ESTIMATING FUNCTIONS AND HIGHER ORDER SIGNIFICANCE

D.A.S. Fraser, N. Reid, and Jianrong Wu
Department of Statistics
University of Toronto
Toronto, Canada

ABSTRACT

Estimating functions provide inference methods based on a model for specified functions of the response variable, such as the mean and variance. The inference methods can use the limiting distribution of the estimating function, or the derived limiting distribution of the estimating function roots, or the derived limiting distribution of the quasi-likelihood function. In fully specified parametric models more accurate inference can be obtained by using recently developed higher order approximations based on likelihood asymptotics. We consider the recent third order methods in the estimating function context using the quasi-likelihood function. We focus on inference for a scalar parameter of interest: profile quasi-likelihood for the scalar parameter is defined and we describe a method for using this to approximate significance probabilities.

Key Words: Ancillary directions; Asymptotic inference; Estimating functions; profile likelihood; Quasi-likelihood; Tail area approximations.

1 Introduction

In a fully specified parametric model, quantities for testing a parameter or parameter component can be constructed from the score function, the maximum likelihood estimator, or the likelihood ratio statistic. These quantities are equivalent in first order asymptotic theory, although examples tend to indicate that the likelihood ratio statistic provides the most reliable assessment, and the score statistic the least reliable assessment of a parameter of interest. It is also possible by considering higher order asymptotics to derive a modification of the likelihood ratio statistic with much better inferential properties.

In the estimating equations context, optimally weighted estimating functions serve as an equivalent to, or extension of, the score function, and the quasi-likelihood function serves as an equivalent to, or extension of, the likelihood function. Our goal in this paper is to explore the extension of likelihood asymptotics to the estimating equation context. We assume that the parameter of interest is a scalar, as this is central to the methods of likelihood asymptotics.

In Section 2 we summarize the main results for third order inference based on the recent likelihood based asymptotics, with emphasis on the approach developed in Fraser and Reid (1995). In Section 3 we discuss the application of these methods to the estimating equation context. Some examples are discussed in Section 4 and limitations of the current work and possible extensions are outlined in Section 5.

2 Likelihood and significance

We assume in this section that we have a log-likelihood function $\ell(\theta) = \ell(\theta; y)$, based on a continuous variable y with n coordinates, and that the parameter is partitioned as $\theta = (\lambda, \psi)$, with ψ a scalar parameter of interest and λ a vector of $p-1$ nuisance parameters. We will denote the observed information function $-\ell''(\theta)$ by $j(\theta)$ and the nuisance parameter submatrix by $j_{\lambda\lambda}(\theta)$. The profile log-likelihood function is $\ell_P(\psi) = \ell(\hat{\lambda}_\psi, \psi)$, where $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ with ψ fixed. We will often write $\hat{\theta}_\psi$ for $(\hat{\lambda}_\psi, \psi)$, and $\ell(\hat{\theta}_\psi)$ for the profile log-likelihood function.

In fairly wide generality, an approximate p-value for testing the hypothesis $H_0 : \psi = \psi_0$ can be computed from either of the following formulas:

$$\Phi_1(r, q) = \Phi(r) + \phi(r)(r^{-1} - q^{-1}) \ , \ \Phi_2(r, q) = \Phi\{r - r^{-1} \log(r/q)\} \tag{1}$$

where ϕ and Φ are the standard normal density and distribution functions. These are approximations to the distribution function of r, typically with relative error $O(n^{-3/2})$. The first formula, an extension of the Lugannani and Rice (1980) approximation, often gives better accuracy with exponential models, while the second formula, due to Barndorff-Nielsen (1986, 1991) avoids anomalous values outside $[0, 1]$. In (1) the quantity $r = r(\psi_0)$ is usually the signed square root of the profile log-likelihood ratio statistic:

$$r = r(\psi) = \mathrm{sgn}(\hat{\psi} - \psi) \cdot [2\{\ell(\hat{\theta}; y) - \ell(\hat{\theta}_\psi; y)\}]^{\frac{1}{2}} \tag{2}$$

and q is a complementary first order quantity, the explicit form of which is determined by the problem.

