UW Biostatistics Working Paper Series

1-24-2003

# Whither PQL?

Norm Breslow
*University of Washington*, norm@u.washington.edu

# Whither PQL?

Norman Breslow

January 24, 2003

ABSTRACT Generalized linear mixed models (GLMM) are generalized linear models with normally distributed random effects in the linear predictor. Penalized quasi-likelihood (PQL), an approximate method of inference in GLMMs, involves repeated fitting of linear mixed models with "working" dependent variables and iterative weights that depend on parameter estimates from the previous cycle of iteration. The generality of PQL, and its implementation in commercially available software, has encouraged the application of GLMMs in many scientific fields. Caution is needed, however, since PQL may sometimes yield badly biased estimates of variance components, especially with binary outcomes.

Recent developments in numerical integration, including adaptive Gaussian quadrature, higher order Laplace expansions, stochastic integration and Markov chain Monte Carlo (MCMC) algorithms, provide attractive alternatives to PQL for approximate likelihood inference in GLMMs. Analyses of some well known datasets, and simulations based on these analyses, suggest that PQL still performs remarkably well in comparison with more elaborate procedures in many practical situations. Adaptive Gaussian quadrature is a viable alternative for nested designs where the numerical integration is limited to a small number of dimensions. Higher order Laplace approximations hold the promise of accurate inference more generally. MCMC is likely the method of choice for the most complex problems that involve high dimensional integrals

# 1 Introduction

> **P**enalized **Q**uasi-**L**ikelihood is a technique for approximate inference in GLMMs and is not a rigorous statistical method in its own right.[33, p. 390, emphasis added]

The generalized linear model or GLM [35] is a prime tool of the applied statistician. It brings the power and flexibility of linear regression modeling to the analysis of data with outcomes, particularly discrete outcomes, that do not satisfy the conventional assumptions of least squares. The linear mixed model or LMM, with its multiple levels of random variation and best linear unbiased prediction of random effects [19], dominates statistical theory and applications in diverse fields including animal breeding and education. During the past decade these two models have been fused into a hybrid body of statistical theory and methodology known as the generalized linear mixed model or GLMM.[46, 40, 5, 49, 12, 31, 16, 29, 30] An even more general formulation, known as the hierarchical generalized linear model or HGLM, encompasses both normal and non-normal probability distributions for the random effects.[22, 23]

GLM and LMM parameter estimates are obtained from estimating equations that are unbiased under simple moment conditions and that may be solved by iterative solution of systems of linear equations. For the GLMM, by contrast, the specification of normally distributed random effects intrinsically defines the marginal likelihood and its logarithmic derivatives. The fact that the integrals in the GLMM estimating equations cannot be evaluated in closed form has seriously limited GLMM applications. Until recently the only available commercial software was the EGRET program [9] that implemented the logistic-normal model for clustered binary outcomes, unit level covariates and a cluster level random intercept. Thus substantial interest was generated by the work of Schall [40], Breslow and Clayton [5], Wolfinger [48] and others who developed a general approach to approximate inference. Their "penalized quasi-likelihood" or PQL procedure involved repeated fitting of the LMM using a working outcome variable and iterative weights that mimicked the standard iterative least squares algorithm used to fit the GLM.[28, §2.5] It was disseminated in macros written for several commercially distributed LMM programs: the GLIMMIX macro for PROC MIXED in SAS [26]; the PQL option for MLwiN [37]; and the HLM series distributed by SSI [38]. The IR-REML macro in GENSTAT [32] facilitated fitting of both GLMMs and HGLMs. This stimulated increasing use of these procedures in old disciplines such as sociology, where hierarchical models were already familiar, and in new ones like epidemiology [17], where they were just being discovered.

As usual when software for complicated statistical inference procedures is broadly disseminated, there is potential for abuse and misinterpretation. In spite of the fact that PQL was initially advertised as a procedure for *approximate* inference in GLMMs, and its tendency to give seriously biased estimates of variance components and *a fortiori* regression parameters with binary outcome data was emphasized in multiple publications [5, 6, 24], some statisticians seemed to ignore these warnings and to think of PQL as synonymous with GLMM.[7] In an apparent reaction to these developments, and to the algorithm's acknowledged shortcomings for binary outcome data, the authors of one recent textbook have recommended that PQL "not be used in practice".[30, p.234]

The purpose of this review is to take stock of PQL as a tool of the applied statistician now that some years have passed since it was first implemented in commercial software. In the interim, substantial advances have taken place in statistical computing. "True" maximum likelihood (ML) estimation is now available for a much wider range of problems by using numerical integration to calculate marginal likelihoods and solve score equations. In particular, the adaptive Gaussian quadrature methods [27, 36] implemented in SAS PROC NLMIXED [45] apply to clustered data problems where the dimensionality of the required integrations is in the low single digits. Higher order Laplace approximations [39], implemented for the logistic-normal model in the latest HLM program [38], may prove to be just as accurate as quadrature and more widely applicable.

