points $\mathbf{z}_i$, $i = 1, \ldots, n$ for which the second-order derivatives $\partial^2 Q_M(\mathbf{z}_i, \boldsymbol{\beta})/\partial\beta_j \partial\beta_k$, $i = 1, \ldots, n$; $j, k = 1, \ldots, p$ exist and are continuous functions of $\boldsymbol{\beta}$. It is assumed that $\lim_{n\to\infty} \mathsf{P}_{\boldsymbol{\beta}}(D_n) = 1$.

[C2]: For any $\mathbf{z} \in D_n$, any positive value δ, and any $\boldsymbol{\beta}_1$ denote by $\eta_{jk}(\mathbf{z}, \boldsymbol{\beta}_1, \delta)$ the least upper bound and by $\gamma_{jk}(\mathbf{z}, \boldsymbol{\beta}_1, \delta)$ the greatest lower bound of $\partial^2 Q_M(\mathbf{z}, \boldsymbol{\beta})/\partial\beta_j \partial\beta_k$, with respect to $\boldsymbol{\beta}$ in the $\boldsymbol{\beta}$ interval $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}|| \leq \delta$.

Moreover, assume that for any sequence $\{\delta_n\}$ for which $\lim_{n\to\infty} \delta_n = 0$,

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}}\big[\eta_{jk}(\mathbf{z}, \boldsymbol{\beta}, \delta_n)\big] = \lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}}\big[\gamma_{jk}(\mathbf{z}, \boldsymbol{\beta}, \delta_n)\big] = \mathsf{E}_{\boldsymbol{\beta}}\big[\partial^2 Q_M(\mathbf{z}, \boldsymbol{\beta})/\partial\beta_j \partial\beta_k\big],$$

and that there exists a positive ϵ such that the expectations $\mathsf{E}_{\boldsymbol{\beta}}\big[\eta^2_{jk}(\mathbf{z}, \boldsymbol{\beta}, \delta)\big]$ and $\mathsf{E}_{\boldsymbol{\beta}}\big[\gamma^2_{jk}(\mathbf{z}, \boldsymbol{\beta}, \delta)\big]$ are bounded functions of $\boldsymbol{\beta}$ and δ for all $\boldsymbol{\beta}$ and $\delta < \epsilon$.

These conditions are obtained by replacing $\log f(z, \boldsymbol{\beta})$ by $Q_M(z, \boldsymbol{\beta})$ in the corresponding classical results for the likelihood ratio test; cf. Rao (1973), Wald (1943).

D Proof of Proposition 1

First, we derive the asymptotic equivalent quadratic form of Λ_{QM}. The proof follows the same lines as in the classical theory.

The first step of the proof consists in approximating Λ_{QM} under conditions [C1]-[C2] by

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}})^T M(\boldsymbol{\psi}, F) \sqrt{n}(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}), \tag{14}$$

via a Taylor expansion and by making use of Slutsky's theorem. Then, under H_0 and by the asymptotic properties of M-estimators which hold under conditions (A.1)-(A.9) of Heritier and Ronchetti (1994), the following distribution equality holds asymptotically

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}) \overset{\mathcal{D}}{\sim} \sqrt{n}\big(M^{-1}(\boldsymbol{\psi}, F) - \tilde{M}^{+}(\boldsymbol{\psi}, F)\big)\mathbf{L}_n, \tag{15}$$

where $\mathbf{L}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mu_i)$ is such that $\sqrt{n}\mathbf{L}_n \sim \mathcal{N}\big(\mathbf{0}, Q(\psi, F)\big)$. Putting (15) in (14), and taking into account the symmetry of $M(\boldsymbol{\psi}, F)$, we finally have, as $n \to \infty$,

$$\Lambda_{QM} \overset{\mathcal{D}}{\sim} n\mathbf{L}_n^T C(\boldsymbol{\psi}, F)\mathbf{L}_n. \tag{16}$$

