Statistica Sinica **14**(2004), 349-360

CONDITIONAL LIKELIHOOD INFERENCE IN GENERALIZED LINEAR MIXED MODELS

N. Sartori and T. A. Severini

University of Padova and Northwestern University

Abstract: Consider a generalized linear model with a canonical link function, containing both fixed and random effects. In this paper, we consider inference about the fixed effects based on a conditional likelihood function. It is shown that this conditional likelihood function is valid for any distribution of the random effects and, hence, the resulting inferences about the fixed effects are insensitive to misspecification of the random effects distribution. Inferences based on the conditional likelihood are compared to those based on the likelihood function of the mixed effects model.

Key words and phrases: Conditional likelihood, exponential family, incidental parameters, random effects, variance components.

1. Introduction

The addition of random effects to a generalized linear model substantially increases the usefulness of such models; however, such an increase comes at a cost. To obtain the likelihood function of the model, we must average over the random effects. In many cases, the resulting integral does not have a closed form expression and, even when one is available, the simple structure of a fixed-effects generalized linear model is generally lost. Furthermore, the resulting inferences may be sensitive to the assumption regarding the random effects distribution (Neuhaus, Hauck and Kalbfleisch (1992)), a choice that is often difficult to verify.

Let y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$, denote independent scalar random variables such that y_{ij} follows an exponential family distribution with canonical parameter θ_{ij}, $\theta_{ij} = x_{ij}\beta + z_{ij}\gamma$ where x_{ij} and z_{ij} are known covariate vectors, β is a parameter vector representing the fixed effects and γ is a vector random variable representing the random effects. We assume that the distribution of γ is known, except for an unknown parameter η.

Consider inference about the fixed effects parameter β. If γ is fixed, rather than random, then the loglikelihood function is of the form

$$\sum_{i,j}\{y_{ij}x_{ij}\beta + y_{ij}z_{ij}\gamma - k(x_{ij}\beta + z_{ij}\gamma)\},$$

where $k(\cdot)$ denotes the cumulant function of the exponential family distribution. In this case, it is well-known that inference about β in the presence of γ may be based on the conditional distribution of the data given the statistic $s = \sum_{i,j} y_{ij} z_{ij}$, which depends only on β. See, e.g., Diggle, Heagerty, Liang and Zeger (2002, Section 9.2.1).

Although this conditional approach is typically used when γ is fixed, the same approach may be used in the model in which γ is random. Let $p(y|\gamma;\beta)$ and $p(s|\gamma;\beta)$ denote the density functions of y and s, respectively, in the model with γ held fixed, and let $\bar{p}(y;\beta,\eta)$ and $\bar{p}(s;\beta,\eta)$ denote the density functions of y and s, respectively, in the random effects model with γ removed by integration with respect to the random effects density $h(\gamma;\eta)$. For example,

$$\bar{p}(y;\beta,\eta) = \int p(y|\gamma;\beta)h(\gamma;\eta)d\gamma.$$

Likelihood inference in the mixed effects model is based on $\bar{L}(\beta,\eta) = \bar{p}(y;\beta,\eta)$, which we call the *integrated likelihood*.

The density $\bar{p}(y;\beta,\eta)$ may also be used to form a conditional likelihood. Let $\bar{p}(y|s;\beta,\eta)$ denote the density of y given s based on $\bar{p}(y;\beta,\eta)$. In Section 2, it is shown that $\bar{p}(y|s;\beta,\eta)$ depends only on β so that likelihood inference for β may be based on the corresponding conditional likelihood. Furthermore, it is shown that $\bar{p}(y|s;\beta) = p(y|s;\beta)$; hence, the conditional likelihood based on $\bar{p}(y|s;\beta)$ does not depend on the specification of $h(\gamma;\eta)$. The purpose of this paper is to consider the properties of the conditional likelihood in the random effects model; that is, we consider conditional likelihood inference under the assumption that γ is a random rather than a fixed effect, as is done, e.g., in Diggle, Heagerty, Liang and Zeger (2002, Chap. 9).