For example, in the case of a canonical exponential family model with no nuisance parameters, the first version of (1) is the Lugannani and Rice (1980) approximation, with q taken to be the standardized maximum likelihood departure for the canonical parameter $(\hat{\psi} - \psi)\{j(\hat{\psi})\}^{1/2}$. In the general one parameter setting q can be taken to be

$$\{\ell_{;y}(\hat{\psi}) - \ell_{;y}(\psi)\}\{j(\hat{\psi})\}^{1/2}\{\ell_{\psi;y}(\hat{\psi})\}^{-1} \tag{3}$$

where it is assumed that there is a one-to-one transformation from y to $(\hat{\psi}, a)$, with a exactly or approximately ancillary, and $\ell_{;y}(\psi; y) = \partial\ell(\psi; y)/\partial y$ and $\ell_{\psi;y}(\psi; y) = \partial\ell_{;y}(\psi; y)/\partial\psi$, with a held fixed.

In the presence of nuisance parameters a general formula for q was established in Barndorff-Nielsen (1986, 1991) under the assumption that there is available an explicit exact or approximate ancillary statistic a such that the conditioned variable given a is a one to one function of $\hat{\theta}$. In the special case of a canonical exponential family $f(y;\theta) = \exp\{\psi s + \lambda^T t - c(\lambda, \psi) - d(y)\}$, the minimal sufficient statistic is a one-to-one function of $\hat{\theta}$ and the expression for q can be simplified to

$$(\hat{\psi} - \psi)\{-\ell_P''(\hat{\psi})\}^{1/2}\rho^{-1}(\psi, \hat{\psi})$$

where $\rho^2(\psi, \hat{\psi}) = |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|/|j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|$.

Equation (1) with (2) also gives an approximation to the Bayes posterior marginal cumulative distribution function with the choice

$$q = -\ell_P'(\psi)\{-\ell_P''(\hat{\psi})\}^{-1/2}\rho(\psi, \hat{\psi})\frac{\pi(\hat{\lambda}, \hat{\psi})}{\pi(\hat{\lambda}_\psi, \psi)}.$$

An alternative expression for q was derived in Fraser and Reid (1995) which provides a more easily implemented calculation for the case with effective variable of the same dimension as the parameter as covered by Barndorff-Nielsen (1986, 1991) and also handles the general case where the dimension of the effective variable is larger than the dimension of θ. The dimension reduction from y to θ is effected by constructing a new parametrization φ, which is obtained as the gradient of the log likelihood function at the observed data taken in p directions $V = (v_1 \dots v_p)$.

$$\begin{aligned} \varphi^T(\theta) &= \frac{\partial}{\partial y^T}\ell(\theta; y)\Big|_{y^0} \cdot V = \ell_{;y}(\theta; y^0) \cdot V = \ell_{;V}(\theta; y^0); \\ V &= \frac{\partial y}{\partial \theta^T}\Big|_{(y^0, \hat{\theta}^0)} . \end{aligned} \tag{4}$$

The latter differentiation is for fixed values of appropriate pivotal quantities, as discussed in Fraser and Reid (1995).

The quantity q can be viewed as a standardized maximum likelihood departure

$$q = q(\psi) = \operatorname{sgn}(\hat{\psi} - \psi) \cdot |\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)|\{|j_{(\theta\theta)}(\hat{\theta})|/|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|\}^{\frac{1}{2}} \tag{5}$$

where $j_{(\theta\theta)}(\hat{\theta})$ is the observed full information matrix and $j_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is the nuisance information matrix, and both are recalibrated in terms of the new parameterization $\varphi(\theta)$: see eq. (17) below. The scalar parameter $\chi(\theta)$ replacing $\psi(\theta)$ is linear in $\varphi(\theta)$

$$\chi(\theta) = \frac{\psi_{\varphi^T}(\hat{\theta}_\psi)}{|\psi_{\varphi^T}(\hat{\theta}_\psi)|}\varphi(\theta) . \tag{6}$$