(16) can be rewritten as

$$\Lambda_{QM} \overset{\mathcal{D}}{\sim} n\mathbf{R}_{n(2)}^T M(\boldsymbol{\psi}, F)_{22.1}\mathbf{R}_{n(2)},$$

where $M(\boldsymbol{\psi}, F)_{22.1} = M(\boldsymbol{\psi}, F)_{22} - M(\boldsymbol{\psi}, F)_{12}M(\boldsymbol{\psi}, F)_{11}^{-1}M(\boldsymbol{\psi}, F)_{12}$, and $\sqrt{n}\mathbf{R}_n$ is distributed according to $\mathcal{N}\big(\mathbf{0}, M^{-1}(\boldsymbol{\psi}, F)Q(\boldsymbol{\psi}, F)M^{-1}(\boldsymbol{\psi}, F)\big)$.

Finally, from (16) we conclude that

$$\Lambda_{QM} \sim \sum_{i=1}^{q} d_i N_i^2,$$

where d_i are the q positive eigenvalues of $Q(\boldsymbol{\psi}, F)C(\boldsymbol{\psi}, F)$ and $N_1, \ldots, N_q$ are independent standard normal variables. Thus, the distribution of Λ_{QM} is a linear combination of χ^2 random variables with 1 degree of freedom.

E Proof of Proposition 2

By using (11) and by standard results on the distribution of quadratic forms in normal variables, we can say that the statistic $n\mathbf{U}_n^T A\mathbf{U}_n$ is asymptotically distributed as $\sum_{i=1}^{q} d_i\chi_1^2(\xi_i^2)$, with $\boldsymbol{\xi}(\epsilon) = (\xi_1(\epsilon), \ldots, \xi_q(\epsilon))^T = \sqrt{n}P\mathbf{U}(F_{\epsilon,n})$. Notice that the distribution depends only on the $\xi_i^2(\epsilon)$ (see Johnson and Kotz (1970), Chapter 29).

Moreover, up to $O(1/n)$, we have that $\alpha(F_{\epsilon,n}) = 1 - H_{d_1,\ldots,d_q}(\eta_{1-\alpha_0}; \boldsymbol{\lambda}(\epsilon))$, with $\boldsymbol{\lambda}(\epsilon) = \mathrm{diag}(\boldsymbol{\xi}(\epsilon)\boldsymbol{\xi}(\epsilon)^T) = n\,\mathrm{diag}\left(P\mathbf{U}(F_{\epsilon,n})\mathbf{U}(F_{\epsilon,n})^T P^T\right)$.

Let $b(\epsilon) = -H_{d_1,\ldots,d_q}(\eta_{1-\alpha_0}; \boldsymbol{\lambda}(\epsilon))$. Then, up to $O(1/n)$, we have

$$\alpha(F_{\epsilon,n}) - \alpha_0 = b(\epsilon) - b(0) = \epsilon b'(0) + \frac{1}{2}\epsilon^2 b''(0) + o(\epsilon^2).$$

But

$$b'(0) = \boldsymbol{\kappa}^T \cdot \frac{\partial}{\partial \epsilon}\boldsymbol{\lambda}\Big|_{\epsilon=0} = 2n\boldsymbol{\kappa}^T \cdot \mathrm{diag}\left(P\Big[\frac{\partial}{\partial \epsilon}\mathbf{U}(F_{\epsilon,n})\Big]_{\epsilon=0}\mathbf{U}(F_{\boldsymbol{\beta}_0})P^T\right) = 0,$$

because $\mathbf{U}(F_{\boldsymbol{\beta}_0}) = 0$.

We also have that

$$\begin{aligned} b''(0) &= \boldsymbol{\kappa}^T \cdot \frac{\partial^2}{\partial \epsilon^2}\boldsymbol{\lambda}\Big|_{\epsilon=0} = \boldsymbol{\kappa}^T \cdot 2n\,\mathrm{diag}\left(P\Big[\frac{\partial}{\partial \epsilon}\mathbf{U}(F_{\epsilon,n})\frac{\partial}{\partial \epsilon}\mathbf{U}(F_{\epsilon,n})^T\Big]_{\epsilon=0}P^T\right) \\ &= 2\boldsymbol{\kappa}^T \cdot \mathrm{diag}\left(P\left(\int \mathsf{IF}(\mathbf{z};\mathbf{U},F_{\boldsymbol{\beta}_0})dG(\mathbf{z})\right)\left(\int \mathsf{IF}(\mathbf{z};\mathbf{U},F_{\boldsymbol{\beta}_0})dG(\mathbf{z})\right)^T P^T\right), \end{aligned}$$

