Robust Inference Based on Quasi-likelihoods for Generalized Linear Models and Longitudinal Data

Eva Cantoni

No 2001.02

Cahiers du département d’économétrie
Faculté des sciences économiques et sociales
Université de Genève

Avril 2001

Département d’économétrie
Université de Genève, 40 Boulevard du Pont-d’Arve, CH-1211 Genève 4
http://www.unige.ch/ses/metri/
Robust Inference Based on Quasi-likelihoods for Generalized Linear Models and Longitudinal Data

Eva Cantoni
Econometrics Department
University of Geneva
1211 GENEVA 4
Switzerland
Eva.Cantoni@metri.unige.ch

April 2001

Abstract

In this paper we introduce and develop robust versions of quasi-likelihood functions for model selection via an analysis-of-deviance type of procedure in generalized linear models and longitudinal data analysis. These robust functions are built upon natural classes of robust estimators and can be seen as weighted versions of their classical counterparts. The asymptotic theory of these test statistics is studied and their robustness properties are assessed for both generalized linear models and longitudinal data analysis. The proposed class of test statistics yields reliable inference even under model contamination. The analysis of a real data set completes the article.

KEY WORDS: Robust inference; quasi-likelihood functions; estimating equations; generalized linear models; longitudinal data.
Quasi-likelihood functions (Wedderburn, 1974, McCullagh, 1983) are flexible tools that can be used for estimation, inference and model selection. They have proved useful in generalized linear model (McCullagh and Nelder, 1989), where they are valid alternative to the likelihood approach, and for longitudinal (or panel) data analysis (Diggle, Liang, and Zeger, 1996), where the lack of likelihood methods for the analysis of multivariate non normal data is a serious problem. But these procedures based on quasi-likelihoods are sensitive to model assumptions: parameter estimation can not only be highly influenced by the presence of unusual data point, but also inference suffers from lack of robustness to influential observations. Fortunately, quasi-likelihood functions have another advantage: they allow to build robust versions. This allows to develop robust estimation, but – more important – robust inference for large classes of models, e.g. generalized linear models.

Consider a setting where $Y_{it}$ is the discrete or continuous outcome for subject $i$ at time $t$, for $i = 1, \ldots, K$ and $t = 1, \ldots, n_i$. For each outcome $Y_{it}$, we also measure a set of covariates $x_{it}$. We write $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$ for the $n_i \times 1$ vector of responses, and $X_i = (x_{i1} \ldots x_{in_i})^T$ for the $n_i \times p$ matrix of covariates of subject $i$. We suppose that $\text{Corr}(Y_i) = A_i^{-1} \text{Var}(Y_i) A_i^{-1}$, with $A_i = \text{diag}(v^{1/2}(\mu_{i1}), \ldots, v^{1/2}(\mu_{in_i}))$, and that the subjects are independent. Purely dependent data are obtained with $K = 1$ (only one cluster) and purely independent data are obtained with $n_i = 1$ for all $i$ (generalized linear models). We model the marginal mean $E(Y_{it}) = \mu_{it}$, and assume that $g(\mu_{it}) = x_{it}^T \beta = \eta_{it}$ for a link function $g$, and that $\text{V}(Y_{it}) = \phi v(\mu_{it})$.

Estimation of the regression parameters in generalized linear models is usually performed by maximum likelihood or by maximum quasi-likelihood. The non-robustness of these estimators has been studied extensively and alternatives has been proposed, see e.g. Pregibon (1982), Stefanski, Carroll, and Ruppert (1986), Künsch, Stefanski, and Carroll (1989), Morgenthaler (1992), and Ruckstuhl and Welsh (1999). In the marginal approach to longitudinal data analysis, the regression parameters are usually estimated via the GEE approach (Liang and Zeger, 1986, Zeger and Liang, 1986), whereas the nuisance parameters are estimated by the method of moments. The lack of robustness of the GEE approach has been discussed by Preisser and Qaqish (1999) and Qaqish and Preisser (1999), who proposed a class of robust estimators for longitudinal data analysis.

Inference and model selection for generalized linear models is in general carried out based on the notion of deviance. Unfortunately this procedure lacks of robustness. We develop in Section 2 a robust approach to inference for generalized linear models based on robust deviances which are natural.
generalizations of classical quasi-likelihood functions. This approach has the advantage to preserve the structure of model building through the analysis of deviance. The same attractive procedure for model building through the analysis of deviance can be applied to longitudinal data analysis. Taking into account the correlation between observations of the same class, we construct in Section 3 a robust class of test statistics based on quasi-likelihood to make inference for panel data. In Section 4 we present an application to a real dataset. It then follows the appendix and the references.

2 Robust Inference for Generalized Linear Models

Let us first consider the case of purely independent data, that is \( n_i = 1 \) for all \( i \), \( Y_i \) is scalar, and \( X_i \) is a vector of dimension \( 1 \times p \). Note that in this section we suppress the useless second subscript on the notation of the variables.

We build upon the following set of estimating equations

\[
\sum_{i=1}^{K} \left[ \nu(y_i, \mu_i)w(X_i)\mu_i' - a(\beta) \right] = \sum_{i=1}^{K} \psi(y_i, \mu_i) = 0, \tag{1}
\]

where \( a(\beta) = K^{-1} \sum_{i=1}^{K} E[\nu(y_i, \mu_i)]w(X_i)\mu_i' \) with the expectation taken with respect to the conditional distribution of \( y|X \), and \( \nu(\cdot, \cdot) \), \( w(X) \) are weight functions. The constant \( a(\beta) \) ensures the Fisher consistency of the estimator.

Equation (1) defines an M-estimator (Huber, 1964) which inherits the properties of this class: the asymptotic normality with asymptotic variance

\[
\Omega = M(\psi, F)^{-1}Q(\psi, F)M(\psi, F)^{-1},
\]

where \( M(\psi, F) = -E[\frac{\partial}{\partial \beta} \psi(y, \mu)] \) and \( Q(\psi, F) = E[\psi(y, \mu)\psi(y, \mu)^T] \), and the influence function (Hampel, 1974) defined by

\[
IF(y; \psi, F) = M(\psi, F)^{-1} \psi(y, \mu),
\]

see Hampel, Ronchetti, Rousseeuw, and Stahel (1986).

The functions \( \nu \) and \( w \) in (1) are weight functions introduced to control outliers and leverage points: a bounded function \( \nu(y, \cdot) \) induces a bounded influence function with respect to \( y \), and \( w(X) \) downweights atypical observations in the design. The particular choice

\[
\nu(y_i, \mu_i) = \psi_c(r_i) \frac{1}{\psi^{1/2}(\mu_i)}, \tag{2}
\]
where \( r_i = (y_i - \mu_i) / v^{1/2}(\mu_i) \) and \( \psi_c(t) = \min(c, \max(t, -c)) \) is the Huber function, defines what we call a Mallows quasi-likelihood estimator, studied in details in Cantoni (1999) and Cantoni and Ronchetti (2001) for binomial and Poisson models. Simple choices for \( w(X) \) can be based on the hat matrix \( H \), for example \( w(X) = \sqrt{1 - h_i} \), where \( h_i \) is the \( i \)-th diagonal element of \( H \).

More sophisticated choices, in particular with high breakdown point, can be defined as the inverse of the Mahalanobis distance defined through a high breakdown estimate of the center and the covariance matrix of the \( X_i \) (see, for example, Rousseeuw and Leroy, 1987, p. 258 ff.). Finally notice that the choice of \( \nu(y_i, \mu_i) = (y_i - \mu_i) / v(\mu_i) \) and \( w(x_i) = 1 \) for all \( i \), recovers the classical quasi-likelihood estimator.

