Working Paper Series, N. 10, May 2005

Adjusted quasi-profile likelihoods from estimating functions

Ruggero Bellio
Department of Statistics
University of Udine
Italy

Luca Greco and Laura Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: Higher-order adjustments for a quasi-profile likelihood for a scalar parameter of interest in the presence of nuisance parameters are discussed. Paralleling likelihood asymptotics, these adjustments aim to alleviate some of the problems inherent to the presence of nuisance parameters. Indeed, the estimating equation for the parameter of interest, when the nuisance parameter is substituted with an appropriate estimate, is not unbiased and such a bias can lead to poor inference on the parameter of interest. Following the approach of McCullagh and Tibshirani (1990), here we propose adjustments for the estimating equation for the parameter of interest. Moreover, we discuss two methods for their computation: a bootstrap simulation method, and a first-order asymptotic expression, which can be simplified under an orthogonality assumption. Some examples, in the context of generalized linear models and of robust inference, are provided.

Keywords: Asymptotics; bootstrap; nuisance parameter; profile and modified profile likelihood; robustness.

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Laura Ventura
tel: +39 049 827 4177
ventura@stat.unipd.it
http://www.stat.unipd.it/~ventura

Adjusted quasi-profile likelihoods from estimating functions

Ruggero Bellio
Department of Statistics
University of Udine
Italy

Luca Greco and Laura Ventura
Department of Statistical Sciences
University of Padua
Italy

Abstract: Higher-order adjustments for a quasi-profile likelihood for a scalar parameter of interest in the presence of nuisance parameters are discussed. Paralleling likelihood asymptotics, these adjustments aim to alleviate some of the problems inherent to the presence of nuisance parameters. Indeed, the estimating equation for the parameter of interest, when the nuisance parameter is substituted with an appropriate estimate, is not unbiased and such a bias can lead to poor inference on the parameter of interest. Following the approach of McCullagh and Tibshirani (1990), here we propose adjustments for the estimating equation for the parameter of interest. Moreover, we discuss two methods for their computation: a bootstrap simulation method, and a first-order asymptotic expression, which can be simplified under an orthogonality assumption. Some examples, in the context of generalized linear models and of robust inference, are provided.

Keywords: Asymptotics; bootstrap; nuisance parameter; profile and modified profile likelihood; robustness.

1 Introduction

Consider a sample $y = (y_1, \ldots, y_n)$ of n independent observations with distribution function $F(y; \theta)$, $\theta \in \Theta \subseteq \mathrm{I\!R}^p$, $p > 1$. Let θ be partitioned as $\theta = (\tau, \lambda)$, into a scalar parameter of interest τ and a $(p-1)$-dimensional nuisance parameter λ. Inference about τ only, based on the observation of y, is a widely encountered problem.

The common approach for classical parametric inference about τ is to resort to the profile loglikelihood for τ, given by

$$\ell_P(\tau) = \ell(\tau, \hat{\lambda}_\tau) = \sum_{i=1}^{n} \ell(\tau, \hat{\lambda}_\tau; y_i) \ , \tag{1}$$

where $\ell(\theta) = \ell(\tau, \lambda)$ denotes the loglikelihood function for θ and $\hat{\lambda}_\tau$ is the maximum likelihood estimate (MLE) of λ for fixed τ. Function (1) is then treated as an ordinary likelihood for inference about τ. However, it is well-known that nuisance parameters may cause difficulties for inference on τ based on (1), particularly when the dimension of λ is large relative to the sample size n. Several examples, such as the estimation of the variance in a normal-theory linear model, suggest that the bias of the profile score as an estimating function for τ can be substantial. To avoid these drawbacks, a general approach is to consider modifications of (1), which attempt to adjust it for nuisance parameters. Various modifications of the profile likelihood have been proposed over the past twenty years; see, for instance, Barndorff-Nielsen and Cox (1994, Chap. 8) and Severini (2000, Chap. 9). All these adjustments are equivalent to second order and share the common feature of reducing the score bias to $O(n^{-1})$ (DiCiccio *et al.*, 1996).

Reduction of the score bias is the key basic motivation for adjusting the profile loglikelihood in McCullagh and Tibshirani (1990). Their goal is to modify (1) so that the mean of the score function is zero and the variance of the score function equals its negative expected derivative matrix. The resulting estimating function has an improved asymptotic behaviour.

Complementary to likelihood-based procedures, in many situations of practical interest it can be preferable to base inference on estimating equations. This is true, for example, in the context of robustness theory when stability with respect to small deviations from the assumed model is required (see Hampel *et al.*, 1986), or in the context of generalized linear models with overdispersion or random effects (see McCullagh and Nelder, 1989). For inference about τ, extending (1) in the estimating functions setting, a quasi-profile loglikelihood function can be defined, with the standard limiting behaviour (Barndorff-Nielsen, 1995; Fraser *et al.*, 1997; Adimari and Ventura, 2002). This approach leads to a profile-type estimating function for τ that has bias of the same order as the profile score function. Several methods to modify an estimating function for the parameter of interest and reduce its bias have been proposed. They include Severini (2002), Wang and Hanfelt (2003) and Jørgensen and Knudsen (2004). However, all these authors focus only on the reduction of the bias of the profile estimating function and do not consider likelihood-based procedures associated to them.