3 Quasi-likelihood and significance

The quasi-score estimating function (Wedderburn, 1974) is denoted here by u_θ to suggest its nominal role as a derivative with respect to θ,

$$u_\theta(\theta; y) = \mu_\theta^T(\theta)\Sigma^{-1}(\theta)\{y - \mu(\theta)\} \ . \tag{7}$$

This is based on n observations y with mean $\mu(\theta)$, variance matrix $\Sigma(\theta)$, and location gradient $\mu_\theta(\theta) = (\partial/\partial\theta^T)\mu(\theta)$. The more general optimally weighted estimating function (McCullagh & Nelder, 1989)

$$u_\theta(\theta; y) = \mu_\theta^T(\theta)\Sigma^{-1}(\theta)d(y;\theta) \tag{8}$$

is based on a vector $d(y;\theta)$ recording some version of departure of y from θ with mean $E\{d(y;\theta);\theta\} = 0$, variance matrix $\Sigma(\theta)$, and location gradient $\mu_\theta(\theta) = E\{(\partial/\partial\theta)\, d(y;\theta);\theta\}$. A further extension to handle conditional means and variances given a conditioning quantity $A(\theta)$ is discussed in Hanfelt and Liang (1995).

Under reasonable conditions a root $\hat\theta$ of the quasi-score estimating equation $u_\theta(\theta; y) = 0$ is asymptotically normal with mean θ and variance $I_{\theta\theta}^{-1}(\theta)$ where

$$I_{\theta\theta}(\theta) = \mu_\theta^T(\theta)\Sigma^{-1}(\theta)\mu_\theta(\theta) \tag{9}$$

is the variance of the estimating function. A quasi-likelihood ratio for θ_2 versus θ_1 is obtained (Wedderburn, 1974) as a line integral

$$Q(\theta_2, \theta_1) = \int_{\theta_1}^{\theta_2} u_\theta^T(\theta; y)d\theta \ \ : \tag{10}$$

this will in general be path dependent if the quasi-score does not form an integrable vector field.

We examine quasi-likelihood as constructed from estimating equations (7) and (8). With θ expressed as (λ, ψ), the equations have coordinates corresponding to λ and ψ:

$$u_\theta(\theta, y) = \begin{pmatrix} u_\lambda(\theta; y) \\ u_\psi(\theta; y) \end{pmatrix} \ . \tag{11}$$

The roots $\hat\theta$ and $\hat\theta_\psi$ of the estimating equations will typically be obtained by iterative solution of $u_\theta = 0$ and $u_\lambda = 0$, for example as

$$\begin{aligned} \hat\theta^{(i+1)} &= \hat\theta^{(i)} + I_{\theta\theta}^{-1}(\hat\theta^{(i)})u_\theta(\hat\theta^{(i)}; y) \\ \hat\lambda_\psi^{(i+1)} &= \hat\lambda_\psi^{(i)} + I_{\lambda\lambda}^{-1}(\hat\theta_\psi^{(i)})u_\lambda(\hat\lambda_\psi^{(i)}, \psi; y) \end{aligned} \tag{12}$$

Now consider the interest parameter ψ. As a definition of the corresponding profile log quasi-likelihood we take $\ell_{PQ}(\psi)$ to be

$$\ell_{PQ}(\psi_0) = \int_{(\hat\lambda,\hat\psi)}^{(\hat\lambda_{\psi_0},\psi_0)} u_\psi(\hat\lambda_\psi, \psi; y) d\psi \; ; \tag{13}$$

where the effective integration curve $C = \{\hat\theta_\psi\} = \{(\hat\lambda_\psi, \psi)\}$ is along the path of constrained solutions $\hat\theta_\psi$ as suggested in Barndorff-Nielsen (1995). It would typically be calculated iteratively from the overall solution $\hat\theta$. The signed likelihood root is then given uniquely by

$$r(\psi) = \mathrm{sgn}(\hat\psi - \psi)[2\{\ell_{PQ}(\hat\psi) - \ell_{PQ}(\psi)\}]^{\frac{1}{2}} \tag{14}$$