Although inference for β may be based on $\bar{p}(y|s;\beta)$, clearly this conditional density is not useful for inference regarding η, the parameter of the random effects distribution; inference regarding η can be carried out using standard methods (see, e.g., Diggle, Heagerty, Liang and Zeger (2002, Chap. 9)). Thus, conditional inference in the mixed-effects model essentially uses a fixed-effects-model approach to inference regarding β, while inference regarding γ is based on the assumption that γ is random. That is, the conditional approach in the mixed effects model is a hybrid between fixed-effects and mixed-effects methods.

In Section 2 the properties of the conditional likelihood function for β are considered and an approximation to the conditional likelihood is presented. In Section 3 the conditional likelihood is compared to the integrated likelihood for β. Sections 2 and 3 consider models in which any possible dispersion parameter is known; in Section 4 we consider models containing an unknown dispersion parameter. Section 5 contains a numerical example.

Many different approaches to inference in generalized linear mixed models have been considered; these approaches generally include some method of avoiding the integration needed to compute the integrated likelihood. See, for example, Schall (1991), Breslow and Clayton (1993), McGilchrist (1994), Engel and Keen (1994) and Lee and Nelder (1996). For inference about population-averaged quantities, the generalized estimating equation approach of Liang and Zeger (1986) may be used. Davison (1988) considers inference based on conditional likelihoods in generalized linear models with fixed effects only; in some sense, the present paper may be viewed as an extension of Davison's work to mixed models. Breslow and Day (1980) use conditional likelihood methods for inference in a mixed effects model for binary data. Another approach to inference in mixed models is to use Bayesian methods; see, for example, Zeger and Karim (1991), Draper (1995) and Gelman, Carlin, Stern and Rubin (1995).

The mixed models considered here are closely related to mixture models in which the random effects distribution is treated as an unknown mixture distribution. Conditional likelihood methods are often used for inference in these models; see, for example, Basawa (1981), Lindsay (1983, 1995), van der Vaart (1988) and Lindsay, Clogg and Grego (1991).

2. Conditional Likelihood

Since $p(y|\gamma;\beta) = p(y|s;\beta)p(s|\gamma;\beta)$, we have that

$$\begin{aligned}\bar{p}(y;\beta,\eta) &= \int p(y|\gamma;\beta)h(\gamma;\eta)d\gamma = \int p(y|s;\beta)p(s|\gamma;\beta)h(\gamma;\eta)d\gamma \\ &= p(y|s;\beta)\bar{p}(s;\beta,\eta).\end{aligned}$$

Hence,

$$\bar{p}(y|s;\beta,\eta) = \frac{\bar{p}(y;\beta,\eta)}{\bar{p}(s;\beta,\eta)} = \frac{p(y|s;\beta)\bar{p}(s;\beta,\eta)}{\bar{p}(s;\beta,\eta)} = p(y|s;\beta).$$

Therefore, the conditional likelihood based on $\bar{p}(y|s;\beta)$ is the the same as that based on $p(y|s;\beta)$ and does not depend on the choice of h. Furthermore, since the conditional likelihood is a genuine likelihood function for β, its properties are not affected by the dimension of γ.

Example 1. *Poisson regression*

Let y_{ij}, $j = 1,\ldots,n_i$, $i = 1,\ldots,m$, denote independent Poisson random variables such that y_{ij} has mean $\exp\{x_{ij}\beta + \gamma_i\}$. The conditional density of the data given $\gamma = (\gamma_1,\ldots,\gamma_m)$ is given by

$$p(y|\gamma;\beta) = \frac{\exp\{\sum_{i,j} y_{ij}x_{ij}\beta + \sum_i \gamma_i y_i - \sum_{i,j}\exp(x_{ij}\beta + \gamma_i)\}}{\prod_{i,j} y_{ij}!},$$

where $y_i = \sum_j y_{ij}$. Hence, in the model with γ held fixed, $(y_1, \ldots, y_m)$ is sufficient for fixed β and the conditional likelihood function is given by

$$\frac{\exp(\sum_{i,j} y_{ij} x_{ij} \beta)}{\prod_i \{\sum_j \exp(x_{ij}\beta)\}^{y_i}}. \tag{1}$$