The primary goal of this paper is to discuss modifications of the quasi-profile loglikelihood. The aim is to adjust the estimating function for τ, when λ is substituted with a suitable estimate, so that it is both unbiased and information unbiased. Two methods for the calculation of the adjustments, one exact and the other approximate, are discussed. The first method uses parametric bootstrap to estimate the moments of the estimating function for τ. The estimated moments are then used to centre and rescale the estimating function for τ. The approximate adjustment is based on first-order asymptotic expressions for the moments of the derivatives of the estimating function for τ. In both these cases, the modified quasi-profile loglikelihood is then given by the integral of the adjusted estimating function and the result aims at correcting the quasi-profile loglikelihood in a manner similar to the ordinary modified profile loglikelihood. On the practical side, the use of adjusted quasi-likelihood ratio statistics may lead to coverage probabilities more accurate than those pertaining to Wald-type or score-type confidence intervals. This was noted by several authors for likelihood-based procedures in important classes of models; see Hanfelt and Liang

(1995, 1998) and references therein.

The outline of the paper is as follows. Section 2 introduces unbiased estimating equations and quasi-profile loglikelihoods. Section 3 discusses the exact modification of the quasi-profile loglikelihood, while the approximate adjustments are given in Section 4. Section 5 and 6 discuss some examples and some final remarks, respectively.

2 Background theory

Let $\tilde{\theta}$ be an estimator for θ defined as a root of the unbiased estimating equation $\Psi(y;\theta) = \sum_{i=1}^{n} \psi(y_i;\theta) = 0$, where $\psi : \mathcal{Y} \times \Theta \to \mathrm{I\!R}$ is a given function, such that $E_\theta\{\Psi(Y;\theta)\} = 0$. In the following, we shall write Ψ_θ and ψ_θ for $\Psi(y;\theta)$ and $\psi(y;\theta)$, respectively. Under broad conditions assumed throughout this paper (e.g., Barndorff-Nielsen and Cox, 1994, Sec. 9.2), $\tilde{\theta}$ is consistent and asymptotically normal with mean θ and variance $V(\theta) = M(\theta)^{-1}\Omega(\theta)M(\theta)^{-\top}$, where $M(\theta) = -E_\theta(\Psi_{\theta/\theta})$, $\Omega(\theta) = Var_\theta(\Psi_\theta) = E_\theta(\Psi_\theta \Psi_\theta^\top)$, and the symbol / as a subscript indicates differentiation. Large-sample tests and confidence regions for θ can be constructed in a standard way using a consistent estimate of $V(\theta)$.

When θ is partitioned as $\theta = (\tau, \lambda)$, the estimating equation Ψ_θ is similarly partitioned as $(\Psi_\tau, \Psi_\lambda)$, where $\Psi_\tau = \Psi_\tau(y;\theta)$ and $\Psi_\lambda = \Psi_\lambda(y;\theta)$ are the estimating functions corresponding to τ and λ, respectively. This means that, for instance, if λ is known, Ψ_τ may be used as an estimating function for τ. Let $\tilde{\lambda}_\tau$ be the estimator derived from Ψ_λ when τ is considered as known, i.e. the solution of $\Psi_\lambda(y;\tau,\tilde{\lambda}_\tau) = 0$.

When inference about τ based on a pseudo-likelihood function is desired, a quasi-profile loglikelihood for τ can be considered (see Adimari and Ventura, 2002), given by

$$\ell_{QP}(\tau) = \int_{\tau_0}^{\tau} w(t, \tilde{\lambda}_t)\Psi_\tau(y; t, \tilde{\lambda}_t)\, dt \ , \tag{2}$$

where τ_0 is an arbitrary constant and, using index notation,

$$w(\tau, \lambda) = \frac{-\nu_{\tau\tau} - \kappa^{ba}\nu_{\tau a}\nu_{b\tau}}{E_\theta(\Psi_\tau^2) + 2\nu_{\tau a}\kappa^{ba}E_\theta(\Psi_\tau\Psi_b) + \nu_{\tau a}\nu_{\tau b}\kappa^{ca}\kappa^{db}E_\theta(\Psi_c\Psi_d)} \ . \tag{3}$$

In the former expression, the components of λ and Ψ_λ are denoted by λ^a and Ψ_a, respectively, and the derivatives of Ψ_τ and Ψ_a with respect to the components of λ are $\Psi_{\tau/a} = (\partial/\partial\lambda^a)\Psi_\tau$, $\Psi_{\tau/ab} = (\partial^2/\partial\lambda^a\partial\lambda^b)\Psi_\tau$, $\Psi_{a/b} = (\partial/\partial\lambda^b)\Psi_a$, and $\Psi_{a/bc} = (\partial^2/\partial\lambda^b\partial\lambda^c)\Psi_a$, etc, where $a, b, c = 1, \ldots, p-1$. Moreover, $\nu_{\tau a} = E_\theta(\Psi_{\tau/a})$, $\nu_{\tau ab} = E_\theta(\Psi_{\tau/ab})$, $\nu_{ab} = E_\theta(\Psi_{a/b})$ and $\nu_{abc} = E_\theta(\Psi_{a/bc})$, etc, and κ^{ab} is the inverse matrix of $-\nu_{ab}$. The scale adjustment $w(\tau, \lambda)$ is obtained so that the rescaled profile estimating equation $w(\tau, \tilde{\lambda}_\tau)\Psi_\tau(\tau, \tilde{\lambda}_\tau)$ has bias and information bias of order $O(1)$, as for the ordinary profile score function. It must be noted that by multiplying Ψ_τ by the factor $w(\tau, \lambda)$, the estimator $\tilde{\tau}$ of τ does not change. In practice, the scale adjustment (3) is necessary to obtain quasi-profile loglikelihood-type tests based on (2) with the classical asymptotic distribution.