The nominal reparameterization $\varphi(\theta)$ is needed only along the profile curve C and is obtained by an integral paralleling (13); it uses an $n \times p$ matrix of ancillary directions $V = (v_1 \ldots v_p)$ which will be discussed later in this section. Again integrating along C, we define

$$\varphi(\hat\lambda_{\psi_0}, \psi_0) = \int_{(\hat\lambda,\hat\psi)}^{(\hat\lambda_{\psi_0},\psi_0)} V^T \Sigma^{-1}(\theta) \mu_\theta(\theta) d\theta$$

in the estimating equation case (7), and

$$\varphi(\hat\lambda_{\psi_0}, \psi_0) = \int_{(\hat\lambda,\hat\psi)}^{(\hat\lambda_{\psi_0},\psi_0)} V^T d_y^T(y;\theta) \Sigma^{-1}(\theta) \mu_\theta(\theta) d\theta \tag{15}$$

in the more general case (8) with the notation

$$d_y(y;\theta) = (\partial/\partial y^T) d(y;\theta) \; .$$

This involves an integration for each of p coordinates but most of the calculations are common and related to (13).

The Jacobian of the parameter change from θ to φ,

$$\varphi_{\theta^T}(\theta) = V^T d_y^T(y;\theta) \Sigma^{-1}(\theta) \mu_\theta(\theta) \; , \tag{16}$$

is needed only at $\hat\theta$ and $\hat\theta_\psi$ and is a by product of the calculation for (15). The inverse of $\varphi_{\theta^T}(\hat\theta_\psi)$ gives the coefficients $\psi_{\varphi^T}(\hat\theta_\psi)$ for the linear parameter $\chi(\theta)$ defined by (6).

The nominal information matrices $j_{\theta\theta}(\hat\theta)$ and $j_{\lambda\lambda}(\hat\theta_\psi)$ are calculated from gradients of the quasi-scores

$$j_{\theta\theta}(\theta) = -\frac{\partial}{\partial\theta^T} u_\theta(\theta) \; , \; j_{\lambda\lambda}(\theta) = -\frac{\partial}{\partial\lambda^T} u_\lambda(\theta) \; .$$

The recalibrated information matrices for use in (7) are then obtained using (16):

$$\begin{aligned} |j_{(\theta\theta)}(\hat\theta)| &= |j_{\theta\theta}(\hat\theta)||\varphi_{\theta^T}(\hat\theta)|^{-2} \; , \\ |j_{(\lambda\lambda)}(\hat\theta_\psi)| &= |j_{\lambda\lambda}(\hat\theta_\psi)||\varphi_\lambda^T(\hat\theta_\psi)\varphi_{\lambda^T}(\hat\theta_\psi)|^{-1} \; . \end{aligned} \tag{17}$$

We now have all the ingredients for using the third order formula (1) with (2), (5), and (6) except the $n \times p$ matrix of ancillary directions $V = (v_1, \ldots, v_p)$ at the observed data point.

In most inference contexts the number of variables n will exceed the number of parameters p and some procedure is needed to effectively reduce the number of variables to p. Higher order asymptotics indicates that the appropriate procedure is to condition on an ancillary of dimension $n-p$, thus giving p free variables.

The higher order approximation described in Section 2 needs only the value of the likelihood function at the observed data point, and the gradient of the likelihood in p directions tangent to the ancillary, which give the new parameter φ in (4). The only information needed concerning the ancillary is thus the array V of tangent directions. For third order inference the vectors V need to be tangent to just a second order ancillary (Fraser and Reid, 1995; Skovgaard, 1986) and for second order inference the vectors V need to be tangent to just a first order ancillary. In the estimating equations context where the model is specifed as $Ey_i = \mu_i$, $\text{var} y_i = \Sigma(\mu_i)$, and the components are independent, a first order ancillary can be derived using results from curved exponential family theory (Amari, 1985). The resulting directions in this case are given by

$$V = \mu_{\theta^T}(\hat{\theta}) = \frac{\partial}{\partial \theta} \mu(\theta)|_{\hat{\theta}}. \tag{18}$$

In the case of the more general estimating equation (8), a separate argument is needed to establish V.