by using again the fact that $\mathbf{U}(F_{\boldsymbol{\beta}_0}) = 0$ and because $\frac{\partial}{\partial \epsilon}\mathbf{U}(F_{\epsilon,n})\big|_{\epsilon=0} = \int \mathsf{IF}(\mathbf{z};\mathbf{U},F_{\boldsymbol{\beta}_0})\frac{1}{\sqrt{n}}dG(\mathbf{z})$ (see Hampel et al., 1986, p. 83). This completes the proof.

Acknowledgments

Eva Cantoni (`Eva.Cantoni@metri.unige.ch`) is *maître-assistante* and Elvezio Ronchetti (`Elvezio.Ronchetti@metri.unige.ch`) is Professor, Department of Econometrics, CH-1211 Geneva 4, Switzerland. The authors would like to thank the Editor, the Associate Editor, the referees, P. Rousseeuw and M.-P. Victoria-Feser for their valuable suggestions, and A. Welsh for providing the data analyzed in Section 5.2. The hospitality of Stanford (E. C.) and MIT (E. R.) during the revision of the manuscript is gratefully acknowledged.

References

Bednarski, T. (1993), "Fréchet Differentiability of Statistical Functionals and Implications to Robust Statistics," in *New Directions in Statistical Data Analysis and Robustness*, eds. Morgenthaler, S., Ronchetti, E., and Stahel, W. A., Basel/Cambridge, MA: Birkhaeuser, pp. 25–34.

Carroll, R. J. and Welsh, A. H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician*, 42, 285–287.

Clarke, B. R. (1986), "Nonsmooth Analysis and Frechet Differentiability of M-functionals," *Probability Theory and Related Fields*, 73, 197–209.

Davies, R. B. (1980), "[Algorithm AS 155] The Distribution of a Linear Combination of χ^2 Random Variables (AS R53: 84V33 P366- 369)," *Applied Statistics*, 29, 323–333.

Farebrother, R. W. (1990), "[Algorithm AS 256] The Distribution of a Quadratic Form in Normal Variables," *Applied Statistics*, 39, 294–309.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

Hanfelt, J. J. and Liang, K.-Y. (1995), "Approximate Likelihood Ratios for General Estimating Functions," *Biometrika*, 82, 461–477.

He, X. and Simpson, D. G. (1993), "Lower Bounds for Contamination Bias: Globally Minimax Versus Locally Linear Estimation," *The Annals of Statistics*, 21, 314–337.

He, X., Simpson, D. G., and Portnoy, S. L. (1990), "Breakdown Robustness of Tests," *Journal of the American Statistical Association*, 85, 446–452.

Heritier, S. and Ronchetti, E. (1994), "Robust Bounded-influence Tests in General Parametric Models," *Journal of the American Statistical Association*, 89, 897–904.

Heyde, C. C. (1997), *Quasi-Likelihood and its Application*, Berlin/New York: Springer-Verlag.

Huber, P. J. (1967), "The Behavior of the Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233.

— (1981), *Robust Statistics*, New York: Wiley.

Imhof, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 352–363.

Johnson, N. L. and Kotz, S. (1970), *Continuous Univariate Distributions*, vol. 2, Boston; Geneva, IL: Houghton-Mifflin.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (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.

Lindenmayer, D. B., Cunningham, R. B., Tanton, M. T., Nix, H. A., and Smith, A. P. (1991), "The Conservation of Arboreal Marsupials in the Montane Ash Forests of the Central Highlands of Victoria, South-East Australia: III. The Habitat Requirements of Leadbeater's Possum *G*ymnobelideus leadbeateri and Models of the Diversity and Abundance of Arboreal Marsupials," *Biological Conservation*, 56, 295–315.