The quasi-profile loglikelihood (2) has properties similar to the ordinary profile loglikelihood $\ell_P(\tau)$. In particular, for setting quasi-likelihood confidence regions or for testing hypotheses, the quasi-likelihood ratio

$$W_{QP}(\tau) = 2\{l_{QP}(\tilde{\tau}) - l_{QP}(\tau)\} \tag{4}$$

may be used. Under the null hypothesis and usual regularity conditions, $W_{QP}(\tau)$ is approximately χ^2_1 distributed.

We note here that Hanfelt and Liang (1995) provided a quasi-likelihood ratio test statistic having the standard χ^2_1 asymptotic distribution, of the form

$$W_{HL}(\tau) = 2\xi(\tilde{\tau}, \tilde{\lambda})\,\{Q(\tilde{\tau}) - Q(\tau)\}\,, \tag{5}$$

where $Q(\tau) = \int_{\tau_0}^{\tau} \Psi_\tau(y; t, \tilde{\lambda}_t)\, dt$ and $\xi(\tau, \lambda)$ is a suitable function. The two versions (4) and (5) are actually asymptotically equivalent. Indeed, note that (2) can be recast in the asymptotically equivalent form $\ell_{QP}(\tau) = w(\tilde{\tau}, \tilde{\lambda}) \int_{\tau_0}^{\tau} \Psi_\tau(y; t, \tilde{\lambda}_t)\, dt$. From the latter expression, we obtain a quasi-likelihood ratio test of the form (5), with $w(\tilde{\tau}, \tilde{\lambda})$ playing the role of $\xi(\tilde{\tau}, \tilde{\lambda})$.

3 A modification of the quasi-profile likelihood

Since λ has to be estimated, similarly to the usual profile loglikelihood function, the quasi-profile loglikelihood (2) does not behave exactly like an ordinary loglikelihood. In particular, in small samples, $\ell_{QP}(\tau)$ does not take properly into account sampling variability of $\tilde{\lambda}_\tau$. For the usual profile loglikelihood $\ell_P(\tau)$ various modifications have been proposed (see, e.g., Barndorff-Nielsen and Cox, 1994, Chap. 8, and Severini, 2000, Chap. 9), leading to modified profile loglikelihoods of the form

$$\ell_{MP}(\tau) = \ell_P(\tau) + M(\tau)\ , \tag{6}$$

where $M(\tau)$ is a suitable smooth function having derivatives of order $O_p(1)$.

Here we discuss modifications of $\ell_{QP}(\tau)$, motivated as in McCullagh and Tibshirani (1990). The modified quasi-profile loglikelihood for τ is given by the integral of the adjusted estimating equation for τ. Both exact and approximate methods for the calculation of the adjustments are discussed. The exact calculation presented in this section is achieved through a simulation process in which the moments of the estimating function for τ are estimated by parametric bootstrap sampling. The approximate adjustment discussed in the next section uses instead first-order analytical approximations to the moments of the derivatives of the estimating function for τ.

A basic property of an ordinary score function is that its mean is zero and its variance is minus the expected derivative matrix, expectations being computed at the true parameter value. Our interest is to adjust $\Psi_\tau(\tau, \tilde{\lambda}_\tau)$ so that these properties hold when expectations and derivatives are computed at $(\tau, \tilde{\lambda}_\tau)$, rather than at the true parameter point. Consider the functions $\mu(\tau, \tilde{\lambda}_\tau)$ and $\omega(\tau, \tilde{\lambda}_\tau)$ and let $\Psi^\dagger_\tau(\tau, \tilde{\lambda}_\tau) = \Psi^\dagger_\tau(y; \tau, \tilde{\lambda}_\tau)$ be equal to

$$\Psi^\dagger_\tau(\tau, \tilde{\lambda}_\tau) = \omega(\tau, \tilde{\lambda}_\tau)\,\{\Psi_\tau(y; \tau, \tilde{\lambda}_\tau) - \mu(\tau, \tilde{\lambda}_\tau)\}\,.$$