4 Examples

As a first example we consider a mean and variance function corresponding to exponential regression: the coordinates y_i have mean μ_i and variance μ_i^2 with $\mu_i = \exp\{\alpha + \beta(x_i - \bar{x})\}$. We assume interest centers on the regression parameter β. The corresponding estimating equations for α and β are

$$\begin{aligned} u_\alpha &= \Sigma \mu_i^{-1}(y_i - \mu_i) \\ u_\beta &= \Sigma (x_i - \bar{x}) \mu_i^{-1}(y_i - \mu_i) \end{aligned} \tag{19}$$

These integrate on a path-free basis to give the quasi-likelihood

$$\ell(\alpha, \beta) = -\Sigma \exp\{-\alpha - \beta(x_i - \bar{x})\} y_i - n\alpha$$

which in fact coincides with the actual likelihood for the exponential regression model. The corresponding profile quasi-likelihood for β thus coincides with the ordinary profile

$$\ell_P(\beta) = \Sigma y_i \hat{\mu}_i^{-1} + n\hat{\alpha} - \Sigma y_i \hat{\mu}_{i\beta}^{-1} - n\hat{\alpha}_\beta$$

where $\hat{\mu}_i = \exp\{\hat{\alpha} + \hat{\beta}(x_i - \bar{x})\}$ and $\hat{\mu}_{i\beta} = \exp\{\hat{\alpha}_\beta + \beta(x_i - \bar{x})\}$.

For this example the ith row of the matrix V of ancillary directions (18) for second order inference is

$$V_i = (\hat{\mu}_i^0, \hat{\mu}_i^0(x_i - \bar{x})).$$

The ancillary directions for third order inference are also available (Fraser et al., 1994): the ith row of V is

$$V_i = (y_i^0, y_i^0(x_i - \bar{x})).$$

The implied statistical model has in fact location model structure with an exact ancillary so exact p-values are available for comparison.

We consider a random sample of size 5 from the Fiegl and Zelen leukemia data given as Set U in Cox and Snell (1981). Table 1 shows the exact and quasi-likelihood based p-values for selected values of β, the regression coefficient. The full sample size is $n = 17$, and for the full sample there is almost no difference among the first order, quasi-likelihood and third order methods. The first order method here refers to using the normal approximation for the profile log-likelihood root.

Table 1. Exact and approximate p-values: exponential regression.

β	first order	using (18)	exact
-4.7	0.9954	0.9952	0.9960
-4.3	0.9899	0.9896	0.9911
-3.8	0.9747	0.9747	0.9776
-3.4	0.9505	0.9514	0.9564
-2.9	0.8956	0.8994	0.8953
1.5	0.0177	0.0240	0.0249
1.0	0.0370	0.0466	0.0513
2.0	0.0096	0.0134	0.0109

Our second example is a binomial regression model: we assume that $Ey_i = \mu_i$, $\text{var} y_i = m_i\mu_i(1-\mu_i)$, with m_i known and $\mu_i = \alpha + \beta x_i$. We are using a non-canonical parametrization, as there is no exact conditional method for inference about β in this setting. For illustration we fit this model to data from Example 1.5 of Cox and Snell (1989). The data values are

i	x_i	y_i	n_i
1	1.0	4	110
2	1.7	4	105
3	2.2	2	62
4	2.8	1	65
5	4.0	1	45

As y is discrete the arguments of Section 2 do not apply directly, but the ancillary directions given by (18) are easily computable. Table 2 compares the significance functions for the first order and higher order approximations

Table 2. First order and second order p-values: binomial example.