Lindenmayer, D. B., Cunningham, R. B., Tanton, M. T., Smith, A. P., and Nix, H. A. (1990), "The Conservation of Arboreal Marsupials in the Montane Ash Forests of the Victoria, South-East Australia, I. Factors Influencing the Occupancy of Trees with Hollows," *Biological Conservation*, 54, 111–131.

Markatou, M., Basu, A., and Lindsay, B. G. (1998), "Weighted Likelihood Equations With Bootstrap Root Search," *Journal of the American Statistical Association*, 93, 740–750.

Markatou, M. and He, X. (1994), "Bounded Influence and High Breakdown Point Testing Procedures in Linear Models," *Journal of the American Statistical Association*, 89, 543–549.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall, 2nd ed.

Morgenthaler, S. (1992), "Least-absolute-deviations Fits for Generalized Linear Models," *Biometrika*, 79, 747–754.

Pearson, E. S. (1959), "Note on the Distribution of Non-central χ^2," *Biometrika*, 46, 364.

Phelps, K. (1982), "Use of the Complementary log-log Function to Describe Dose-response Relationships in Insecticide Evaluation Field Trials," in *Lecture Notes in Statistics, No. 14. GLIM.82: Proceedings of the International Conference on Generalized Linear Models*, ed. Gilchrist, R., Berlin/New York: Springer-Verlag.

Pregibon, D. (1982), "Resistant Fits for Some Commonly Used Logistic Models With Medical Applications," *Biometrics*, 38, 485–498.

Preisser, J. S. and Qaqish, B. F. (1999), "Robust Regression for Clustered Data with Applications to Binary Regression," *Biometrics*, 55, 574–579.

Rao, C. R. (1973), *Linear Statistical Inference*, New York: Wiley, 2nd ed.

Ronchetti, E. (1997), "Robustness Aspects of Model Choice," *Statistica Sinica*, 7, 327–338.

Ronchetti, E. and Staudte, R. G. (1994), "A Robust Version of Mallows' C_p," *Journal of the American Statistical Association*, 89, 550–559.

Ronchetti, E. and Trojani, F. (2001), "Robust Inference with GMM Estimators," *Journal of Econometrics*, 101, 36–79.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Ruckstuhl, A. F. and Welsh, A. H. (1999), "Robust Fitting of the Binomial Model," Manuscript.

Sommer, S. and Huggins, R. (1996), "Variable Selection using the Wald Test and a Robust C_p," *Applied Statistics*, 45, 15–29.

Staudte, R. G. and Sheather, S. J. (1990), *Robust Estimation and Testing*, New York: Wiley.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986), "Optimally Bounded Score Functions for Generalized Linear Models With Applications to Logistic Regression," *Biometrika*, 73, 413–424.

Wald, A. (1943), "Test for Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large." *Transactions of the American Mathematical Society*, 54, 426–482.

Wedderburn, R. W. M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61, 439–447.

Weisberg, S. and Welsh, A. H. (1993), "Missing Links," in *New Directions in Statistical Data Analysis and Robustness*, eds. Morgenthaler, S., Ronchetti, E., and Stahel, W. A., Basel/Cambridge, MA: Birkhaeuser, pp. 275–284.

Williams, D. A. (1987), "Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions," *Applied Statistics*, 36, 181–191.

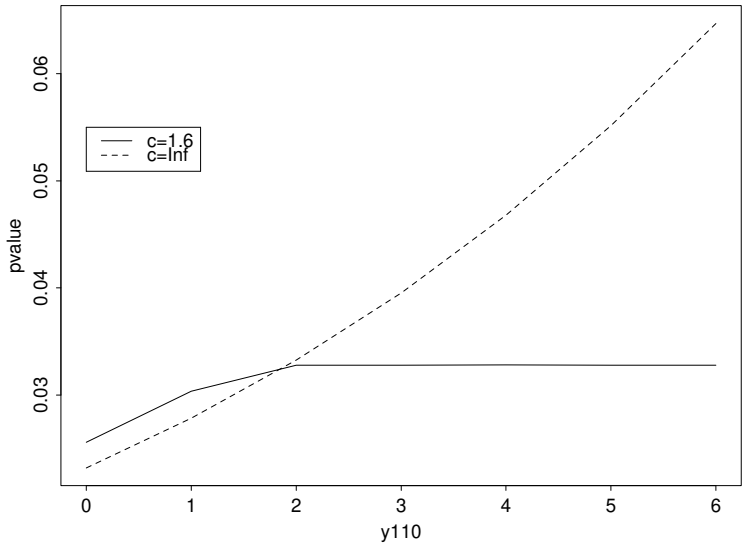

Figure 1: Sensitivity curves of the p-value for Mallows quasi-likelihood tests with $c = 1.6$ (solid line) and $c = \infty$ (dashed line).

	Max. likelihood	Huber quasi-likelihood
Intercept	1.480 (0.66)	1.939 (0.70)
logdose	-1.817 (0.34)	-2.049 (0.37)
block2	0.843 (0.23)	0.685 (0.24)
block1	0.542 (0.23)	0.450 (0.24)

Table 1: Estimation of $\boldsymbol{\beta}$ by maximum likelihood and by the Huber quasi-likelihood estimator with $c = 1.2$. Standard errors are indicated within parentheses.

	Resid. Deviance	Resid. Huber quasi-deviance
NULL	83.34	60.46
logdose	54.73 (0.000)	39.94 (0.000)
block2	45.59 (0.003)	35.21 (0.017)
block1	39.98 (0.018)	32.74 (0.085)

Table 2: Residual deviance and residual Huber quasi-deviance with $c = 1.2$. p-values are indicated within parentheses.

Variable	Coefficient		Standard Error	
Intercept	-0.8978	(-0.947)	0.2682	(0.265)
`shrubs`	0.0099	(0.012)	0.0222	(0.022)
`stumps`	-0.2514	(-0.272)	0.2876	(0.286)
`stags`	0.0402	(0.040)	0.0113	(0.011)
`bark`	0.0400	(0.040)	0.0145	(0.014)
`acacia`	0.0178	(0.018)	0.0107	(0.011)
`habitat`	0.0714	(0.072)	0.0385	(0.038)
`eucalyptus nitens`	0	(0)	-	(-)
`eucalyptus regnans`	-0.020	(-0.015)	0.1938	(0.192)
`eucalyptus delegatensis`	0.127	(0.115)	0.2738	(0.272)
`aspect NW-NE`	0	(0)	-	(-)
`aspect NW-SE`	0.0601	(0.067)	0.1913	(0.190)
`aspect SE-SW`	0.0949	(0.117)	0.1920	(0.190)
`aspect SW-NW`	-0.5079	(-0.489)	0.2505	(0.247)

Table 3: Coefficients estimation and corresponding standard errors for the Poisson model with log-link of the possum dataset.

	Classical QL	Robust QL
shrubs	0.0871	0.3642
stumps	0.0646	0.2988
stags	**0.0000**	**0.0000**
bark	**0.0035**	**0.0039**
acacia	**0.0002**	**0.0009**
habitat	**0.0500**	**0.0443**
eucalyptus regnans	0.8754	0.8030
eucalyptus delegatensis	0.8591	0.8074
eucalyptus nitens	0.5681	0.6461
aspect NW-NE	0.8336	0.9100
aspect NW-SE	0.2612	0.3462
aspect SE-SW	0.1996	0.1646
aspect SW-NW	**0.0012**	**0.0023**

Table 4: p-values of a forward stepwise procedure for the Poisson model with log-link of the possum dataset.

Variable	Coefficient		Standard Error	
Intercept	-0.7981	(-0.8213)	0.2030	(0.2000)
`stags`	0.0406	(0.0410)	0.0104	(0.0103)
`bark`	0.0410	(0.0406)	0.0126	(0.0125)
`habitat`	0.0143	(0.0136)	0.0098	(0.0097)
`acacia`	0.0776	(0.0782)	0.0371	(0.0367)
`aspect SW-NW`	-0.6044	(-0.5968)	0.2121	(0.2086)

Table 5: Coefficients estimation and corresponding standard errors for the final Poisson model with log-link of the possum dataset.