Now consider a distribution for the random effects. Suppose that $\exp\{\gamma_1\}, \ldots, \exp\{\gamma_m\}$ are independent random variables, each with an exponential distribution with mean η. Then

$$\bar{p}(y; \beta, \eta) = \exp\{\sum_{i,j} y_{ij} x_{ij} \beta\} \prod_i \eta^{y_i} \frac{\Gamma(y_i + 1)}{\{\eta \sum_j \exp(x_{ij}\beta) + 1\}^{y_i+1}} \prod_{i,j} \frac{1}{y_{ij}!}.$$

Clearly, $(y_1, \ldots, y_m)$ is sufficient for η; i.e., it is sufficient in the model with β taken to be known. Given γ, $y_1, \ldots, y_m$ are independent Poisson random variables with means $\exp\{\gamma_i\} \sum_j \exp\{x_{ij}\beta\}$, $i = 1, \ldots, m$, respectively. Hence, the marginal density of y_i is

$$\{\sum_j \exp(x_{ij}\beta)\}^{y_i} \prod_i \frac{\Gamma(y_i + 1)}{\{\eta \sum_j \exp(x_{ij}\beta) + 1\}^{y_i+1}} \frac{1}{y_i!}$$

and the conditional likelihood given $y_1, \ldots, y_n$ is identical to (1). The argument given earlier in this section shows that the same result holds for any random effects distribution.

Some functions of β may not be identifiable based on the conditional distribution given s. Let X denote the $n \times p$ matrix, $n = \sum n_i$, $p = \dim(\beta)$, given by

$$X = M(x_{ij}) \equiv (x_{11}^\top \; x_{12}^\top \; \cdots \; x_{1n_1}^\top \; \cdots \; x_{m1}^\top \; x_{m2}^\top \; \cdots \; x_{mn_m}^\top)^\top;$$

similarly, let $Z = M(z_{ij})$ and $y = M(y_{ij})$ so that Z is $n \times q$, $q = \dim(\gamma)$ and y is $n \times 1$. The sufficient statistic in the full model is given by $(X^\top y, Z^\top y)$ and the conditioning statistic s is equivalent to $Z^\top y$. Hence, if there exists a vector b such that $Xb = Za$ for some vector a, the corresponding linear function of β will not be identifiable in the conditional model. Therefore we assume that $Xb = Za$ only if a and b are both zero vectors, so that the entire vector β is identifiable in the conditional model. If, for a given model, this condition is not satisfied, the results based on the conditional likelihood given below may be interpreted as applying only to those components of β that are identifiable in the conditional model.

Those linear functions of β not identifiable in the conditional model are also not identifiable in the model with γ treated as fixed effects. However, they may be identifiable in the model with γ taken to be random effects; in this case, those parameters may be viewed as being parameters of the random effects distribution and, hence, inferences regarding those parameters may be particularly sensitive to assumptions regarding the random effects distribution.

Example 2. *Poisson regression* (continued).

Suppose that $x_{ij} = 1$ for all i, j. Then $(y_1, \ldots, y_m)$ is still sufficient in the model with β held fixed; however, $(y_1, \ldots, y_m)$ is also sufficient in the general model with parameter $(\beta, \gamma_1, \ldots, \gamma_m)$ so that the conditional likelihood does not depend on β.

The identifiability of β in the mixed model depends on the distribution of the random effects. In the model with γ held fixed, the likelihood function may be written $\exp[\sum_i \{(\gamma_i + \beta) y_i - n_i \exp(\beta + \gamma_i)\}]$. Hence, $\gamma_i + \beta$, $i = 1, \ldots, m$ may be viewed as the random effects and β is therefore a parameter of the random effects distribution. If $\exp\{\gamma_1\}, \ldots, \exp\{\gamma_m\}$ are taken to be independent exponential random variables with mean η, then $\exp\{\gamma_1 + \beta\}, \ldots, \exp\{\gamma_m + \beta\}$ are independent exponential random variables with mean $\eta \exp(\beta)$ so that β is not identifiable in this model.

However, if $\exp\{\gamma_1\}, \ldots, \exp\{\gamma_m\}$ are independent random variables each with a gamma distribution with mean 1 and variance $1/\eta$, then the integrated likelihood is given by

$$\exp\{\sum_i y_i \beta\} \frac{\eta^{m\eta}}{\Gamma(\eta)^m} \prod_i \frac{\Gamma(\eta + y_i)}{(\eta + n_i \exp\{\beta\})^{\eta + y_i}}$$

and β is identifiable.

If exact computation of the conditional likelihood is difficult, an approximation may be used. Using a saddlepoint approximation (e.g., Daniels (1954) and Jensen (1995)) to the density of s, an approximation to the conditional likelihood given $y_1, \ldots, y_m$ is given by

$$\hat{L}(\beta) = \Big|\{\sum_{i,j} z_{ij}^{\mathsf{T}} k''(x_{ij}\beta + z_{ij}\hat{\gamma}_\beta) z_{ij}\}\Big|^{\frac{1}{2}} \exp[\sum_{i,j} \{y_{ij} x_{ij} \beta + z_{ij}\hat{\gamma}_\beta - k(x_{ij}\beta + z_{ij}\hat{\gamma}_\beta)\}],$$

where $\hat{\gamma}_\beta$ is the maximum likelihood estimator of γ for fixed β. Note that since $\hat{L}(\beta)$ is based on a saddlepoint approximation to the density of s, this approximation does not depend on the choice of random effects density h; hence, $\hat{L}(\beta)$ is not related to Laplace approximations to the integrated likelihood (e.g., Breslow and Lin (1995) and Booth and Hobert (1998)). If the dimension of γ is fixed, the error of the approximation is $O(n^{-1})$. If m, the dimension of γ, increases with n, then the error is $o(1)$, provided that $m = o(n^{3/4})$; see Sartori (2003) for further details.

The approximation $\hat{L}(\beta)$ is identical to the one given by Davison (1988) for inference in a fixed-effects generalized linear model. Note that $\hat{L}(\beta) = |j_{\gamma\gamma}(\beta, \hat{\gamma}_\beta)|^{1/2} L_p(\beta)$ where $L_p(\beta)$ denotes the profile likelihood and $j_{\gamma\gamma}(\beta, \gamma)$ denotes the observed information for fixed β; in both cases, γ is treated as fixed effects. Hence, $\hat{L}(\beta)$ is also identical to the modified profile likelihood function (Barndorff-Nielsen (1980, 1983)). That is, the modified profile likelihood based on treating γ as a fixed effect is also valid if γ is modeled as a random effect.

Since ℓ_c is also a conditional loglikelihood in the model with parameters (β, η), under standard regularity conditions $\hat{\beta}$, the maximizer of ℓ_c, is asymptotically distributed according to a multivariate normal distribution (Andersen (1970)). The asymptotic covariance matrix of $\hat{\beta}$ may be estimated using $\hat{j}_c$, the observed information based on ℓ_c evaluated at $\hat{\beta}$. Furthermore, Andersen (1970) shows that the convergence of the normalized $\hat{\beta}$ to a normal distribution holds conditionally on γ. Hence, the asymptotic normality of $\hat{\beta}$ is valid for any random effects distribution.

A confidence region for β may be based on $W = 2\{\ell_c(\hat{\beta}) - \ell_c(\beta)\}$. Under standard conditions, W is asymptotically distributed according to a chi-squared distribution with p degrees-of-freedom (Andersen (1971)). As with the asymptotic normality of $\hat{\beta}$, this result holds conditionally on γ and, hence, the result is valid for any random effects distribution.

3. Relationship between the Conditional and Integrated Likelihoods

Let $\ell_c(\beta)$ denote the conditional loglikelihood for β and let $\bar{\ell}(\beta, \eta) = \log \bar{p}(y; \beta, \eta)$ denote the integrated loglikelihood based on a particular choice for the random effects distribution. Since $\bar{\ell}(\beta, \eta)$ depends on η and β, for inference about β, we may consider the profile integrated loglikelihood, $\bar{\ell}_p(\beta) = \bar{\ell}(\beta, \hat{\eta}_\beta)$; for instance, β may be estimated by maximizing $\bar{\ell}_p(\beta)$.

In general, $\bar{\ell}_p(\beta) - \ell_c(\beta) = \bar{\ell}_p(\beta; s)$ where $\bar{\ell}(\beta, \eta; s)$ denotes the integrated loglikelihood function based on the marginal distribution of s and $\bar{\ell}_p(\beta; s)$ is the corresponding profile loglikelihood function. Hence, the difference between $\ell_c(\beta)$ and $\bar{\ell}_p(\beta)$ depends on how $\bar{\ell}_p(\beta; s)$ varies with β. Since ℓ_c does not depend on the choice of h, the sensitivity of $\bar{\ell}_p(\beta)$ to choice of h is measured by the sensitivity of $\bar{\ell}_p(\beta; s)$ to the choice of h.

If $\bar{\ell}_p(\beta; s)$ does not depend on β, then $\bar{\ell}_p(\beta) = \ell_c(\beta)$. This occurs, e.g., if the statistic s is S-ancillary for β based on the density $\bar{p}(s; \beta, \eta)$ (Severini (2000, Section 9.2)). Recall that s is S-ancillary for β if, for each β_1, β_2, η_1, there exists η_2 such that

$$\int p(s|\gamma; \beta_2) h(\gamma; \eta_2) d\gamma = \int p(s|\gamma; \beta_1) h(\gamma; \eta_1) d\gamma;$$

this holds, in particular, if $\bar{p}(s;\beta,\eta)$ depends on (β,η) only through a function of lower dimension. Hence, this condition depends on both $p(s|\gamma;\beta)$ and $h(\gamma;\eta)$.

Example 3. *Matched pairs of Poisson random variables*

Consider the following special case of the Poisson regression model in which $n_i = 2$ for all i and $x_{ij} = 1$ if $j = 1$ and $x_{ij} = 0$ if $j = 2$. In this model, $s = (y_1,\ldots,y_m)$ where $y_1,\ldots,y_m$ are independent Poisson random variables such that y_i has mean $\omega_i T(\beta)$, $\omega_i = \exp(\gamma_i)$ and

$$T(\beta) = \sum_j \exp(x_{ij}\beta) = \exp(\beta) + 1.$$

Assume that $\omega_1,\ldots,\omega_m$ are independent identically distributed random variables and let $g(\cdot;\eta)$ denote the density of ω_i. Then

$$\bar{p}(s;\beta,\eta) = \prod_i \frac{1}{y_i!}\int \{\omega T(\beta)\}^{y_i}\exp\{-\omega T(\beta)\}g(\omega;\eta)d\omega.$$

If η is a scale parameter, then $g(\omega;\eta) = g(\omega/\eta;1)/\eta$ and

$$\bar{p}(s;\beta,\eta) = \prod_i \frac{1}{y_i!}\int \{(\omega/\eta)\eta T(\beta)\}^{y_i}\exp\{-(\omega/\eta)\eta T(\beta)\}g(\omega/\eta;1)/\eta d\omega.$$

Therefore, $\bar{p}(s;\beta,\eta)$ depends on (β,η) only through $\eta T(\beta)$ and, hence s is S-ancillary. Thus, in the two-sample model, any integrated likelihood function based on a scale model for the $\exp(\gamma_i)$ yields the same estimate of β and that estimate is identical to the one based on $\ell_c(\beta)$.

This same result holds in a general Poisson regression model provided that the design is balanced in the sense that x_{ij}, $j = 1,\ldots,n_i$, are the same for each i.

Exact agreement between $\ell_c(\beta)$ and $\ell_p(\beta)$ occurs only in exceptional cases. It is straightforward to show that the Laplace approximation to the integrated likelihood function is given by

$$\Big|\{\sum_{i,j} z_{ij}^{\mathsf{T}} k''(x_{ij}\beta + z_{ij}\hat{\gamma}_\beta)z_{ij}\}\Big|^{-\frac{1}{2}}\exp\{\sum_{i,j}[y_{ij}x_{ij}\beta + y_{ij}z_{ij}\hat{\gamma}_\beta - k(x_{ij}\beta + z_{ij}\hat{\gamma}_\beta)]\}h(\hat{\gamma}_\beta;\eta)$$

so that $\ell_p(\beta)$ may be approximated by

$$\hat{\ell}_c(\beta) - \log\Big|\{\sum_{i,j} z_{ij}^{\mathsf{T}} k''(x_{ij}\beta + z_{ij}\hat{\gamma}_\beta)z_{ij}\}\Big| + \log h(\hat{\gamma}_\beta;\tilde{\eta}_\beta),$$

where $\hat{\ell}_c(\beta)$ denotes the saddlepoint approximation to the conditional loglikelihood given in Section 2 and $\tilde{\eta}_\beta$ maximizes $h(\hat{\gamma}_\beta;\eta)$ with respect to η for fixed

β. When the dimension of γ is fixed, the relative error of this approximation is $O(n^{-1})$. In this case,

$$\frac{1}{\sqrt{n}}\bar{\ell}_p'(\beta) = \frac{1}{\sqrt{n}}\ell_c'(\beta) + O_p(\frac{1}{\sqrt{n}}). \tag{2}$$

That is, $\ell_c(\beta)$ provides a first-order approximation to $\bar{\ell}_p(\beta)$ based on any non-degenerate random effects distribution. It is important to note that the O_p term in this expression refers to the distribution of the data corresponding to the random effects distribution h.

The analysis above is based on the assumption that m, the dimension of γ, remains fixed as $n \to \infty$ and the conclusions do not necessarily hold when m increases with n. For instance, the saddlepoint approximation and Laplace approximation used are valid only when m grows very slowly with n, specifically when $m = o(n^{1/3})$ (Shun and McCullagh (1995) and Sartori (2003)).

Example 4. *Poisson regression* (continued)

Suppose $\bar{\ell}(\beta,\eta)$ is based on the assumption that $\exp(\gamma_1),\ldots,\exp(\gamma_m)$ are independent exponential random variables with mean η. It follows that

$$\bar{\ell}_p'(\beta) - \ell_c'(\beta) = \sum_i \frac{y_i - \hat{\eta}_\beta \sum_j \exp(x_{ij}\beta)}{\sum_j \exp(x_{ij}\beta)} \frac{\sum_j x_{ij}\exp(x_{ij}\beta)}{\hat{\eta}_\beta \sum_j \exp(x_{ij}\beta) + 1}.$$

For each $i = 1,\ldots,m$,

$$\frac{y_i - \hat{\eta}_\beta \sum_j \exp(x_{ij}\beta)}{\sum_j \exp(x_{ij}\beta)} = O_p(1) \quad \text{as} \quad n_i \to \infty.$$

Hence, under the assumption that each $n_i \to \infty$ while m stays fixed, $\bar{\ell}_p'(\beta)/\sqrt{n} = \ell_c'(\beta)/\sqrt{n} + O_p(1/\sqrt{n})$, in agreement with the general result given above.

Now suppose the n_i remain fixed while $m \to \infty$. Since $\hat{\eta}_\beta = \eta + O_p(n^{-\frac{1}{2}})$,

$$\frac{y_i - \hat{\eta}_\beta \sum_j \exp(x_{ij}\beta)}{\sum_j \exp(x_{ij}\beta)} = \frac{y_i - \eta \sum_j \exp(x_{ij}\beta)}{\sum_j \exp(x_{ij}\beta)} + O_p(n^{-1}),$$

and, hence,

$$\sum_{i=1}^m \frac{y_i - \hat{\eta}_\beta \sum_j \exp(x_{ij}\beta)}{\sum_j \exp(x_{ij}\beta)} = O_p(\sqrt{m}).$$

It follows that, in this case, $\bar{\ell}_p'(\beta)/\sqrt{n} = \ell_c'(\beta)/\sqrt{n} + O_p(1)$, as described above.

For cases in which ℓ_c and $\bar{\ell}_p$ lead to different estimators of β, an important question is the relative efficiency of those estimators. It follows from (2) that, if m is considered fixed as $n \to \infty$, then $\hat{\beta}$, the maximizer of ℓ_c, is asymptotically efficient (Liang (1983)). However, this is not necessarily true if $m \to \infty$

as $n \to \infty$. It has been shown that, if the set of possible random effects distributions is sufficiently broad, then $\hat{\beta}$ is asymptotically efficient (Pfanzagl (1982, Chap. 14), Lindsay (1980)). However, when there is a parametric family of random effects distributions, $\hat{\beta}$ is not necessarily asymptotically efficient (Pfanzagl (1982), Chap. 14). Hence, if the dimension of γ is large relative to n, there may be some sacrifice of efficiency associated with the use of $\hat{\beta}$. However, the estimator of β is valid under any random effects distribution and any loss of efficiency must be viewed in that context.