β	first order	using(18)
-0.0245	0.9953	0.9939
-0.0225	0.9894	0.9867
-0.02	0.9735	0.9678
-0.018	0.9487	0.9398
-0.0155	0.8948	0.8826
-0.0115	0.7454	0.7375
-0.0065	0.4956	0.5061
0.000	0.2413	0.2587
0.0055	0.0956	0.1074
0.009	0.0512	0.0591
0.0125	0.0259	0.0307
0.017	0.0010	0.0122
0.0195	0.0056	0.0071

In both these examples the quasi-likelihood obtained from integrating the score equation is identical to the log-likelihood. It is necessary to introduce some dependence in the score equations, along the lines of Liang and Zeger's generalized estimating equations, or to use an optimally weighted estimating function as at (8), to derive a quasi-likelihood that is not also a log-likelihood. Unfortunately, we have not as yet determined a way to compute the ancillary directions V for these more general problems. Further comment on this point is given in the next section.

5 Discussion

In the application of formula (1) the choice of q seems to be crucial, and seems to need an argument based on approximate ancillarity in the nominal model leading to the estimating equation. For example, Hanfelt and Liang (1995) give an example using a moment type estimating equation for the shape parameter of a gamma distribution, and compare the first order approximations based on the quasi-likelihood ratio statistic and the standardized root of the estimating equation. We tried combining the two statistics, essentially using the standardized root as the q in (1), but the approximation was worse than either of the first order approximations.

The derivation of V above uses the fact that μ_i is a location parameter, and we do not at the moment have an argument that applies to non-location parameters, such as over-dispersion parameters or additional dependence parameters in Σ. It may be possible to incorporate over-dispersion parameters using Nelder and Pregibon's (1987) extended quasi-likelihood, which essentially gives a REML-type marginal likelihood for the scale parameters. Alternatively it might be possible to substitute a consistent estimate of the scale parameter into the profile log quasi-likelihood, and still have an improved approximation, but we have not investigated this.

The recent third order asymptotic methods address various model features not handled by typical first order methods:

(i) Third and fourth order moments of the distributions of the component variables are implicitly involved.

(ii) The coordinates means $\mu_i(\theta)$ can measure θ and thus ψ on differing measurement scales.

(iii) Non linearity in the parameter of interest as a function of θ or $\mu(\theta)$ is allowed for.

While the first of these was the initial stimulus for recent asymptotics, (ii) and (iii) are perhaps of greater importance.

In the estimating equation context, (ii) and (iii) are of particular interest while information for (i) is generally not available. Accordingly we have adapted the asymptotic methods to the estimating equation context primarily to handle the complications (ii) and (iii). However concerning (i) we note that the application of the asymptotic methods starts with the quasi likelihood and we feel it is appropriate to make the most accurate extraction of significance probabilities from that likelihood.

References

Amari, S.-I. (1985) *Differential Geometric Methods in Statistics.* New York: Springer-Verlag.

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–322.

Barndorff-Nielsen, O.E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557—64.

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

Cox, D.R. and Snell, E.J. (1981). *Applied Statistics*. London: Chapman and Hall.

Cox, D.R. and Snell, E.J. (1989). *The Analysis of Binary Data*. London: Chapman and Hall.

Fraser, D.A.S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica* **47**, 33–53.

Fraser, D.A.S., Monette, G., Ng, K.W., and Wong, A. (1994). Higher order approximations with generalized linear models. *Multivariate Analysis and its Applications*. eds. T.W. Anderson, K.T. Fang and I. Olkin. IMS Lecture Notes Monograph Series, **24**, 253–262.

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

Lugannani, R. and Rice, S.O. (1980). Saddlepoint approximation for the distribution of the sums of independent random variables. *Adv. Appl. Prob.* **12**, 475–490.

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

Nelder, J.A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221–232.

Skovgaard, I.M. (1986). Successive improvements of the order of ancillarity. *Biometrika* **74**, 516–519.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **82**, 439–447.

Acknowledgements

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada.