ROBUST QUASI-LIKELIHOODS FOR MODEL SELECTION IN LONGITUDINAL DATA ANALYSIS OR OTHERWISE CLUSTERED DATA

by

Eva Cantoni

September 2000

Department of Statistics
STANFORD UNIVERSITY
Stanford, California
ROBUST QUASI-LIKELIHOODS FOR MODEL SELECTION IN
LONGITUDINAL DATA ANALYSIS OR OTHERWISE
CLUSTERED DATA

by

EVA CANTONI
Econometrics Department
University of Geneva
Switzerland

SEPTEMBER 2000

This research was supported by the
Swiss National Science Foundation

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Robust Quasi-likelihoods for Model Selection in Longitudinal Data Analysis or Otherwise Clustered Data

Eva Cantoni

Statistics Department
Stanford University
390 Serra Mall
Stanford, CA 94305
September 11, 2000

Abstract
In this paper we introduce an extension of robust techniques for inference in generalized linear models to the analysis of longitudinal data. Robust versions of quasi-likelihood functions are obtained by applying robustness weights to their classical counterpart, and are used to construct a class of test statistics for model selection. The asymptotic theory of this class of test statistics is studied, and its robustness properties are assessed. The proposed class of test statistics yields reliable inference even in presence of deviating data points. The application to a real dataset confirms the benefit of such a robust analysis.

KEY WORDS: robust inference; longitudinal data; estimating equations; quasi-likelihood.
1 Introduction

Longitudinal data commonly arise in medical studies and in economics, where they are more often called panel data. The goal of longitudinal studies is to describe a response (or outcome) as a function of some covariates. The particularity of longitudinal data (as opposed to cross-sectional data) lies in the fact that they usually consist of repeated observations (over time) for each subject. Even though the measurements for different subjects can be considered independent, this is not the case for repeated measurements on the same subject, and this within-cluster correlation must be taken into account. Clusters can sometimes be defined by different criteria: for example the set of individuals in a medical practice (see, for instance, the example in Section 4).

In situations where there is only one measurement for each subject (and therefore no correlation structure), generalized linear models (McCullagh and Nelder, 1989) provide a powerful tool to analyze a broad variety of continuous and discrete data. There are several strategies to extend generalized linear models such that they account for the correlation within subjects: marginal mean models, random-effects models and transition models (see the general reference Diggle, Liang, and Zeger, 1996 for details). We choose here to follow the first approach, where we model the population-averaged response as a function of the covariates assuming independence between subjects. This approach describes the average population response with respect to changes in the covariates. If the goal is to describe the individual changes, a mixed-effects model should be used instead (see Zeger, Liang, and Albert, 1988 for details on the distinction between subject-specific and population-averaged models).

Because of the lack of likelihood methods for the analysis of multivariate non Gaussian data, essentially due to the lack of a rich class of joint distributions, longitudinal data analysis is successfully performed via the generalized estimating equations (GEE) approach of Liang and Zeger (1986) and Zeger and Liang (1986). Under mild regularity conditions, this procedure yields consistent and asymptotically Gaussian estimates, even when the correlation structure is misspecified, as it is often the case
in practice. The GEE procedure is a quasi-likelihood type procedure and therefore inference and model selection can be carried out using differences of quasi-likelihoods (see, for instance, Wedderburn, 1974, McCullagh, 1983, and McCullagh and Nelder, 1989). This is possible even though the derivation of a quasi-likelihood function from a set of estimating equations in the case of longitudinal data is more delicate than in the case of generalized linear models, due to the correlation within subjects. This issue is discussed in Section 2.3 below.

Because of the quasi-likelihood nature of the GEE procedure, the parameter estimates can be highly influenced by the presence of unusual data points and attention has to be payed to avoid misleading conclusions due to few outlying points. To cope with this problem in the estimation of the parameters, Qaqish and Preisser (1999) introduce a generalization of GEE which yields regression parameter estimates resistant to influential data (see Section 2.1). The robust estimation of the nuisance parameters has been proposed by the same authors in some particular cases. We address this problem in Section 2.2.

If quasi-likelihood functions are used for model selection and inference, attention has also to be payed to the robustness properties of the quasi-likelihood function itself in order to ensure reliable conclusions. The robustness of quasi-likelihood in generalized linear models have been studied and robust quasi-deviances for these models have been proposed by Cantoni and Ronchetti (2000). In Section 3 of the present paper we address the problem of influential observations in inference for longitudinal data, and we define a class of robust quasi-deviances functions that can be used for safe inference even in presence of outliers and leverage points. The asymptotic distribution and the robustness properties are derived. Beside the advantages of having a test statistics for model selection, quasi-likelihood functions give the possibility to distinguish between multiple roots, a known drawback of estimating equations (see Hanfelt and Liang, 1995).