Using a definition of finite-sample efficiency based on estimating functions, Godambe (1976) shows that the estimating function based on ℓ_c is optimal if either the conditioning statistic s is S-ancillary or the set of possible distributions of s is complete for fixed β.

4. Models with a Dispersion Parameter

Generalized linear models often have an unknown dispersion parameter as well, so that, conditional on γ, the loglikelihood function is of the form

$$\sum_{i,j} \frac{y_{ij}x_{ij}\beta + y_{ij}z_{ij}\gamma - k(x_{ij}\beta + z_{ij}\gamma)}{a(\sigma)} + \sum_{i,j} c(y_{ij}, \sigma),$$

where $\sigma > 0$ is an unknown parameter and c and a are known functions. The conditional likelihood given $\sum y_{ij}z_{ij}$ is still independent of γ, although it now depends on σ.

Inference about β may be based on the profile conditional loglikelihood, $\ell_c(\beta, \hat{\sigma}_\beta)$ where $\hat{\sigma}_\beta$ is the value of σ that maximizes $\ell_c(\beta, \sigma)$ for fixed β. Note that $\hat{\sigma}_\beta$ is valid estimator of σ for fixed β for any random effects distribution.

Now consider inference about σ. For fixed σ and γ, the statistics $t = \sum_{i,j} y_{ij}x_{ij}$ and $s = \sum_{i,j} y_{ij}z_{ij}$ are sufficient; hence, we may form a conditional likelihood for σ by conditioning on these statistics. The argument given in Section 2 showing that the conditional likelihood function given s is valid in the random effects model, for any random effects distribution, is valid for the conditional likelihood given s, t as well. Hence, the conditional likelihood estimator of σ is a valid estimator of σ in the random effects model for any random effects distribution.

Example 5. *Normal distribution*

Let y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$, denote independent normal random variables such that y_{ij} has mean $x_{ij}\beta + z_{ij}\gamma$ and variance σ^2. The conditional loglikelihood function given $\sum_{i,j} y_{ij}x_{ij}, \sum_{i,j} y_{ij}z_{ij}$ is given by $-\sum_{i,j}(y_{ij} - x_{ij}\hat{\beta} - z_{ij}\hat{\gamma})^2/(2\sigma^2) - (n-p-q)\log\sigma$, where $\hat{\beta}$ and $\hat{\gamma}$ are the least-squares estimators of β and γ, respectively. Hence, the conditional maximum likelihood estimator of σ^2 is the usual unbiased estimator: $s^2 = \sum_{i,j}(y_{ij} - x_{ij}\hat{\beta} - z_{ij}\hat{\gamma})^2/(n-p-q)$.

5. An Example

Consider the data in Table 1 of Booth and Hobert (1998, p.263). These data describe the effectiveness of two treatments administered at eight different clinics. For clinic i and treatment j, n_{ij} patients are treated and y_{ij} patients respond favorably. Following Beitler and Landis (1985), we model the clinic effects as random effects. Given the random effects, the y_{ij} are taken to be independent binomial random variables such that y_{i1} has index n_{i1} and mean $n_{i1}\exp(\gamma_i + \beta_0 + \beta_1)/[1 + \exp(\gamma_i + \beta_0 + \beta_1)]$ and y_{i2} has index n_{i2} and mean $n_{i2}\exp(\gamma_i + \beta_0)/[1 + \exp(\gamma_i + \beta_0)]$.

Let $y_i = y_{i1} + y_{i2}$. The conditional loglikelihood for β_1 is given by

$$\beta_1 \sum_i y_{i1} - \sum_i \log\{\sum_u \binom{n_{i1}}{u}\binom{n_{i2}}{y_i - u}\exp(\beta_1 u)\},$$

where the summation with respect to u is from $\max(0, y_i - n_{i2})$ to $\min(y_i, n_{i1})$.

The random effects $\gamma_1, \ldots, \gamma_8$ are taken to be independent and identically distributed, each with density $h(\cdot;\eta)$. Several choices were considered for the random effects distribution: a normal distribution, a logistic distribution, and an extreme value distribution for γ_i and a gamma distribution for $\exp(\gamma_i)$. In each case, γ_i has mean 0 and standard deviation η.

Table 1. Parameter estimates in the example.

Likelihood		Parameter		
		β_0	β_1	η
Conditional	Exact		0.756 (.303)	
	Saddlepoint		0.755 (.303)	
Integrated	Normal	-1.20 (0.549)	0.739 (0.300)	1.40 (0.430)
	Logistic	-1.22 (0.582)	0.738 (0.300)	1.52 (0.510)
	Extreme value	-1.15 (0.580)	0.743 (0.301)	1.49 (0.526)
	Gamma	-1.23 (0.643)	0.729 (0.299)	1.67 (0.653)

Table 1 contains parameter estimates based on the conditional likelihood as well as on the integrated likelihood for each of the four random effects distributions. In addition, estimates based on the saddlepoint approximation to the conditional likelihood function are given. The integrated likelihood functions were computed numerically using Hardy quadrature. Standard errors of the estimates are given in parentheses. Inferences for β_1 based on the conditional likelihood are essentially the same as those based on the integrated likelihood for each choice of the random effects distribution; note, however, that the conditional likelihood eliminates the need for numerical integration.

Acknowledgements

The work of N. Sartori was partially supported by a grant from Ministero dell'Istruzione, dell'Universitá e della Ricera, Italy; the work of T. A. Severini was partially supported by the National Science Foundation. N. Sartori gratefully acknowledges the hospitality of the Department of Statistics, Northwestern University.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimates. *J. Roy. Statist. Soc. Ser. B* **32**, 283-301.

Andersen, E. B. (1971). The asymptotic distribution of conditional likelihood ratio tests. *J. Amer. Statist. Assoc.* **66**, 630-633.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika* **67**, 293-310.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343-365.

Basawa, I. V. (1981). Efficiency of conditional maximum likelihood estimators and confidence limits for mixtures of exponential families. *Biometrika* **68**, 515-523.

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

Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93**, 262-272.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research,* 1*: The Analysis of Case-Control Studies*. International Agency for Cancer Research, Lyon.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81-91.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631-650.

Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 445-461.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. 2nd edition. Oxford University Press, Oxford.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences (with discussion). *J. Educ. Behav. Statist.* **20**, 115-147, 228-233.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statist. Neerland.* **48**, 1-22.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277-284.

Jensen, J. L. (1995). *Saddlepoint Approximations.* Oxford University Press, Oxford.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. Ser. B* **58**, 619-678.

Liang, K.-Y. (1983). The asymptotic efficiency of conditional likelihood methods. *Biometrika* **71**, 305-313.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lindsay, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc.* **296**, 639-665.

Lindsay, B. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11**, 486-497.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*. IMS, Hayward.

Lindsay, B., Clogg, C. C. and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* **86**, 96-107.

McGilchrist, C. A. (1994). Estimation in generalized linear mixed models. *J. Roy. Statist. Soc. Ser. B* **56**, 61-69.

Neuhaus, J. M., Hauck, W. W. and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755-762.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Theory.* Springer-Verlag, Heidelberg.

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

Schall, R. (1991). Estimation in generalized linear model with random effects. *Biometrika* **78**, 719-727.

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

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *J. Roy. Statist. Soc. Ser. B* **57**, 749-760.

van der Vaart, A. W. (1988). Estimating a real parameter in a class of semiparametric models. *Ann. Statist.* **16**, 1450-1474.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**, 79-86.

Department of Statistics, University of Padova, via C. Battisti 241-243, 35121 Padova, Italy.

E-mail: sartori@stat.unipd.it

Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, Illinois 60208-4070. U.S.A.

E-mail: severini@northwestern.edu

(Received October 2002; accepted June 2003)