The estimating equations discussed above correspond to the minimization with respect to \( \beta \) of the quantity \( \sum_{i=1}^{K} Q_M(y_i, \mu_i) \), where

\[
Q_M(y_i, \mu_i) = \int_{\hat{s}}^{\mu_i} \nu(y_i, t)w(X_i)dt - \frac{1}{K} \sum_{j=1}^{K} \int_{\hat{\mu}}^{\mu_i} E[\nu(y_j, t)w(X_j)]dt,
\]

with \( \hat{s} \) such that \( \nu(y_i, \hat{s}) = 0 \), and \( \hat{\mu} \) such that \( E[\nu(y_i, \hat{\mu})] = 0 \). Note that differences of quasi-likelihoods, as in (4), are independent of \( \hat{s} \) and \( \hat{\mu} \).

Denote by \( a = (a_1^T, a_2^T)^T \) the partition of a vector of dimension \( p \) into \( (p-q) \) and \( q \) components, and consider the partition of matrices accordingly. Let \( M_p \) be a model with \( p \) regression parameters and \( M_{p-q} \) a submodel with only \( (p-q) \) parameters. The corresponding set of parameter is \( \beta = (\beta_1, \ldots, \beta_p) = (\beta_1^T, \beta_2^T)^T \) for model \( M_p \) and \( (\beta_1^T, 0^T)^T \) for model \( M_{p-q} \).

Based on the function \( Q_M(y_i, \mu_i) \) defined by (3), we build the following test statistic to test the null hypothesis that \( H_0 : \beta_2 = 0 \), that is the null hypothesis that the submodel describes the data sufficiently well:

\[
\Lambda_{QM} = 2 \left[ \sum_{i=1}^{K} Q_M(y_i, \hat{\mu}_i) - \sum_{i=1}^{K} Q_M(y_i, \hat{\mu}_i) \right],
\]

where \( \hat{\mu}_i \) and \( \hat{\mu}_i \) are the estimates under model \( M_p \) and \( M_{p-q} \), associated with \( \beta \) and \( \beta \) respectively.

Statistic (4) defines a measure of discrepancy between models \( M_p \) and \( M_{p-q} \). It is a generalization of the quasi-deviance test for generalized linear models, which is recovered by taking \( Q_M(y_i, \mu_i) = \int_{y_i}^{\mu_i} (y_i - t) / v(t)dt \). Moreover, for identity link functions \( g \), (4) is the \( \tau \)-test statistic defined in Hampel et al. (1986), Chapter 7.

The next proposition establishes the asymptotic distribution of the test statistic (4).
Proposition 1 Under conditions (A.1)-(A.9) in Heritier and Ronchetti (1994), \([C1], [C2]\) in Cantoni and Ronchetti (2001), and under \(H_0: \beta(2) = 0\), the test statistic \(\Lambda_{QM}\) defined by (4) equals

\[
n\mathbf{L}_n^T C(\psi, F) \mathbf{L}_n + o_p(1) = n\mathbf{R}_{n(2)}^T M(\psi, F)_{22.1} \mathbf{R}_{n(2)} + o_p(1),
\]

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

Moreover, \(\Lambda_{QM}\) 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 \(Q(\psi, F)(M^{-1}(\psi, F) - \tilde{M}^+(\psi, F))\), and \(\tilde{M}^+(\psi, F)\) is such that \(\tilde{M}^+(\psi, F)_{11} = M(\psi, F)_{11}^{-1}\) and \(\tilde{M}^+(\psi, F)_{12} = 0\), \(\tilde{M}^+(\psi, F)_{21} = 0\), \(\tilde{M}^+(\psi, F)_{22} = 0\).

The proof of this proposition can be found in Cantoni and Ronchetti (2001). A similar result can be obtained for the distribution of \(\Lambda_{QM}\) under contiguous alternatives \(\beta(2) = n^{-1/2}\Delta\). In such a case \(\Lambda_{QM}\) is asymptotically distributed as

\[
\sum_{i=1}^q (d_i^{1/2} N_i + S^T \Delta)^2,
\]

where \(D\) is the diagonal matrix with elements \(d_1, \ldots, d_q\), \(S\) is such that \(SS^T = M_{22.1}\) and \(S^T(M^{-1}(\psi, F_{\beta_0})Q(\psi, F_{\beta_0})M^{-1}(\psi, F_{\beta_0}))_{22} S = D\).

2.1 Robustness properties and tuning constant selection

The robustness properties of the test based on (4) can be investigated by showing that a small amount of contamination at a point \(z\) 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 investigate the asymptotic level of the test statistic (4) under the sequence of \(\epsilon\)-contamination \(F_\epsilon,n = (1 - \frac{\epsilon}{\sqrt{n}})F_{\beta_0} + \frac{\epsilon}{\sqrt{n}} G\), with \(G\) an arbitrary distribution.
Proposition 2 Consider a parametric model $F_{\beta_0}$ and the null hypothesis $H_0 : \beta(2) = 0$. Denote by $F^{(n)}$ the empirical distribution and by $U_n$ the functional $U(F^{(n)})$ such that $U(F_{\beta_0}) = 0$, $IF(z; U, F_{\beta_0})$ is bounded and $\sqrt{n}(U_n - U(F^{(n)})) \sim \mathcal{N}(0, \Sigma)$ uniformly over the $\epsilon$-contamination $F_{\epsilon,n}$. Let $\alpha(F)$ be the level of the test based on the quadratic form $nU_n^T A U_n$ when the underlying distribution is $F$. The nominal level is $\alpha(F_{\beta_0}) = \alpha_0$.

Then, under the $\epsilon$-contamination $F_{\epsilon,n}$, we have
\[
\lim_{n \to \infty} \alpha(F_{\epsilon,n}) = \alpha_0 + \epsilon^2 \kappa^T \cdot \text{diag} \left( P \left( \int IF(z; U, F_{\beta_0}) dG(z) \right) \left( \int IF(z; U, F_{\beta_0}) dG(z) \right)^T P^T \right) + o(\epsilon^2),
\]
where $\kappa = -\frac{\partial}{\partial \lambda} H_{d_1, \ldots, d_q}(\eta_{1-\alpha_0}; \lambda)|_{\lambda = 0}$, $\lambda = (\lambda_1, \ldots, \lambda_q)^T = (\xi_1^2, \ldots, \xi_q^2)^T$, $H_{d_1, \ldots, d_q}(:, \lambda)$ 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 $\Sigma A$. Moreover, $\text{diag}(R)$ indicates the vector with components the diagonal elements of the matrix $R$.

The proof of this result is given in Cantoni and Ronchetti (2001). If the influence function of the functional $U$ is bounded, then the asymptotic level under contamination is also bounded. A similar result can be obtained for the power, showing that the asymptotic power is stable under contamination.

The function $\nu(y, \mu)$ is often tuned by a constant. One can either choose the tuning constant that ensures a certain level of asymptotic efficiency, or, as suggested in Ronchetti and Trojani (2001), choose the constant that controls the maximal bias on the asymptotic level of the test in a neighborhood of the model. The following corollary is needed.

Corollary 1 Under conditions (A.1)-(A.9) in Heritier and Ronchetti (1994), for any $M$-estimator $\hat{\beta}(2)$ with bounded influence function, the asymptotic level of the robust quasi-likelihood test statistic (4) under a point mass contamination is given by
\[
\lim_{n \to \infty} \alpha(F_{\epsilon,n}) = \alpha_0 +
\epsilon^2 \kappa^T \cdot \text{diag} \left( P \left( IF(z; \hat{\beta}(2), F_{\beta_0}) IF(z; \hat{\beta}(2), F_{\beta_0})^T P^T \right) \right) + o(\epsilon^2),
\]
where $P$ is an orthogonal matrix such that $P^T D P = \Omega_{22} M_{22,1}$, $\Omega$ is the asymptotic variance of $\hat{\beta}$, and $D$ is the diagonal matrix with elements $d_1, \ldots, d_q$ defined in Proposition 1.
This corollary is obtained by applying Proposition 2 with \( G(z) = \Delta z \), \( U = \beta(2) \), \( \Sigma = \Omega_{22} \), \( A = M_{22,1} \), and by using the Fréchet differentiability of \( \hat{\beta}(2) \); see Heritier and Ronchetti (1994). It proves that a bounded influence M-estimator \( \hat{\beta}(2) \) ensures a bound on the asymptotic level of the robust quasi-likelihood test under contamination.

The maximal level \( \alpha \) of the robust quasi-likelihood test statistic in a neighborhood of the model of radius \( \epsilon \) is given by

\[
\alpha = \alpha_0 + \epsilon^2 \gamma(\hat{\beta}(2), F_{\beta_0})^2 \kappa^T \text{diag} \left( P11^TP^T \right),
\]

where \( \gamma(\hat{\beta}(2), F_{\beta_0}) = \sup_z ||IF(z; \hat{\beta}(2), F_{\beta_0})|| \) and \( 1 = (1, \ldots, 1)^T \), which gives

\[
b = \frac{1}{\epsilon} \sqrt{\frac{\alpha - \alpha_0}{\kappa^T \text{diag}(P11^TP^T)}},
\]

where \( b \) is the bound on the influence function of the estimator \( \hat{\beta}(2) \). Then, for a fixed amount of contamination \( \epsilon \) 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 \( b = \gamma(\hat{\beta}(2), F_{\beta_0}) = \gamma_c \) with respect to \( c \). For example, if \( q = 1 \) we have \( P = 1 \), \( \text{diag}(P11^TP^T) = 1 \), and \( \kappa = 0.1145 \), see Ronchetti and Trojani (2001). In practice, the supremum on \( z = (y, x) \) is taken as the maximum over the sample of the supremum on \( y|x \). Note also that the solution depends on the unknown parameter \( \beta_0 \); our experience shows that it does not vary much for different values of \( \beta \), so that one can safely plug-in a reasonable (robust) estimate. This is valid for a single test. However, in a stepwise procedure one would have to choose a different value of \( c \) for each test. As a practical rule, we suggest to choose a global value of \( c \) by solving \( b = \sup_z ||IF(z; \hat{\beta}, F_{\beta_0})|| \), based on the fact that \( \gamma(\hat{\beta}(2), F_{\beta_0}) = \sup_z ||IF(z; \hat{\beta}(2), F_{\beta_0})|| \leq || \sup_z IF(z; \hat{\beta}, F_{\beta_0})|| \).

### 3 Robust Inference for Longitudinal Data

The lack of independence present in longitudinal must be taken into account. This is done by modeling the correlation structure either parametrically or nonparametrically. Beside this, marginal models for longitudinal data fit in a similar framework as generalized linear models with a linear predictor related to the marginal mean through a link function. Therefore, differences of quasi-likelihood can be used for model selection as in generalized linear models. As a starting point we use the set of estimating equations, suggested
by the general theory of optimal estimating functions, see Hanfelt and Liang (1995):

\[ \sum_{i=1}^{K} D_i^T \Gamma_i^T V_i^{-1} (\psi_i - c_i) = 0, \]  

(9)

where \( D_i = D_i(X_i, \beta) = \frac{\partial \mu_i}{\partial \beta} \) is a \( n_i \times p \) matrix, \( V_i = V_i(\mu_i, \alpha) = A_i R_i(\alpha) A_i \) is a \( n_i \times n_i \) matrix. \( R_i(\alpha) \), for an s-parameter \( \alpha \), is said to be the “working” correlation matrix, as opposed to the “true” correlation matrix \( \text{Corr}(Y_i) = A_i^{-1} \text{Var}(Y_i) A_i^{-1} \). Moreover, \( \psi_i = W_i \cdot (Y_i - \mu_i) \), where \( W_i = W_i(X_i, y_i, \mu_i) \) is a diagonal \( n_i \times n_i \) weight matrix containing robustness weights \( w_{it} \) for \( t = 1, \ldots, n_i \), and \( c_i = E(\psi_i) \). Finally, \( \Gamma_i = E(\tilde{\psi}_i - \tilde{c}_i) \) with \( \tilde{\psi}_i = \partial \psi_i / \partial \mu_i \) and \( \tilde{c}_i = \partial c_i / \partial \mu_i \).

The weights \( W_i \), which may also depend on \( \alpha \) and \( \phi \), downweight each observation separately, but it is also possible to consider a cluster down-weighting scheme, each element of the cluster is assigner the same weight \( w_i \). The same choices as in generalized linear models arise here for the weights: \( w_{it} \) as a function of \( r_{it} = (y_{it} - \mu_{it}) / \sqrt{\mu_{it}} / \sqrt{\phi} \) to ensure robustness with respect to outlying points in the \( y \)-space, or \( w_{it} \) as a function of the diagonal elements of the hat matrix \( h_{it} \) to handle leverage points. In practice, it makes often sense to combine both types of weights.

Note that the estimating equations in (9) are a slightly modified version of the estimating equations defined in Qaqish and Preisser (1999) and Preisser and Qaqish (1999). In this last paper robust estimators of the nuisance parameters \( \phi \) and \( \alpha \) (exchangeable correlation) based on the method of moments are also proposed.

Under some regularity conditions, the estimator obtained by solving (9) is asymptotically normally distributed with asymptotic variance \( M^{-1} Q M^{-1} \), where

\[ M = \lim_{K \to \infty} \frac{1}{K} \sum_{i=1}^{K} D_i^T \Gamma_i^T V_i^{-1} \Gamma_i D_i \]

and

\[ Q = \lim_{K \to \infty} \frac{1}{K} \sum_{i=1}^{K} D_i^T \Gamma_i^T V_i^{-1} \text{Var}(\psi_i) V_i^{-1} \Gamma_i D_i. \]

The estimator defined by (9) belongs to the class of M-estimators, and its influence function is therefore given by

\[ IF((X, y); T, F_\beta) = M^{-1} D^T(X, \beta) \Gamma^T V^{-1} (\psi - c) \]

\[ = M^{-1} D^T(X, \beta) \Gamma^T V^{-1} (W(y - \mu) - c), \]
for a generic observation \((X, y)\), and where \(T\) is the functional representing the estimator \(\hat{\beta}\). This influence function is bounded with respect to contaminations in the outcome as long as \(\psi\) is bounded, and with respect to contamination in the design if \(D^T \Gamma^T V^{-1} W \psi\) is bounded.

As in Section 2, the goal is to compare the adequacy of a submodel \(M_{p-q}\) with \((p-q)\) regression parameters to the adequacy of a larger model \(M_p\) with \(p\) regression parameters. A test statistic playing the same role as (4) can be defined by

\[
\Lambda_t(s) = 2 \left( \sum_{i=1}^{K} Q_t(s)(y_i, \hat{\mu}_i) - \sum_{i=1}^{K} Q_t(s)(y_i, \hat{\mu}_i) \right),
\]

where \(\hat{\mu}_i = \mu_i(X_i, \hat{\beta})\) is the estimation under model \(M_p\), and \(\mu_i = \mu_i(X_i, \hat{\beta})\) is the estimation under model \(M_q\), and where

\[
Q_t(s)(y_i, \mu_i) = 
\begin{aligned}
&\frac{1}{\phi} \int_{t_i(s)=\mu_i} (y_i - t_i)^TW(X_i, y_i, t_i)V^{-1}(t_i, \alpha) \Gamma_i(t_i)dt_i(s) \\
&- \frac{1}{\phi} \int_{t_i(s)=\mu_i} E((y_i - t_i)^TW(X_i, y_i, t_i))V^{-1}(t_i, \alpha) \Gamma_i(t_i)dt_i(s),
\end{aligned}
\]

with the integrals possibly path-dependent. A typical set of integration paths is given for example by \(t_{ii}(s) = y_{it} + (\mu_{it} - y_{it})s_{c_{it}}\), for \(c_{it} \geq 1\) and \(t = 1, \ldots, n_i\).

For the test statistic in (10) we have the following proposition:

**Proposition 3** Under conditions (A.1)-(A.9) in Heritier and Ronchetti (1994), [C1], [C2] in Appendix A, and under \(H_0 : \beta(2) = 0\), the test statistic \(\Lambda_t(s)\) defined by (10) equals

\[
KL_K^T(M_1 - \hat{M}^+)L_K + o_p(1) = KR_K^T M_{22} R_K(2) + o_p(1),
\]

where \(M_{22} = M_{22} - M_{12}^T M_{11}^{-1} M_{12}\), \(\sqrt{K} U_K\) and \(\sqrt{K} R_K\) are asymptotically normally distributed \(N(0, Q)\) and \(N(0, M^{-1} Q M^{-1})\) respectively.

Moreover, \(\Lambda_t(s)\) 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 \(Q(M^{-1} - \hat{M}^+)\), and \(\hat{M}^+\) is such that \(\hat{M}_{11}^+ = M_{11}^{-1}\) and \(\hat{M}_{12}^+ = 0\), \(\hat{M}_{21}^+ = 0\), \(\hat{M}_{22}^+ = 0\).
Table 1: Residual deviance and robust quasi-deviance ($c = 1.2$ and $w(X_i) = 1$ for all $i$). $p$-values are indicated within parentheses

<table>
<thead>
<tr>
<th></th>
<th>Resid. Deviance</th>
<th>Resid. robust quasi-deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>83.34</td>
<td>60.46</td>
</tr>
<tr>
<td>logdose</td>
<td>54.73 (0.000)</td>
<td>39.94 (0.000)</td>
</tr>
<tr>
<td>block2</td>
<td>45.59 (0.003)</td>
<td>35.21 (0.017)</td>
</tr>
<tr>
<td>block1</td>
<td>39.98 (0.018)</td>
<td>32.74 (0.085)</td>
</tr>
</tbody>
</table>

The proof of this proposition is given in Appendix B. Beside giving the asymptotic distribution and an asymptotically equivalent quadratic form to $\Lambda_t(s)$, Proposition 3 shows that the path-dependence of the integrals in (11) vanishes asymptotically. It makes then sense to use the difference of robust quasi-likelihood for inference. Moreover, thanks to the quadratic form in (12), Proposition 2 applies, and a similar result as Corollary 1 can be proved for statistic (10). The result ensures that the asymptotic level of $\Lambda_t(s)$ under contamination is under control.

4 Example

We analyze the so-called damaged carrots dataset (Phelps, 1982), issued from a soil experiment. The proportion of carrots showing insect damage in a trial with three blocks and eight dose levels of insecticide is recorded. The sample size is 24. A binomial model with log-link is fitted to the linear predictor made up by the intercept ($\beta_0$), the logarithm of the dose ($\beta_1$), a dummy variable indicating block 2 ($\beta_2$) and a dummy variable indicating block 1 ($\beta_3$). The Mallows quasi-likelihood estimates of the regression parameter with $c = 1.2$ and $w(X_i) = 1$ for all $i$ are given by (standard errors within parentheses): $\hat{\beta}_0 = 1.939$ (0.70), $\hat{\beta}_1 = -2.049$ (0.37), $\hat{\beta}_2 = 0.685$ (0.24) and $\hat{\beta}_3 = 0.450$ (0.24). Note that the constant $c = 1.2$ is obtained by applying the results in Section 2.1 with $\alpha - \alpha_0 = 0.02$, $\epsilon = 0.04$ and $\kappa = 0.1145$. The large outlier of this dataset (observation 14) is automatically downweighted with a weights of 0.26, whereas most of the observations receive a weight equal to 1 or at least greater than 0.70. As a comparison, classical maximum quasi-likelihood would have yield the following estimates: $\tilde{\beta}_0 = 1.480$ (0.66), $\tilde{\beta}_1 = -1.817$ (0.34), $\tilde{\beta}_2 = 0.843$ (0.23) and $\tilde{\beta}_3 = 0.542$ (0.23). The outlying observation has the effect of substantially increasing the estimated value of $\beta_2$. 

9
In order to better investigate these questions related to inference, we consider an analysis of robust difference of quasi-likelihood (equation (4)) to assess the significance of the variables used for modeling the response. We build the model via a Mallows-type difference of quasi-deviances (function (2) with \( c = 1.2 \), and \( w(X_i) = 1 \) for all \( i \)). As reported in the second column of Table 4, the robust procedure allows to identify that the dummy variable indicating block 1 does not contribute significantly to the model. At the contrary, a classical analysis would have kept all the variables, as one can see from the first column of Table 4.

5 Acknowledgments

This work has been partially supported by the Swiss National Science Foundation through its Prospective Researchers Fellowships Program.

A Assumptions for Proposition 3

[C1]: Denote by \( D_n \) the set of all sample points \( z_i, i = 1, \ldots, n \) for which the second-order derivatives \( \partial^2 Q_t(s)(z_i, \beta)/\partial \beta_j \partial \beta_k, i = 1, \ldots, n; j, k = 1, \ldots, p \) are continuous functions of \( \beta \). It is assumed that \( \lim_{n \to \infty} P_\beta(D_n) = 1 \).

[C2]: For any \( z \in D_n \), any positive value \( \delta \), and any \( \beta_1 \) denote by \( \eta_{jk}(z, \beta_1, \delta) \) the least upper bound and by \( \gamma_{jk}(z, \beta_1, \delta) \) the greatest lower bound of \( \partial^2 Q_t(s)(z, \beta)/\partial \beta_j \partial \beta_k \), with respect to \( \beta \) in the \( \beta \) interval \( ||\beta_1 - \beta|| \leq \delta \). Moreover, assume that for any sequence \( \{\delta_n\} \) for which \( \lim_{n \to \infty} \delta_n = 0 \),

\[
\lim_{n \to \infty} E_\beta[\eta_{jk}(z, \beta, \delta_n)] = \lim_{n \to \infty} E_\beta[\gamma_{jk}(z, \beta, \delta_n)] = E_\beta[\partial^2 Q_t(s)(z, \beta)/\partial \beta_j \partial \beta_k],
\]

and that there exists a positive \( \epsilon \) such that the expectations \( E_\beta[\eta_{jk}^2(z, \beta, \delta)] \) and \( E_\beta[\gamma_{jk}^2(z, \beta, \delta)] \) are bounded functions of \( \beta \) and \( \delta \) for all \( \beta \) and \( \delta < \epsilon \).

B Proof of Proposition 3

By considering a second order Taylor expansion of \( \sum_{i=1}^K Q_{t_i(s)}(y_i, \hat{\mu}_i) \) around \( \sum_{i=1}^K Q_{t_i(s)}(y_i, \hat{\mu}_i) \), and by the fact that \( K^{-1} \partial^2 / \partial \beta \partial \beta^T \sum_{i=1}^K Q_{t_i(s)}(y_i, \hat{\mu}_i) \) tends to \( M \) when \( K \to \infty \), we have by Slutsky’s theorem that

\[
\Lambda_{t(s)} \approx K(\hat{\beta} - \beta)^T M (\hat{\beta} - \beta).
\]
Moreover, for $K \to \infty$, under $H_0$ and by the asymptotic properties of the estimators $\hat{\beta}$ and $\hat{\beta}$, the following distribution equality holds under conditions (A1)-(A9) in Heritier and Ronchetti (1994):

$$\sqrt{K}(\hat{\beta} - \hat{\beta}) = \sqrt{K}(M^{-1} - \tilde{M}^+)L_K,$$

where $L_K = K^{-1}\partial/\partial \beta \sum_{i=1}^{K} Q_{t_i(s)}(y_i, \mu_i)$. This implies that

$$\Lambda_{t(s)} \simeq KL_T^T(M^{-1} - \tilde{M}^+)L_K,$$

or, equivalently, that

$$\Lambda_{t(s)} \simeq KR_{K(2)}T_{22}R_{K(2)}K^T221HR_{222}M_{221}1M_{22}1,$$

with $M_{221} = M_{22} - M_{12}^T M_{11}^{-1} M_{12}$.

The distributional statement follows from standard results on the distribution of quadratic forms in normal variables (see Johnson and Kotz, 1970).

References