We illustrate the use of the new technique in Section 4, where we study in details the GUIDE dataset.
2 Resistant GEE and Quasi-likelihood Functions

Consider the following general setting for longitudinal data. Let $Y_{it}$ be the outcome for subject $i$ at time $t$, for $i = 1, \ldots, K$ and $t = 1, \ldots, n_i$. Attached to each outcome, a $p \times 1$ vector of covariates $x_{it}$ is also measured. We will 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 model the marginal mean and assume that $E(Y_{it}) = \mu_{it}$, and that the relationship of $\mu_{it}$ with the linear predictor $\eta_{it}$ can be described in the same way for each subject by means of a regression parameter $\beta$ of dimension $p \times 1$. Formally, this means that $g(\mu_{it}) = x_{it}^T \beta = \eta_{it}$ for a known function $g$, usually called the link function. We further assume that the variance of $Y_{it}$ can be expressed through $\mu_{it}$ by a function $v$, that is $V(Y_{it}) = \phi v(\mu_{it})$. The parameter $\phi$ is a scale parameter and it is regarded as a nuisance parameter, to be estimated. Note, however, that for binary and counts data (without over-dispersion), $\phi$ is equal to 1 and no estimation is required.

2.1 Resistant GEE

The set of resistant generalized estimating equations as defined by Qaqish and Preissler (1999) are given by

$$
\sum_{t=1}^{K} D_i^T V_i^{-1} (\psi_i - c_i) = 0,
$$

where $D_i = D_i(X_i, \beta) = \frac{\partial \mu_i}{\partial \beta}$ is a $n_i \times p$ matrix, and $V_i = V_i(\mu_i, \alpha) = A_i R_i(\alpha) A_i$ is a $n_i \times n_i$ matrix with $A_i = A_i(\mu_i) = \text{diag}(v^{1/2}(\mu_{i1}), \ldots, v^{1/2}(\mu_{in_i}))$. $R_i(\alpha)$ – which depends on a $s$-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 $c_i = E(\psi_i)$. Note that $W_i$ may also depend on $\alpha$ and $\phi$, but these arguments have been dropped for sake
of simplicity. The weights \( w_{it} \) down-weight each observation separately. It is also possible to consider a cluster down-weighting scheme, where each element of the cluster is assigned the same weight \( w_i \).

Equation (1) defines a set of estimating equations for the parameter \( \beta \) based only on the distributional assumptions about the first two moments of the response. Note, however, that a fully specified distribution is needed to compute \( c_i \), if \( W_i \) depends on \( Y_i \). Calibration of these terms is usually done with the most close distribution matching the first two moments assumptions (e.g. Bernoulli, binomial, Poisson or normal distribution). This calibration is not necessary in the case of classical GEE as defined in Liang and Zeger (1986), which can be reproduced with \( W_i = I \) in (1).

Define the \( (\sum n_i) \times p \) matrix \( X_{tot} = (X_{i1}^T, \ldots, X_{iK}^T)^T \), and the \( (\sum n_i) \times 1 \) vectors \( \psi_{tot} = (\psi_{i1}^T, \ldots, \psi_{iK}^T)^T \) and \( c_{tot} = (c_{i1}^T, \ldots, c_{iK}^T)^T \). The estimation of \( \beta \) can be performed via iterative reweighted least squares by regressing the adjusted dependent variable \( Z = X_{tot} \hat{\beta} + D^*(\psi_{tot} - c_{tot}) \) on \( X_{tot} \) with block diagonal weight matrix \( W^* \), where the \( i \)-th block of \( W^* \) is the \( n_i \times n_i \) matrix \( W_i^* = D_i^{-1}A_i^{-1}R_i^{-1}A_i^{-1}D_i^{-1} \), and where \( D^* \) is a block diagonal matrix with blocks \( D_i^* = \text{diag}(\partial \eta_{i1}/\partial \mu_{i1}, \ldots, \partial \eta_{in_i}/\partial \mu_{in_i}) \). Remark that \( D_i = D_i^{-1}X_i \). The matrix \( H_i = X_i(X^TW^*X)^{-1}X_i^TW_i^* \) defines the hat matrix for subject \( i \).

There are several ways of choosing the weights \( w_{it} \). To ensure robustness with respect to outlying points in the \( Y \)-space, one can consider weights of the form \( w_{it} = w(r_{it}/\sqrt{\phi}) \), where \( r_{it} = (y_{it} - \mu_{it})/\nu^{1/2}(\mu_{it}) \) are the Pearson residuals. For example, Preisser and Qaqish (1999) use the redescending weight function \( w(r_{it}) = \exp(-a(r_{it}/a)^2) \), for a given value of the tuning constant \( a \). Another choice is \( w(r_{it}) = c/|r_{it}/\sqrt{\phi}| \), if \( |r_{it}/\sqrt{\phi}| > c \), and \( w(r_{it}) = 1 \) otherwise. This comes down to the choice \( w(r_{it}) = \psi_c(r_{it}/\sqrt{\phi})/(r_{it}/\sqrt{\phi}) \), where \( \psi_c \) is the Huber function with tuning constant \( c \), when there is only one observation per subject (generalized linear models), as in Cantoni and Ronchetti (2000). For protection against leverage points, the weights can be defined as a function of the hat matrices \( H_i \), that is \( w_{it} = w(h_{it}) \), where \( h_{it} \) is the \( t \)-th diagonal element of \( H_i \). For example, a common choice is \( w(h_{it}) = \sqrt{1 - h_{it}} \), or \( w(h_{it}) = \exp(-a(h_{it}/a)^2) \) as in Preisser and Qaqish.
In practical situations, it is often sensible to combine both types of weights—on the design and on the outcome—to obtain global robustness.

Qaqish and Preissir (1999) prove that under some regularity conditions, the estimator \( \hat{\beta} \) which solves the set of equations (1) is consistent, and that \( \sqrt{K}(\hat{\beta} - \beta) \) is asymptotically normal distributed with variance \( \lim_{K \to \infty} A_{K}^{-1} G_{K} A_{K}^{-1} \), where

\[
A_{K} = \frac{1}{K} \sum_{i=1}^{K} D_{i}^{T} V_{i}^{-1} \Gamma_{i} D_{i} \quad \text{and} \quad G_{K} = \frac{1}{K} \sum_{i=1}^{K} D_{i}^{T} V_{i}^{-1} \text{Var}(\psi_{i}) V_{i}^{-1} D_{i},
\]

with \( \Gamma_{i} = E(\tilde{\psi}_{i} - \tilde{c}_{i}), \tilde{\psi}_{i} = \partial \psi_{i} / \partial \mu_{i} \) and \( \tilde{c}_{i} = \partial c_{i} / \partial \mu_{i} \).

The variance of \( \beta \) can be estimated by

\[
\left( \sum_{i=1}^{K} D_{i}^{T} V_{i}^{-1} \Gamma_{i} D_{i} \right)^{-1} \left( \sum_{i=1}^{K} D_{i}^{T} V_{i}^{-1}(\psi_{i} - c_{i})(\psi_{i} - c_{i})^{T} V_{i}^{-1} D_{i} \right) \left( \sum_{i=1}^{K} D_{i}^{T} V_{i}^{-1} \Gamma_{i} D_{i} \right)^{-T},
\]

where the estimates of \( \phi, \alpha \) and \( \beta \) are plugged-in. In the case of a single observation per subject, one can also take \( \text{Var}(\psi_{i}) = E(\psi_{i}^{2}) \), calibrated with the most close distribution. Note that both estimators \( \hat{\beta} \) and \( \text{Var}(\hat{\beta}) \) are consistent regardless of the choice of the working correlation matrix. The equivalent results for classical GEE are obtained by Liang and Zeger (1986).

In fact, the estimator defined by (1) is a M-estimator (Huber, 1964), and its influence function (Hampel, 1974) is therefore given by

\[
IF((X, y); T, F_{\beta}) = \left( \lim_{K \to \infty} A_{K} \right)^{-1} D^{T}(X, \beta)V^{-1}(\psi - c)
\]

\[
= \left( \lim_{K \to \infty} A_{K} \right)^{-1} D^{T}(X, \beta)V^{-1}(W(y - \mu) - c),
\]

which is bounded with respect to contaminations in the outcome as long as \( \psi \) is, and with respect to contamination in the design if \( D^{T}V^{-1}W\psi \) is.

From general theory (see Godambe and Heyde, 1987 and Morton, 1981), it is known that the optimal estimating functions based on \( (Y_{i} - \mu_{i}) \) are given by \( \sum_{i=1}^{K} D_{i}^{T} V^{-1}(Y_{i} - \mu_{i}) = 0 \). These estimating functions are optimal in the sense that they produce an estimator which has the smallest variance among the class of estimators that are linear combinations of the elementary functions \( (Y_{i} - \mu_{i}) \). With
resistant GEE, two distinct sources of loss of efficiency arise: the use of the working
correlation matrix instead of the true one, and the introduction of the robustness
weights. The first point – shared with classical GEE – has no practical solution, and
simply calls for a sensible choice of the working correlation structure. The second
point is the usual trade-off between robustness and efficiency, and it is the price to
pay to achieve safe inference. This trade-off is usually tuned with the tuning parame-
ters of the robustness weights such that the loss of efficiency is around a certain
target, 5% say.

2.2 Estimation of \( \phi \) and \( \alpha \)

The estimation of the nuisance parameter \( \phi \) and of the correlation parameter \( \alpha \) has
also to be made robust, to avoid harmful consequences on the estimation of the
regression parameters. With classical GEE, \( \sqrt{K} \)-consistent estimators of \( \phi \) and \( \alpha \)
are derived by the method of moments based on Pearson residuals, and are plugged-in
in the estimating equations. The robust counterpart to Pearson residuals are the
weighted residuals \( r_{it}^* = (\psi_{it} - c_{it})/v^{1/2}(\mu_{it}) \). By analogy with the classical case,
Preisser and Qaqish (1999) suggest to robustly estimate \( \phi \) by

\[
\hat{\phi} = \frac{K}{\sum_{i=1}^{K} \sum_{t=1}^{n_i} r_{it}^*} \left( \frac{\sum_{i=1}^{K} \sum_{t=1}^{n_i} b_{it}^2}{p} \right),
\]

where \( B_i = \text{diag}(b_{i1}, \ldots, b_{in_i}) \) is such that \( \text{Var}(\psi_i) = B_i \text{Var}(Y_i) B_i \), whenever de-
defined. Formula (2) gives the estimator by Liang and Zeger (1986) when \( W_i = I \),
which also implies \( B_i = I \).

The specific estimator of \( \alpha \) depends upon the parametric choice of \( R_i(\alpha) \). The
case of exchangeable correlation, that is \( R_i = \alpha 1_{n_i} 1_{n_i}^T + (1 - \alpha) I_{n_i} \), has been studied
by Preisser and Qaqish (1999), and their proposal is

\[
\hat{\alpha} = \frac{1}{\hat{\phi}} \frac{K}{\sum_{i=1}^{K} \sum_{t>s} r_{it}^* r_{is}^*} \left( \frac{\sum_{i=1}^{K} \sum_{t>s} b_{it} b_{is}}{p} \right),
\]

which again yields the classical estimator as a particular case.
Both proposals (2) and (3) rely on the condition that $\text{Var}(\psi_i)$ can be decomposed as $B_i \text{Var}(Y_i) B_i$. This is not always possible, and even when it is, the joint distribution of $(Y_{it}, Y_{is})$ is often required. Moreover, this joint distribution is not completely defined given only the marginal distribution and the correlation $\rho_{ts}$, except in the Gaussian case and for Bernoulli trials. However, it has to be noticed that the estimator (3) of $\alpha$, as well as its classical counterpart, can handle an arbitrary number of observations per subject, and different observation times for each subject. In situation where $n_i = n$ for all $i$, and where the measurement times are the same for all subjects, one can also consider to let $R(\alpha)$ totally unspecified, in which case $s = n(n-1)/2$. The estimation of $R(\alpha)$ takes the form

$$\hat{R} = \frac{1}{\hat{\phi}} \text{Var}(A^{-1}(y - \mu)).$$

Classical GEE make use of the variance estimator $1/K \sum_{i=1}^{K} A_i^{-1}(Y_i - \mu_i)(Y_i - \mu_i)^T A_i^{-1}$, but other robust estimators can be used instead: for example, if the covariates are continuous, the minimum volume ellipsoid estimate of Rousseeuw and van Zomeren (1990) can be considered.

### 2.3 Estimating Equations and Quasi-likelihood Functions

Estimating equations, like the set in (1), are very attractive because they avoid to assume strict hypotheses on the underlying process, other than the two first moments of the outcome. Quasi-likelihood functions (Wedderburn, 1974 and McCullagh, 1983) go a step further and construct a function to be used as a likelihood, even when there is not enough information to construct a proper likelihood function. Moreover, estimating equations are known to have multiple roots, and quasi-likelihood functions can help in distinguish between them.

We consider the general setting where a response vector $Y$ is described by its mean $\mu = E(Y)$ and by its covariance structure $\text{Var}(Y) = \phi V(\mu)$. We assume that $\mu = \mu(x, \beta)$. Let us consider the set of estimating equations

$$U(\beta) = D^TV^{-1}(Y - \mu) = 0,$$

(4)
where $D = \partial \mu / \partial \beta$.

A quasi-likelihood function $Q_{t(s)}$ is defined by the (possibly path-dependent) line integral

$$Q_{t(s)}(y, \mu) = \frac{1}{\phi} \int_{t(s)=y}^{t(s)=\mu} (y - t)^TV^{-1}(t)dt(s),$$  \hspace{1cm} (5)

along a path $t(s)$ in $\mathbb{R}^n$ from $t(s_0) = y$ to $t(s_1) = \mu$. If the integral in (5) is path-independent, then the gradient of $Q_{t(s)}(y, \mu)$ with respect to $\beta$ is given by (4). In fact, the necessary and sufficient condition for $U(\beta)$ to be the gradient of $Q_{t(s)}(y, \mu)$ is that $\partial U(\beta) / \partial \beta$ is symmetric – see McCullagh (1983), McCullagh and Nelder (1989) p. 332-339, and Hanfelt and Liang (1995) for additional details. This is in general not true for arbitrary $V(\mu)$, even though the expectation of $\partial U(\beta) / \partial \beta$ is symmetric. Note that if $V(\mu) = \text{diag}(V(\mu_1), \ldots, V(\mu_n))$ this condition is fulfilled, and the quasi-likelihood function is uniquely defined and given by

$$Q(y, \mu) = \frac{1}{\phi} \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt.$$ 

In the case of GEE (classical and resistant), the estimating equations are sum of $U(\beta)$ type functions as defined in (4), one for each subject of the sample. Due to the correlation between measurements on a same subject, the covariance matrix is not diagonal, and moreover the symmetry condition does not hold in general. However, results in Hanfelt and Liang (1995) show that the asymptotic distribution of $Q_{t(s)}(y, \mu)$ is path-independent, and that the path of integration does not play an important role in finite-sample situations. Therefore, the function in (5) can nevertheless be used to carry out inference. Hanfelt and Liang (1995) also prove that, even under path-dependence, the quasi-likelihood function $Q_{t(s)}(y, \mu)$ can be used to distinguish between multiple roots. In fact, inconsistent roots of the estimating functions eventually behave like local minimizer of the quasi-likelihood function as $n \to \infty$, whereas consistent roots eventually behave like local maximizer.