As in McCullagh as Tibshirani (1990), we require that

$$E_{(\tau,\tilde{\lambda}_\tau)}\{\Psi^\dagger_\tau(\tau,\tilde{\lambda}_\tau)\} = 0 \text{ and } Var_{(\tau,\tilde{\lambda}_\tau)}\{\Psi^\dagger_\tau(\tau,\tilde{\lambda}_\tau)\} = -E_{(\tau,\tilde{\lambda}_\tau)}\left\{\frac{\partial\Psi^\dagger_\tau(\tau,\tilde{\lambda}_\tau)}{\partial\tau}\right\}\,, \quad (7)$$

with $E_{(\tau,\tilde{\lambda}_\tau)}(\cdot)$ denoting $E_\theta(\cdot)|_{\lambda=\tilde{\lambda}_\tau}$. Solving for $\mu(\tau,\tilde{\lambda}_\tau)$ and $\omega(\tau,\tilde{\lambda}_\tau)$, we find

$$\begin{aligned}\mu(\tau,\tilde{\lambda}_\tau) &= E_{(\tau,\tilde{\lambda}_\tau)}\{\Psi_\tau(\tau,\tilde{\lambda}_\tau)\} & (8)\\ \omega(\tau,\tilde{\lambda}_\tau) &= \left[\frac{\partial}{\partial\tau}\mu(\tau,\tilde{\lambda}_\tau) - E_{(\tau,\tilde{\lambda}_\tau)}\{\Psi_{\tau/\tau}(\tau,\tilde{\lambda}_\tau)\}\right]/Var_{(\tau,\tilde{\lambda}_\tau)}\{\Psi_\tau(\tau,\tilde{\lambda}_\tau)\}\ . & (9)\end{aligned}$$

Thus a modified quasi-profile loglikelihood for τ is given by

$$\ell_{MQP}(\tau) = \int_{\tau_0}^{\tau}\Psi^\dagger_\tau(y;t,\tilde{\lambda}_t)\,dt = \int_c^\tau \omega(\tau,\tilde{\lambda}_t)\,\{\Psi_\tau(y;t,\tilde{\lambda}_t) - \mu(\tau,\tilde{\lambda}_t)\}dt\ . \quad (10)$$

The required quantities for the exact computation of $\ell_{MQP}(\tau)$ involve expectations evaluated at $(\tau,\tilde{\lambda}_\tau)$, and can be computed analytically only in very simple special cases. This was already noted by McCullagh and Tibshirani (1990) for the likelihood-based case, and for general estimating functions the problems to be faced are not likely to be easier. When the required expectations cannot be computed analytically, we must resort to Monte Carlo simulation. Even if the computation of the adjustment for many realistic models may seen rather complicated, it can be implemented in modern statistical environments, such as R. The computation requires an algorithm similar to that used by McCullagh and Tibshirani (1990). In particular, a suitable grid of Q values for τ is considered, and at each point of the grid τ_q, $q = 1,\ldots,Q$, the values of $\mu(\tau_q,\tilde{\lambda}_{\tau_q})$ and $\omega(\tau_q,\tilde{\lambda}_{\tau_q})$ are estimated by parametric boostrap of B samples generated under the model $F(y;\tau_q,\tilde{\lambda}_{\tau_q})$.

4 A first-order approximation to the modification

A first-order approximation to $\ell_{MQP}(\tau)$ can be obtained using results given in the Appendix of Adimari and Ventura (2002); see also Severini (2002). Taking termwise expectation of a Taylor expansion for $\Psi_\tau(\tau,\tilde{\lambda}_\tau)$ about the true parameter value, an expansion for the additive adjustment (8) can be obtained from $E_\theta\{\Psi_\tau(\tau,\tilde{\lambda}_\tau)\} = m(\tau,\lambda) + O(n^{-1})$, where $m(\tau,\lambda)$ is of order $O(1)$ and is given by

$$\begin{aligned}m(\tau,\lambda) &= \kappa^{ba}E_\theta(\Psi_b\Psi_{\tau a}) + \nu_{\tau a}\kappa^{ca}\kappa^{db}E_\theta(\Psi_d\Psi_{cb})\\ &+ \frac{1}{2}\nu_{\tau a}\nu_{dbc}\kappa^{da}\kappa^{eb}\kappa^{fc}E_\theta(\Psi_e\Psi_f) + \frac{1}{2}\nu_{\tau ab}\kappa^{da}\kappa^{cb}E_\theta(\Psi_d\Psi_c)\ . \quad (11)\end{aligned}$$

The first-order bias correction (11) involves only the first two derivatives with respect to λ of (Ψ_τ,Ψ_λ). There is a formal similarity between equation (11) and the expression for the bias of the ordinary profile score function given in McCullagh and Tibshirani (1990).

The expansion for the scale adjustment (9) is less straightforward. However, from Taylor expansions it turns out that it can be approximated with error $O(n^{-1})$

by (3). For a single parameter of interest, we find that the adjusted score function has the form

$$w(\tau, \tilde{\lambda}_\tau) \left\{ \Psi_\tau(y; \tau, \tilde{\lambda}_\tau) - m(\tau, \tilde{\lambda}_\tau) \right\} \ , \tag{12}$$

so that an approximate modified quasi-profile loglikelihood, aiming at correcting the score bias, is given by

$$\begin{aligned} \ell_{AQP}(\tau) &= \int_{\tau_0}^{\tau} w(t, \tilde{\lambda}_t) \left\{ \Psi_\tau(y; t, \tilde{\lambda}_t) - m(t, \tilde{\lambda}_t) \right\} dt \\ &= \ell_{QP}(\tau) - \int_{\tau_0}^{\tau} w(t, \tilde{\lambda}_t) m(t, \tilde{\lambda}_t) \, dt \ . \end{aligned} \tag{13}$$

The adjusted quasi-profile score function (12) has score bias which vanishes asymptotically. There is again a close connection between (13) and the proposal by McCullagh and Tibshirani (1990); see the similarities between their $\ell_{AP}(\tau)$ and (13).

There are situations where the use of $\ell_{AQP}(\tau)$ becomes compelling. In fact, the scale and additive modifications of (12) can be simplified under the conditions of nuisance parameter insensitivity of Ψ_τ or of G-orthogonality (Godambe and Thompson, 1989, Godambe, 1991, and Jørgensen and Knudsen, 2004), which are suitable extensions for estimating functions of the orthogonality condition, studied by Cox and Reid (1987) in the likelihood-based framework. In particular, the condition of λ-insensitivity (Jørgensen and Knudsen, 2004) is $E_\theta(\Psi_\tau \, \ell_\lambda) = 0$, where ℓ_λ is the score function for λ. The condition of λ-insensitivity of Ψ_τ has several consequences, and in particular it implies that

$$\nu_{\tau a} = 0 \, , \qquad a = 1, \ldots, p-1 \, . \tag{14}$$

Under conditions (14), the expressions for (3), (11) and (13) can be simplified. In particular, we obtain $w(\tau, \lambda) = -\nu_{\tau\tau} \left\{ E_\theta(\Psi_\tau^2) \right\}^{-1}$ and $m(\tau, \lambda) = \kappa^{ba} E_\theta(\Psi_b \Psi_{\tau a}) + \frac{1}{2} \nu_{\tau ab} \kappa^{da} \kappa^{cb} E_\theta(\Psi_d \Psi_c)$. Note that the latter expression $m(\tau, \lambda)$ equals the expression (27) of Jørgensen and Knudsen (2004).

5 Examples

The theory presented in the previous sections applies to any kind of estimating functions, provided that some regularity conditions hold. In the computation of both $\ell_{MQP}(\tau)$ and $\ell_{AQP}(\tau)$ there is a clear dependence on the model assumptions, as we need either to specify the distribution of y for the resampling or to be able to compute moments of the estimating functions and its derivatives. In the following we will focus on two situations where $\ell_{MQP}(\tau)$ and $\ell_{AQP}(\tau)$ are useful.

(i) In the context of generalized linear models with overdispersion or random effects (see McCullagh and Nelder, 1989). In this case, there is a model for the data, but maximum likelihood is not straightforward, and it may be helpful to resort to some estimating equations.

(ii) In the context of robustness theory (*e.g.* Hampel *et al.*, 1986), when M-estimators are considered and stability with respect to small deviations from the assumed model is required. In this case, there is a reasonable model for the bulk of the data, but we wish to protect the inferential conclusions from model misspecification or from the presence of few outlying observations. In this class of situations, it is sensible to use robustified versions of the likelihood score equations as estimating functions, but it makes sense to compute the higher-order adjustment at the true model.

Example 1: Overdispersion in count data. Let us consider Poisson regression for count data. The responses y_i are realizations of independent Poisson random variables with mean $\mu_i = \exp(x_i^\mathsf{T}\beta)$, $\beta \in \mathrm{I\!R}^k, k \geq 1$, $i = 1, \ldots, n$. In many applications with discrete data, overdispersion can be encountered. This means that more variability then would be expected from the Poisson model is observed, and it has to be taken into account.

We focus on two cases. In the first, the variance is assumed to be a quadratic function of μ_i of the form $Var_\theta(Y_i) = \mu_i(1+\alpha\mu_i)$, whereas in the second the variance is proportional to the mean, $Var_\theta(Y_i) = \mu_i\,(1+\alpha)$. In either case, the estimating function for β is the score function from the Poisson likelihood

$$\Psi_\beta(y;\beta) = \sum_{i=1}^{n} (y_i - \mu_i)\, x_i^\mathsf{T} \ , \tag{15}$$

which still provides an unbiased estimating equation. An estimating function for α can be obtained from the method of moments, as shown in Lawless (1987). With quadratic variance function we get

$$\Psi_\alpha(y;\beta,\alpha) = \sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{\mu_i(1+\alpha\mu_i)} - (n-k)\,, \tag{16}$$

whereas, with linear variance function

$$\Psi_\alpha(y;\beta,\alpha) = \sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{\mu_i(1+\alpha)} - (n-k)\ . \tag{17}$$

The use of $\ell_{MQP}(\tau)$ allows us to quantify the consequences of overdispersion for inference on a regression coefficient. Since the assumptions on the variance are typical of a negative binomial model for the response, this distribution can be used for computing the adjustments. We apply this procedure to the Ames Salmonella data, already analysed by Lawless (1987). The response is the number of revertant colonies observed on a plate, and covariates are based on the dose level of quinoline on the plate (x). We assume the following model for the response

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 \log(x_i + 10)\,,\ i = 1, \ldots, 18\,,$$

where the interest lies on $\tau = \beta_2$. Figure 1 compares the normed modified quasi profile loglikelihoods $\ell_{MQP}(\tau) - \ell_{MQP}(\tilde{\tau})$ for the three models considered, namely Poisson and negative binomial with the two different variance functions. In all cases, the boostrap computation was applied with B=1,000 repetitions.

[Figure 1 about here.]

We note that the adjustment for the Negative Binomial case is much larger than for the Poisson case, while the choice of the variance function has a limited influence. In this example the modified quasi-likelihood function provides a useful display of the effect of accounting for overdispersion.

Example 2. Quasi-likelihood estimation in GLMMs

Consider a generalized linear mixed model obtained by adding random effects to the linear predictors for clustered data. Here we consider a random intercepts model for independent clusters, with linear predictor $\eta_{ij}^u = x_{ij}^\intercal \beta + u_i$, where $i = 1, \ldots, n$ is the number of clusters, $j = 1, \ldots, m_i$ the number of observations in the i-th cluster, x_{ij} a vector of covariates, and the cluster-specific intercepts u_i are independent normally distributed random effects, $u_i \sim N(0, \sigma^2)$. The model is completed by assuming a link function g, such that $g(E_\theta[y_{ij}|u_i, x_{ij}]) = \eta_{ij}^u$. The estimation of $\theta = (\beta, \sigma)$ is usually hampered by the necessity of integrating out the random effects, hence estimation methods other than maximum likelihood are often considered; McCulloch and Searle (2001) provide a comprehensive survey of the subject.

Here we consider the estimation approach based on quasi-likelihood, as proposed by McCullagh and Nelder (1989, Sec. 14.4). The method is based on the unconditional mean and variance of Y, $E_\theta(Y) = \mu(\theta)$ and $Var_\theta(Y) = V(\theta)$. If the random effect variance σ^2 is known, then the quasi-likelihood estimating function for β is given by

$$\Psi_\beta(y; \beta, \sigma) = D(\theta)^\intercal \, V(\theta)^{-1} \, \{y - \mu(\theta)\} \,, \tag{18}$$

where $D(\theta) = \partial\mu(\theta)/\partial\beta$. Sutradhar and Rao (2003) show that, for given values of σ, the resulting estimator for β is quite efficient. For the estimation of σ it is necessary to use a supplementary estimating equation. Following McCullagh and Nelder (1989), we may use a moment approach based on equating a quadratic form to its expected value, namely

$$\Psi_\sigma(y; \beta, \sigma) = \{y - \mu(\theta)\}^\intercal \, P \, \{y - \mu(\theta)\} - \mathrm{tr}\{P \, V(\theta)\} \,, \tag{19}$$

where P is a suitable matrix. Notice that Sutradhar and Rao (2003) show that (19) can be quite inefficient, but it is very simple as, like (18), it requires only the computation of $\mu(\theta)$ and $V(\theta)$. If the interest lies on a scalar component of β, the method proposed in this paper is an appealing choice, allowing us to recover the possible inaccuracy of (19).

As an illustrative example we consider data from Beitler and Landis (1985) on a multicentre trial to compare the efficacy of two topical cream preparations. Here we fit a probit model with random effects for the different clinics, namely $\eta_{ij}^u = \beta_0 + \beta_1 \, t_{ij} + u_i$, where t_{ij} is a binary treatment indicator. The inferential interest lies on the treatment effect, thus $\tau = \beta_1$. The estimating equations (18) and (19) represent a convenient choice, as $\mu(\theta)$ and $V(\theta)$ are not difficult to obtain with probit link, for example by using the same approach of Drum and McCullagh (1993). In (19) we set P equal to a block-diagonal matrix with blocks corresponding to the different clinics and $P_i = 1_i \, 1_i^\intercal$, where 1_i is a m_i-dimensional vector of ones. The resulting estimate of τ is hardly significant at the 0.05 level, since $\tilde{\tau}$=0.428, with s.e.($\tilde{\tau}$)=0.229. There is a clear amount of between-clinic heterogeneity as $\tilde{\sigma}$=0.752. For comparison, we obtained the MLEs by integrating out the random effects using Monte-Carlo integration (see McCulloch and Searle, 2001, Sec. 10.3). MLEs are not so different, with $\hat{\tau}$=0.439 and $\hat{\sigma}$=0.813. However, s.e.($\hat{\tau}$)=0.175 and the treatment effect appears much more significant, with a two-sided P-value decreasing from 0.06 to about 0.01.

The difference between the two results requires further investigation, and $\ell_{MQP}(\tau)$ seems quite helpful. We applied the boostrap method described in Section 3 with B=500 simulations, obtaining the plot of $\ell_{MQP}(\tau)$ of Figure 2. Clearly, the point $\tau = 0$ is not supported by $\ell_{MQP}(\tau)$.

[Figure 2 about here.]

The 95% confidence interval for τ derived from $\ell_{MQP}(\tau)$ is (0.039, 0.808), which is slightly larger of that derived from the profile likelihood, (0.097, 0.787). Note that the computation of MLE is too burdensome to obtain an approximation to $\ell_{MP}(\tau)$ of Section 3; see (6). However, a simple run of parametric bootstrap of 1,000 repetitions obtained by setting $\theta = \hat{\theta}$ gives a studentised bootstrap 95% confidence interval based on the MLE equal to (0.082, 0.804), even closer to the interval from $\ell_{MQP}(\tau)$. As studentised bootstrap confidence intervals are second order accurate (Davison and Hinkley, 1997, Chapter 5), this shows that there is little need to adjust the results from the profile likelihood.

To wind up, in this example $\ell_{MQP}(\tau)$ based on simple estimating equations provides results comparable to more demanding likelihood-based methods, recovering the inaccuracy of the Wald statistic.

Example 3. Bounded-influence inference in linear regression.

Let y_i, $i = 1, \ldots, n$, be independent observations such that $y_i = x_i^{\mathsf{T}} \beta + \varepsilon_i$, where β is a k-dimensional vector of coefficients and ε_i has mean 0 and variance σ^2.

Let us consider Mallows's M-estimator for β, with the scale parameter estimated by weighted Huber's Proposal 2 (see Marazzi, 1993, Sec. 2.1). Hence, if $r_i = (y_i - x_i^{\mathsf{T}} \beta)/\sigma$ are the standardized residuals, the estimating functions for $\theta = (\beta, \sigma)$ have the form

$$\begin{aligned} \Psi_\beta(y;\beta,\sigma) &= \frac{1}{\sigma}\sum_{i=1}^n \psi_{HF}(r_i;c_1)\, w_i\, x_i^{\mathsf{T}}\,, \\ \Psi_\sigma(y;\beta,\sigma) &= \sum_{i=1}^n \psi_{HF}^2(r_i;c_2)\, w_i - (n-k)\,\gamma\,, \end{aligned} \tag{20}$$

where $\psi_{HF}(\cdot;c)$ is the Huber function, c_1, c_2 are tuning constants, and γ is a suitable constant such that the solution $\tilde{\sigma}$ is asymptotically consistent at the normal model. The weights w_i are chosen to get a bounded influence function for the resulting estimator.

Suppose we are interested in drawing inference about a regression coefficient. For illustrative purposes, let us consider the mortality data already analysed by several authors, including Krasker and Welch (1982). The data consist of $n = 60$ observations about age-adjusted mortality in several cities in the U.S., with some available covariates. The data present several high-leverage points, hence bounded-influence estimation seems a sensible choice. Let us focus in particular on the coefficient of the covariate EDUC. After setting $w_i = \sqrt{1 - h_i}$, where h_i is the i-th diagonal element of the hat matrix, $c_1 = 1.345$ and $c_2 = 1.5$, the Mallows estimate of τ is $\tilde{\tau} = -13.17$ (5.90), not so different from the least squares estimate (OLS), $\hat{\tau} = -13.30$ (6.97). The difference between standard errors is due to the different estimate of the scale, since $\tilde{\sigma} = 30.50$ and $\hat{\sigma} = 36.39$. Now we can perform a simple sensitivity analysis on the results, as often done in robust statistics. For instance, let us take a single point (observation 37) and change the values of its response from 1113.16 to 1000. As shown in Figure 3, the change in the response seems relatively modest. Note that observation 37 is an influential point, and actually the change causes a large change in its value of the Cook's distance.

[Figure 3 about here.]

Hence, the change causes a large variation in the OLS estimates, and actually now $\hat{\tau} = -8.75$ (6.20), while the variation is much smaller for the bounded-influence estimator, as $\tilde{\tau} = -11.76$ (5.62). The same occurs to test statistics. Figure 4 shows the change occurred to $W_{MQP}(\tau)$ and to the loglikelihood ratio test for τ derived from the profile loglikelihood.

[Figure 4 about here.]

The different degree of sensitivity to a single change in the response is striking, and totally in favor of bounded-influence estimation. The modified quasi-profile likelihood allows a more precise and stable inference.

Example 4: Stratified linear model. Let us consider a normal linear regression model with stratum nuisance parameters, of the form $y_{ij} = \lambda_i + \tau x_{ij} + \varepsilon_{ij}$, where $i = 1, \ldots, q$ and $j = 1, \ldots, m$. We consider Huber estimation of the regression coefficients, with scale parameter estimated by Huber's Proposal 2; see (20) with $w_i = 1$. We set $c_1 = 1$ and $c_2 = 1.345$.

The bias adjustment for the profile estimating function for the parameter of interest τ is null and $\ell_{AQP}(\tau) = \ell_{QP}(\tau)$. In Table 3.5, the performance of the directed quasi-profile likelihood, i.e. $r_{MQP}(\tau) = \text{sgn}(\tilde{\tau} - \tau)\sqrt{W_{MQP}(\tau)}$, with $w(\tau, \lambda)$ evaluated by 100 bootstrap samples, is compared with its analytical counterpart, $r_{AQP}(\tau)$, and the Wald statistic. The numerical study was carried out under four error distributions, chosen to represent different departures from normality (see Ronchetti *et al.*, 1997). These distributions are: (i) **e1**: standard normal; (ii) **e2**: 93% from a standard normal and 7% from a normal with $\sigma = 5$; (iii) **e3**: slash distribution; (iv) **e4**: 90% from a standard normal and 10% from a normal with $\mu = 30$. The table is based on $5,000$ simulations for each setting.

[Table 1 about here.]

The results indicate that there $r_{MQP}(\tau)$ has a reasonably accurate behavior and tends to improve slightly on $r_{AQP}(\tau)$. The Wald statistic works well too, but it seems generally less accurate than the other two.

6 Final remarks

This paper discusses adjustments for a quasi-profile likelihood for a parameter of interest, paralleling the classical likelihood-based approach. The adjustments alleviate some of the problems inherent to the presence of nuisance parameters. Other modifications of the estimating function for the parameter of interest have been proposed in the literature (see Severini, 2002, and Wang and Hanfelt, 2003). However, their aim is only to reduce the bias of the profile estimating function and do not consider quasi-likelihood procedures associated to them. The use of quasi-profile likelihood ratio statistics may lead to improved coverage probabilities with respect to Wald-type or score-type confidence intervals in many cases.

Moreover, the possibility of representing graphically the quasi-profile likelihood is a good point of the method, reducing the gap between maximum likelihood estimation and the estimating equation approach. In fact, the method presented in this paper allows us to supplement the plot of the quasi-profile likelihood with its adjusted counterparts, thus visualizing the effect of the nuisance parameter estimation and the correction for the small-sample bias of the profile estimating equations.

If τ has $r > 1$ components, then the modified quasi-profile loglikelihood (10) can be generalized in a straightforward manner, but its computation will be more burdensome. Aside from computational aspects, the main difficulty is that when $r > 1$ a modified quasi-profile loglikelihood for τ of the form (10) does not exist in general. A necessary and sufficient condition for the existence of $\ell_{MQP}(\tau)$ is that the matrix $\tilde{\Psi}^{\dagger}_{\tau/\tau}$ be symmetric. Alternatively, a possible solution is to operate in a componentwise fashion, considering each component of τ separately.

Finally, we note that he theory developed in this paper may be useful with sparse data structures, like the stratified settings considered in Wang and Hanfelt (2003). Sartori (2003) studied thoroughly the behavior of the modified profile likelihood for stratified data. It would be interesting to extend the study to the modified quasi-profile likelihood.

References

Adimari, G., Ventura, L. (2002). Quasi-profile loglikelihood for unbiased estimating functions. *Ann. Inst. Statist. Math.*, **54**, 235–244.

Barndorff-Nielsen, O.E (1995). Quasi profile and directed likelihoods from estimating functions. *Ann. Inst. Statist. Math.*, **47**, 461–464.

Barndorff-Nielsen, O. E., Cox, D. R. (1994). *Inference and Asymptotics.* Chapman and Hall, London.

Beitler, P.J., Landis, J. R. (1985). A mixed-effects model for categorical data. *Biometrics*, **41**, 991–1000.

Cox, D. R., Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc.* B, **49**, 1–39.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge.

DiCiccio, T. J., Martin, M. A., Stern, S. E., Young, G. A. (1996). Information bias and adjusted profile likelihoods. *J. R. Statist. Soc.* B, **58**, 189–203.

Drum, M. L., McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics*, **49**, 667–689.

Fraser, D. A. S., Reid, N., Wu, J. (1997). Estimating functions and higher order significance. *Selected Proceedings of the Symposium on Estimating Functions*, IMS Lecture Notes Monograph Series, **32**, 105–114.

Godambe, V. P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika*, **78**, 143–151.

Godambe, V. P., Thompson, M. E. (1989). An extension of quasi-likelihood (with discussion). *J. Statist. Plann. Inference* , **2**, 568–571.

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

Hanfelt, J. J., Liang, K-. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461–477.

Hanfelt, J. J., Liang, K-. Y. (1998). Inference for odds ratio regression models with sparse dependent data. *Biometrics*, **54**, 136–147.

Jørgensen, B., Knudsen, S. J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scand. J. Statist.*, **31**, 93–114.

Krasker, W. S., Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association*, **77** , 595–604.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 9–225.

Marazzi, A. (1993). *Algorithms, Routines and S Functions for Robust Statistics.* Pacific Grove, California: Wadsworth and Brooks/Cole.

McCullagh, P., Nelder, J.A. (1989), *Generalized linear models*, Chapman and Hall, London.

McCullagh, P., Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc.* B **52**, 325–344.

McCulloch, C. E., Searle, S. R. (2001). *Generalized, Linear, and Mixed Models.* Wiley, New York.

Ronchetti, E., Field, C., Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, **92**, 1017–1023.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, **90**, 533–549.

Severini, T. A. (2000). *Likelihoods Methods in Statistics.* Oxford University Press, Oxford.

Severini, T. A. (2002). Modified estimating functions. *Biometrika*, **89**, 333–343.

Sutradhar, B. C, Rao, R. P. (2003). On quasi-likelihood inference in generalized linear mixed models with two components of dispersion. *The Canadian Journal of Statistics*, **31**, 415–435.

Wang, M., Hanfelt, J. J. (2003). Adjusted profile estimating function. *Biometrika*, **90**, 845–858.

Acknowledgements

We thank Luigi Pace for helpful comments. This work was supported by MIUR COFIN 2004-2005.