An alternative approach would consist of constructing potential function for projected scores as in Li and McCullagh (1994). Note, however, that the choice of the
path of integration is replaced by the choice of a prior measure on the parameter $\beta$, which can be even more arbitrary.

3 Robust Quasi-deviances for Inference in GEE

Let $M_p$ a model for longitudinal data as specified in Section 2 involving $p$ covariates, and $M_{p-q}$ a sub-model involving only $(p - q)$ covariates. We consider the partition of the parameter $\beta^T$ as $(\beta_{(1)}^T, \beta_{(2)}^T)^T$ into $(p - q)$ and $q$ components. Matrices will be partitioned accordingly.

We would like to establish whether the sub-model is rich enough to explain the variation in the outcome, or whether the larger model is needed. We are going to use the quasi-likelihood approach for robust model selection between nested models, and construct a class of likelihood-ratio type test statistic. This approach has already proven useful in obtaining robust inference for generalized linear models (see Cantoni and Ronchetti, 2000). The starting point will be the set of estimating equations (1).

We define the test statistic

$$
\Lambda_{t(s)} = 2 \left( \sum_{i=1}^{K} Q_{t_i(s)}(y_i, \hat{\mu}_i) - \sum_{i=1}^{K} Q_{t_i(s)}(y_i, \hat{\mu}_i) \right),
$$

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

$$
Q_{t_i(s)}(y_i, \mu_i) = \frac{1}{\phi} \int_{\substack{t_i(s) = \mu_i \\ t_i(s) = y_i}} (y_i - t_i)^T W(X_i, y_i, t_i)V^{-1}(t_i, \alpha)dt_i(s) + \frac{1}{\phi} \int_{\substack{t_i(s) = \mu_i \\ t_i(s) = y_i}} E((y_i - t_i)^T W(X_i, y_i, t_i)V^{-1}(t_i, \alpha)dt_i(s),
$$

with the integrals possibly path-dependent. A typical set of integration paths is given by $t_i(s) = y_i + (\mu_i - y_i)s^c$, for $c \geq 1$. For example, when $c = 1$, we have that

$$
Q_{t_i(s)}(y_i, \mu_i) = -\frac{1}{\phi}(y_i - \mu_i)^T \left( \int_0^1 sW(X_i, y_i, t_i(s))V^{-1}(t_i(s), \alpha)ds \right)(y_i - \mu_i) + \frac{1}{\phi} \int_0^1 E((y_i - t_i(s))^T W(X_i, y_i, t_i(s)))V^{-1}(t_i(s), \alpha)(y_i - \mu_i)ds,
$$
so that the integrations are univariate.

In the following proposition, we establish the asymptotic distribution of $\Lambda_{t(s)}$.

**Proposition 1** Under conditions (A.1)-(A.9) in Héritier and Ronchetti (1994), [C1], [C2] in Appendix B, and under $H_0 : \beta_{(2)} = 0$, the test statistic $\Lambda_{t(s)}$ defined by (6) equals

$$KU_K^T(A_K^{-1} - \tilde{A}^+)U_K + o_p(1) = KR_{K(2)}^T A_{K(2)} R_{K(2)} + o_p(1),$$

where $A_{K(2),1} = A_{K2} - A_{K21}A_{K11}^{-1}A_{K12}$, $\sqrt{K}U_K$ and $\sqrt{K}R_K$ are asymptotically normally distributed $N(0, \lim_{K \to \infty} G_K)$ and $N(0, \lim_{K \to \infty} A_K^{-1}G_KA_K^{-1})$ 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 $G_K(A_K^{-1} - \tilde{A}^+)$, and $\tilde{A}^+$ is such that $\tilde{A}_{11}^+ = A_{K11}^{-1}$ and $\tilde{A}_{12}^+ = 0$, $\tilde{A}_{21}^+ = 0$, $\tilde{A}_{22}^+ = 0$.

A sketch of the proof is given in Appendix C. The asymptotic distribution of the statistic (6) is a linear combination of central $\chi^2$ variables with one degree of freedom (see Chapter 29 in Johnson and Kotz, 1970, and also Davies, 1980 and Farebrother, 1990 for algorithms to compute probabilities accordingly).

Proposition 1 shows that the path-dependence of the integrals defining (6) vanishes asymptotically, diminishing the first sight concern about the consequences of the choice of the path. Proposition 1 also implies that the numerically expensive integrations involved in the computation of the test statistic (6) can be avoided by using the asymptotically equivalent quadratic form $R_{K(2)}^T A_{K(2)} R_{K(2)}$, with $R_{K(2)} = \hat{\beta}_{(2)}$.

Thanks to the quadratic form (7), we can apply Proposition 2 and Corollary 1 in Cantoni and Ronchetti (2000). This means that the asymptotic level of the test statistic under $\varepsilon$-contaminations around the model is bounded as long as a M-estimator $\hat{\beta}_{(2)}$ with bounded influence function is used to build the test statistic. A
similar result holds for the asymptotic power. This formally proves that inference based on the difference of quasi-deviances (6) is safe even under contamination, if the test statistics is used with a bounded influence M-estimator \( \hat{\beta}_{(2)} \).

The result about the asymptotic level can also be used to choose the tuning constants of the robustness weights in such a way to control the maximal bias on the asymptotic level of the test in a neighborhood of the model (see Cantoni and Ronchetti, 2000 for details).

4 Application to the GUIDE Study

We will study the dataset of the GUIDE study (Guidelines for Urinary Incontinence Discussion and Evaluation), as used in Preisser and Qaqish (1999). The data are available at http://www.phs.wfubmc.edu/data/uipreiss.html. The outcome variable is the coded answer (bothered: 1 for “yes”, 0 for “no”) of a patient to the question: “Do you consider this accidental loss of urine a problem that interferes with your day to day activities or bother you in other ways?” There are 5 explanatory variables: the gender, coded as an indicator for women (female), the age (minus 76, divided by 10: age), the average number of leaking accidents per day (dayacc), the degree of the leak (severe: 1 = just create some moisture, 2 = wet their underwear (or pad), 3 = trickle down their thigh, 4 = wet the floor), and the daily number of visit to the toilet to urinate (toilet). There are 137 patients in the study, divided into 38 clusters. The clusters are defined by the practice, which means that patients from different practices are assumed independent. We assume common (exchangeable) correlation \( \alpha \) within any two patients of each cluster. The model considered for this dataset is a logit-link model without overdispersion (\( \phi = 1 \)), where the linear predictor takes the form:

\[
\eta = \beta_0 + \beta_1 x_{\text{female}} + \beta_2 x_{\text{age}} + \beta_3 x_{\text{dayacc}} + \beta_4 x_{\text{severe}} + \beta_5 x_{\text{toilet}}.
\]

[Table 1 about here.]
Table 1 reports the estimated coefficients by the robust method and by the classical method. Robustness weights of the form \( w_{it} = w_x(h_{it})w_y(r_{it}/\sqrt{\phi}) \) are used (see Appendix A for detailed computations). Several combinations of weighting schemes have been listed (see description in Section 2.1), along with the classical analysis which corresponds to the situation with no weights at all. Note that columns 3 and 4 of the table reproduce the results of Preisser and Qaqish (1999).

A close look to the robustness weights on the design and on the response (not shown) suggests that both types of weights should be used simultaneously to obtain a safe analysis. In fact, the dataset seems to contain both outliers and leverage points. Therefore, the analyses reported in the three last columns of Table 1 have to be preferred. It has to be noticed that the large difference between the classical and the robust analysis is essentially due to few observations. For example, in the analysis with Huber type weights on the Y-space and square-root-weights on the design (column 5 of Table 1), there are 10 observations with combined weight \( w_{it} = w_x(h_{it})w_y(r_{it}/\sqrt{\phi}) \) lower than 0.8. These observations represent only 7.3% of the dataset, but they affect the coefficient estimates in a sensible manner, as one can see from Table 1.

With respect to the classical analysis, the robust estimates suggest that the variable toilet could potentially play an important role, whereas some doubts arise on the significance of the variable severe. The importance of dayacc seems to be equally well assessed in the classical analysis as in the robust analysis. But these questions related to model selection need to be studied in a more precise manner, and we further investigate them by means of the test statistic developed in Section 3.

We consider a stepwise forward procedure starting from the empty model. At each step we add the variable that is mostly significant, that is the variable that has the smallest p-value, or, equivalently, that gives the largest value of the test statistic (6) (standardized with respect to \( d_1 \), see Proposition 1). The procedure is stopped when the test statistic is no longer significant at, say, the 5% level. This happens when the standardized test statistic is lower than 3.84, which is the 0.95-quantile of a \( \chi^2_1 \)-variable. To compute the test statistic (6), we use the asymptotic
equivalent quadratic form (7) with Huber type weights \( w.y(r_{it}/\sqrt{\phi}) \) (\( c = 1.5 \)), and square-root type weights \( w.x(h_{it}) \).

[Table 2 about here.]

Table 2 lists the result of this analysis. The final model retains three variables: \texttt{dayacc}, \texttt{severe} and \texttt{toilet}, which are clearly significant at the 5% level. It is also clear from the analysis that the age and the sex of the patient do not contribute in a significant manner to the variability of the response.

The test statistic (6) allows to test not only univariate hypotheses – as we did in the stepwise procedure – but also multivariate hypotheses. For example, in view of the results of the first step of the forward procedure reported in Table 2, one could have decided to test whether the addition to the empty model of the three variables \texttt{dayacc}, \texttt{severe} and \texttt{toilet} simultaneously would have been relevant. This particular test yields a \( p \)-value of 0.0001, strongly supporting the important contribution of the three variables to the model.

The analysis carried out with the robust techniques protects against outliers and leverage points, yielding safe inference. As a comparison, the same analysis with no robustness weights (corresponding to a classical analysis) would have built a model with the variables \texttt{dayacc} and \texttt{severe} only, neglecting the importance of the variable \texttt{toilet} (results not shown). The outlying points of the data set seem to hide the contribution of this variable to the model.

5 Conclusions

This paper develops robust quasi-likelihood functions to use for model selection in a class of likelihood-ratio type test statistics, which can automatically and safely handle data containing outlying points. The asymptotic distribution of this class of test statistics is derived. This class of test statistics enjoys good robustness properties on the asymptotic level and on the asymptotic power. The study of a real data set supports the relevance of the addressed problem, and the benefits of
the new technique. With respect to its classical counterpart, the robust analysis allows to identify the importance of an additional variable in the model.

A general proposal for the robust estimation of the nuisance parameters $\phi$, and the correlation parameter $\alpha$ is left for further work.

6 Acknowledgment

The author would like to thank Scott Zeger for providing bibliographic references at the early stages of this work, and Elvezio Ronchetti for helpful discussions. This research has been carried out during the postdoctoral year of the author at Stanford University. The financial support of the Swiss National Science Foundation and the hospitality of Stanford University are gratefully acknowledged.
Appendix

A Bernoulli Trials in Details

Let $Y_{it}$ and $Y_{is}$ be Bernoulli distributed with probability of success equal to $\mu_{it}$, and $\mu_{is}$ respectively. Further assume that the correlation between these two random variables is $\rho_{ts}$. We also suppose that the robustness weight $w_{it}$ associated with subject $i$ at time $t$ can be factorized as $w \cdot x(h_{it})w \cdot y(\tau_{it})$. The joint distribution of $(Y_{it}, Y_{is})$ is multinomial with set of probabilities $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$, where $\pi_{11} = \rho_{ts} v_{it}^{1/2} v_{is}^{1/2} + \mu_{it} \mu_{is}$, $\pi_{10} = \mu_{it} - \pi_{11}$, $\pi_{01} = \mu_{is} - \pi_{11}$ and $\pi_{00} = 1 - \mu_{it} - \mu_{is} + \pi_{11}$.

The consistency correction $c_{it} = E(\psi_{it})$ takes the form:

$$c_{it} = w \cdot x(h_{it}) (w \cdot y_{it}^{(1)} - w \cdot y_{it}^{(0)}) v(\mu_{it}),$$

where $w \cdot y_{it}^{(j)} = w \cdot y((j - \mu_{it})/v(\mu_{it})/\sqrt{\phi})$ is the weight for the $t$-th observation of subject $i$ evaluated at $y_{it} = j$.

Moreover, by using the set of probabilities $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ as specified above, we have that

$$Cov(\psi_{it}, \psi_{is}) = \rho_{ts} v_{it}^{1/2}(\mu_{it}) v_{is}^{1/2}(\mu_{is}) b_{it} b_{is},$$

which implies $Var(\psi_{it}) = b_{it}^2 v(\mu_{it})$.

Finally, the matrix $\Gamma_i = E(\tilde{\psi}_i - \tilde{c}_i)$, with $\tilde{\psi}_i = \partial \psi_i / \partial \mu_i$ and $\tilde{c}_i = \partial c_i / \partial \mu_i$, has diagonal elements $\Gamma_{it} = -b_{it}$.

B Assumptions for Proposition 1

[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 $z$, 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 $\partial^2 Q_t(s)(z_i, \beta)/\partial \beta_j \partial \beta_k$, with respect to $\beta$ in the $\beta$ interval $||\beta_1 - \beta|| \leq \delta$.  

15
Moreover, assume that for any sequence \( \{\delta_n\} \) for which \( \lim_{n \to \infty} \delta_n = 0 \),
\[
\lim_{n \to \infty} E_\beta [\eta_{jk}(z, \beta, \delta_n)] = \lim_{n \to \infty} E_\beta [\gamma_{jk}(z, \beta, \delta_n)] = E_\beta [\partial^2 Q_t(z, \beta, \delta_n) / \partial \beta_j \partial \beta_k],
\]
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 \( \delta < \epsilon \).

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

**C  Proposition 1: Sketch of the Proof**

By considering a second order Taylor expansion of \( \sum_{i=1}^K Q_t(y_i, \mu_t) \) around \( \sum_{i=1}^K Q_t(y_i, \mu_t) \), and by the fact that \( K^{-1} \frac{\partial^2}{\partial \beta \partial \beta^T} \sum_{i=1}^K Q_t(y_i, \mu_\iota) \) tends to \( \lim_{K \to \infty} A_K \) when \( K \to \infty \), we have – by Slutsky’s theorem – that
\[
\Lambda_t(z) \simeq K(\hat{\beta} - \beta)^T \left( \lim_{K \to \infty} A_K \right) (\hat{\beta} - \beta).
\]

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 hold under conditions (A1)-(A9) in Hérétier and Ronchetti (1994):
\[
\sqrt{K}(\hat{\beta} - \beta) = \sqrt{K}(A_K^{-1} - \bar{A}^+) U_K,
\]
where \( U_K = K^{-1} \frac{\partial}{\partial \beta} \sum_{i=1}^K Q_t(y_i, \mu_i) \). This implies that
\[
\Lambda_t(z) \simeq K U_K^T (A_K^{-1} - \bar{A}^+) U_K,
\]
or, equivalently, that
\[
\Lambda_t(z) \simeq K R_{K(2)}^T \left( \lim_{K \to \infty} A_{K_{22.1}} \right) R_{K(2)},
\]
with \( A_{K_{22.1}} = A_{K_{22}} - A_{K_{12}} A_{K_{11}}^{-1} A_{K_{12}} \).

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


<table>
<thead>
<tr>
<th>W.y</th>
<th>Huber (c=1.465)</th>
<th>redesc. (a=4)</th>
<th></th>
<th>Huber (c=1.445)</th>
<th>Huber (c=1.5)</th>
<th>redesc. (a=4)</th>
<th>redesc. (a=4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W.x</td>
<td>--</td>
<td>--</td>
<td></td>
<td>redesc. (a=1.8)</td>
<td>sqroot</td>
<td>redesc. (a=1.8)</td>
<td>redesc. (a=1.8)</td>
</tr>
<tr>
<td>$\hat{\alpha}$</td>
<td>0.10</td>
<td>0.07</td>
<td>0.03</td>
<td>0.11</td>
<td>0.07</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>(0.96)</td>
<td>(0.97)</td>
<td>(1.21)</td>
<td>(1.03)</td>
<td>(0.97)</td>
<td>(1.00)</td>
<td>(1.10)</td>
</tr>
<tr>
<td>female</td>
<td>-0.75</td>
<td>-1.06</td>
<td>-1.33</td>
<td>-0.72</td>
<td>-1.05</td>
<td>-0.98</td>
<td>-1.14</td>
</tr>
<tr>
<td></td>
<td>(0.60)</td>
<td>(0.62)</td>
<td>(0.75)</td>
<td>(0.62)</td>
<td>(0.62)</td>
<td>(0.61)</td>
<td>(0.66)</td>
</tr>
<tr>
<td>age</td>
<td>-0.68</td>
<td>-1.02</td>
<td>-1.49</td>
<td>-0.56</td>
<td>-0.99</td>
<td>-0.80</td>
<td>-1.02</td>
</tr>
<tr>
<td></td>
<td>(0.56)</td>
<td>(0.60)</td>
<td>(0.70)</td>
<td>(0.57)</td>
<td>(0.60)</td>
<td>(0.58)</td>
<td>(0.64)</td>
</tr>
<tr>
<td>dayacc</td>
<td>0.39</td>
<td>0.44</td>
<td>0.56</td>
<td>0.40</td>
<td>0.44</td>
<td>0.44</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.10)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>severe</td>
<td>0.81</td>
<td>0.66</td>
<td>0.71</td>
<td>0.80</td>
<td>0.66</td>
<td>0.65</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>(0.36)</td>
<td>(0.34)</td>
<td>(0.37)</td>
<td>(0.36)</td>
<td>(0.34)</td>
<td>(0.34)</td>
<td>(0.35)</td>
</tr>
<tr>
<td>toilet</td>
<td>0.11</td>
<td>0.21</td>
<td>0.36</td>
<td>0.15</td>
<td>0.21</td>
<td>0.21</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.13)</td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.11)</td>
</tr>
<tr>
<td></td>
<td>Step 1</td>
<td></td>
<td>Step 2</td>
<td></td>
<td>Step 3</td>
<td></td>
<td>Step 4</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>----------</td>
<td>--------</td>
<td>----------</td>
<td>--------</td>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_t$</td>
<td>p-value</td>
<td>$\Lambda_t$</td>
<td>p-value</td>
<td>$\Lambda_t$</td>
<td>p-value</td>
<td>$\Lambda_t$</td>
</tr>
<tr>
<td>female</td>
<td>3.027</td>
<td>(0.0819)</td>
<td>1.925</td>
<td>(0.1654)</td>
<td>2.336</td>
<td>(0.1264)</td>
<td>1.956</td>
</tr>
<tr>
<td>age</td>
<td>1.773</td>
<td>(0.1830)</td>
<td>3.456</td>
<td>(0.0630)</td>
<td>2.734</td>
<td>(0.0983)</td>
<td>1.721</td>
</tr>
<tr>
<td>dayacc</td>
<td>19.652</td>
<td>(0.0000)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>severe</td>
<td>14.085</td>
<td>(0.0002)</td>
<td>7.853</td>
<td>(0.0051)</td>
<td>4.914</td>
<td>(0.0266)</td>
<td>–</td>
</tr>
<tr>
<td>toilet</td>
<td>11.198</td>
<td>(0.0008)</td>
<td>7.855</td>
<td>(0.0051)</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 2: Forward stepwise analysis for model selection. Underlined values highlight the added variable at each step.