Recent Monte Carlo approaches to numerical integration include Monte Carlo relative likelihood [13], Monte Carlo EM [29, 3] and Monte Carlo

Newton-Raphson [20]. Kuk and Cheng [21] provide an excellent, comprehensive review of these stochastic procedures. Their use in practice to date has been limited by their longer computing times and the fact that none have yet been implemented in standard software packages. Booth and Hobert [3] argue that their "automated" Monte Carlo EM algorithm is an improvement on the Markov chain Monte Carlo (MCMC) version. It facilitates assessment of convergence and thus removes one of the main impediments to commercial implementation. Hierarchical Bayes procedures, which also depend on MCMC to evaluate posterior distributions, have been implemented in available, supported software and are increasingly used in applications. [8, 44, 34] These Monte Carlo methods will undoubtedly see much greater use with continuing improvements in computing technology. In view of their greater complexity, however, and the desire to keep this review focussed on the most immediate competitors to PQL, further discussion of Monte Carlo methods is left to investigators who are more familiar with their properties. Comparisons with the "h-likelihood" methodology of Lee and Nelder [22, 23] for inference in HGLMs also have been left for others.

## 2 GLMMs and PQL

The GLMM is a model for the hierarchical regression analysis of a series of $n$ univariate response measurements $y_i$ on $p$-dimensional covariates $x_i$ associated with fixed effects and $q$-dimensional covariates $z_i$ associated with random effects of interest ($i = 1, \dots, n$). Conditional on the unobserved values of a $q$-vector $b$ of random effects, and on all the covariates, the $y_i$ are assumed to be independent observations with means and variances specified by a GLM.[28] Specifically we suppose

$$\mathrm{E}(y_i|b) = \mu_i^b = h(\eta_i^b) = h(x_i^T\alpha + z_i^T b)$$

$$\mathrm{Var}(y_i|b) = \frac{\phi}{a_i} v(\mu_i^b)$$

where $g = h^{-1}$ is the link function that relates the conditional means $\mu_i^b$ to the linear predictors $\eta_i^b$; $v(\cdot)$ is the variance function that relates the conditional means and variances to each another; $\phi$ is a scale factor assumed equal to one for the standard binomial and Poisson models; and $a_i$ is a prior weight such as a binomial denominator. Specification of the model is completed by the assumption that $b$ follows a $q$-dimensional normal distribution with mean 0 and variance matrix $D(\theta)$ depending on a vector of dispersion parameters $\theta$. Examples of typical GLMM applications are considered in Sections 4 and 5.

The objective function for estimation of the GLMM parameters is the integrated quasi-likelihood $\mathrm{L}(\alpha, \theta)$ given by

$$L = \frac{1}{\sqrt{(2\pi)^q |D(\theta)|}} \int_{R^q} \exp\left[-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i, \mu_i^b) - \frac{1}{2} b^T D^{-1}(\theta) b\right] db \quad (1)$$

where

$$d_i(y, \mu) = -2a_i \int_y^\mu \frac{y-u}{v(u)} du$$

denotes the weighted deviance.[28] If $Y$ is Gaussian and $g(\cdot)$ the identity, the integral in (1) is normal and may be evaluated in closed form. Otherwise, maximization of this expression is intrinsically complicated by the integrations that must be performed numerically at each cycle of iteration. One approach to the integration, which eventually leads to the PQL algorithm, is to make a Laplace approximation. The term in square brackets in (1), the logarithm of the "penalized quasi-likelihood", is replaced by its quadratic expansion in $b$ about the value $\tilde{b}$ at which it is maximized. Components of $\tilde{b}$ serve as predictors of the random effects. After some adjustments to the resulting normal integral, application of Fisher scoring to determine $(\hat{\alpha}, \tilde{b})$ as a function of $\theta$ leads to the familiar mixed model equations for joint estimation of fixed and random effects, as originally derived by Henderson [19], but now involving a working vector $Y^*$ and iterative weights $w_i$. Further approximations lead to the standard REML equations for $\theta$. Specifically, with $\hat{\mu}_i^b = h(x_i^T \hat{\alpha} + z_i^T \tilde{b})$,

$$Y_i^* = x_i^T \hat{\alpha} + z_i^T \tilde{b} + (y_i - \hat{\mu}_i^b) g'(\hat{\mu}_i^b)$$

and

$$w_i = \phi a_i [g'(\hat{\mu}_i^b)]^2 v(\hat{\mu}_i^b)^{-1},$$

the algorithm repeatedly applies mixed model REML estimation to the normal theory problem

$$Y^* = X\alpha + Zb + \varepsilon,\ b \sim \mathcal{N}(0, D(\theta)),\ \varepsilon \sim \mathcal{N}(0, W^{-1})$$

where $W$=diag$(w_i)$. See Breslow and Clayton [5] for details.

Although PQL yields REML estimates of variance components and regression coefficients in the Gaussian linear case, in general it only provides an approximation to these quantities. For the simplest GLMM involving clustered data with a single dispersion component $\theta$, Breslow and Lin [6] expanded both the efficient score based on the true profile log-likelihood function, and the PQL variance estimating equation, in Taylor series about $\theta = 0$. They thereby showed that the asymptotic bias in the PQL estimator $\hat{\theta}_{\mathrm{P}}$ was a nearly linear function of $\theta$ in a neighborhood of the origin. By determining the slope of this linear relationship, which is estimable from the standard GLM fit assuming $\theta = 0$, they derived a correction factor for $\hat{\theta}_{\rm p}$ that removed the asymptotic bias for small $\theta$ at the cost of some increase in variability. Lin and Breslow [24] extended this work for models with multiple variance components, deriving a matrix correction factor, and termed the resulting procedure corrected PQL or CPQL.

An alternative derivation of the PQL algorithm developed by Schall [40] and others uses a linearization of the conditional mean as a function of fixed and random effects. Consider, for example, the two-level model with $I$ clusters having $n_i$ observations per cluster, $i = 1, \ldots, I$, and random effects $b_i$ assumed independent between clusters. The $j^{th}$ observation in cluster $i$ may be written

$$y_{ij} = \mu^b_{ij} + \varepsilon_{ij} = h(x^T_{ij}\alpha + z^T_{ij}b_i) + \varepsilon_{ij}$$

with $\mathrm{var}(\varepsilon_{ij}) = \phi v(\mu^b_{ij})/a_i$, $j = 1, \ldots, n_i$. Expanding $h$ about the current estimates $(\hat{\alpha}, \tilde{b})$ based on the current $\hat{\theta}$ gives

$$y_{ij} \approx \hat{\mu}^b_{ij} + h'(\hat{\eta}^b_{ij})[x^T_{ij}(\alpha - \hat{\alpha}) + z^T_{ij}(b_i - \tilde{b_i})] + \varepsilon_{ij} \tag{2}$$

which implies that the “working” observation $Y^*_{ij} = \hat{\eta}^b_{ij} + g'(\hat{\mu}^b_{ij})(y_{ij} - \hat{\mu}^b_{ij})$ satisfies

$$Y^*_{ij} = x^T_{ij}\alpha + z^T_{ij}b_i + \varepsilon^*_{ij} \tag{3}$$

where, at least to an approximation for the $\varepsilon^*_{ij}$,

$$b_i \sim \mathcal{N}(0, D(\theta)) \quad \text{and} \quad \varepsilon^*_{ij} \sim \mathcal{N}\left(0, \phi[g'(\hat{\mu}^b_{ij})]^2 v(\hat{\mu}^b_{ij})/a_i\right). \tag{4}$$

Updated estimates of $(\alpha, b, \theta)$ are obtained by solving for them in the LMM defined by (3) and (4), *i.e.*, by using the PQL algorithm.

A further expansion of the conditional mean in terms involving $b_i$ alone adds $\frac{1}{2}h''(\hat{\eta}^b_{ij})z^T_{ij}(b_i - \hat{b_i})(b_i - \hat{b_i})^T z_{ij}$ to the right hand side of (2). Goldstein and Rasbash [14, 16, 15] suggested that one ignore the cross-products involving different components of $b_i$, add the mean values of the resulting quadratic terms as offsets to the regression model and treat their residuals as additional random error terms with known variance. This modified procedure, implemented as PQL2 in MLwiN [37], is also intended to improve the estimates of variance components.

## 3 Adaptive Gauss-Hermite Quadrature

Consider the two-level GLMM with $I$ independent clusters of observations $\{y_{ij},\ j = 1, \ldots, n_i\}$, $i = 1, \ldots, I$ and a random intercept so that

$$\mu^b_{ij} = \mathrm{E}(y_{ij}|b_i) = h(x^T_{ij}\alpha + b_i), \quad b_i \overset{i.i.d}{\sim} \mathcal{N}(0, \theta).$$

To simplify matters , suppose $g = h^{-1}$ is the canonical link function so that $v(\mu) = [g'(\mu)]^{-1}$ and furthermore that the scale factor and prior weights are all unity. This setup applies, for example, to two-level log-linear modeling of Poisson data and to logistic regression for clustered binary outcome data. The contribution to the marginal likelihood (integrated quasi-likelihood) for the $i^{\text{th}}$ cluster is

$$\begin{aligned} L_i &= \frac{1}{\sqrt{2\pi\theta}} \int L_i^{\text{c}}(b) e^{-\frac{b^2}{2\theta}} db \\ &= \mathrm{E}_{\mathcal{N}(0,\theta)} L_i^{\text{c}}(b) \end{aligned} \tag{5}$$

where $\mathrm{E}_{\mathcal{N}(\mu,\theta)}$ denotes expectation with respect to the $\mathcal{N}(\mu,\theta)$ distribution and $L_i^{\text{c}}$ is the conditional quasi-likelihood contribution

$$L_i^{\text{c}}(b) = \exp\{-\frac{1}{2}\sum_{j=1}^{n_i} d_{ij}(y_{ij}, \mu_{ij}^b)\}.$$

Ordinary Gauss-Hermite quadrature approximates the integral in (5) with the sum

$$L_i \simeq \frac{1}{\sqrt{\pi}} \sum_{r=1}^{R} \omega_r L_i^{\text{c}}(\sqrt{2\theta} t_r)$$

where the $t_r$ are the $R$ quadrature points, roots of the $R$-degree Hermite polynomial, and the $\omega_r$ denote the associated weights.[11, p. 924] The problem with this approach is that the same quadrature points are used for each cluster, irrespective of the cluster outcomes. Thus, for some $i$, the conditional quasi-likelihoods $L_i^{\text{c}}(b)$ may take large values for $b$ well outside the range covered by the points $\{\sqrt{2\theta} t_r,\ r = 1, \ldots, R\}$.

Let $\phi(b; \mu, \sigma^2)$ denote the density of the normal distribution with mean $\mu$ and variance $\sigma^2$. The basic idea behind adaptive quadrature as introduced by Liu and Pierce [27] is the same one that underlies the Laplace integral approximation, namely, to determine the normal density $\phi(b; \tilde{b}_i, \tilde{\sigma}_i^2)$ that best approximates the entire integrand $L_i^{\text{c}}(b)\phi(b; 0, \theta)$ in (5). The value that maximizes the integrand, $\tilde{b}_i$, is obtained as the solution (in $b$) to $\sum_j (y_{ij} - \mu_{ij}^b) + b/\theta = 0$. The curvature in the log integrand at its maximum is the inverse of $\tilde{\sigma}_i^2 = [\sum_j v(\mu_{ij}^{\tilde{b}_i}) + \theta^{-1}]^{-1}$ [5, §2.1]. Once these are computed, the marginal likelihood contribution is approximated via

$$\begin{aligned} L_i &= \mathrm{E}_{\mathcal{N}(\tilde{b}_i, \tilde{\sigma}_i^2)} \left[ \frac{L_i^{\text{c}}(b)\phi(b; 0, \theta)}{\phi(b; \tilde{b}_i, \tilde{\sigma}_i^2)} \right] \\ &\simeq \frac{1}{\sqrt{\pi}} \sum_{r=1}^{R} \omega_r \frac{L_i^{\text{c}}(\tilde{b}_i + \sqrt{2}\tilde{\sigma}_i t_r)\phi(\tilde{b}_i + \sqrt{2}\tilde{\sigma}_i t_r; 0, \theta)}{\phi(\tilde{b}_i + \sqrt{2}\tilde{\sigma}_i t_r; \tilde{b}_i, \tilde{\sigma}_i^2)} \\ &= \sqrt{2}\tilde{\sigma}_i \sum_{r=1}^{R} \omega_r e^{t_r^2} L_i^{\text{c}}(\tilde{b}_i + \sqrt{2}\tilde{\sigma}_i t_r)\phi(\tilde{b}_i + \sqrt{2}\tilde{\sigma}_i t_r; 0, \theta). \end{aligned} \tag{6}$$

The Laplace approximation is given by (6) for $R = 1$, $\omega_1 = 1$ and $t_1 = 0$.

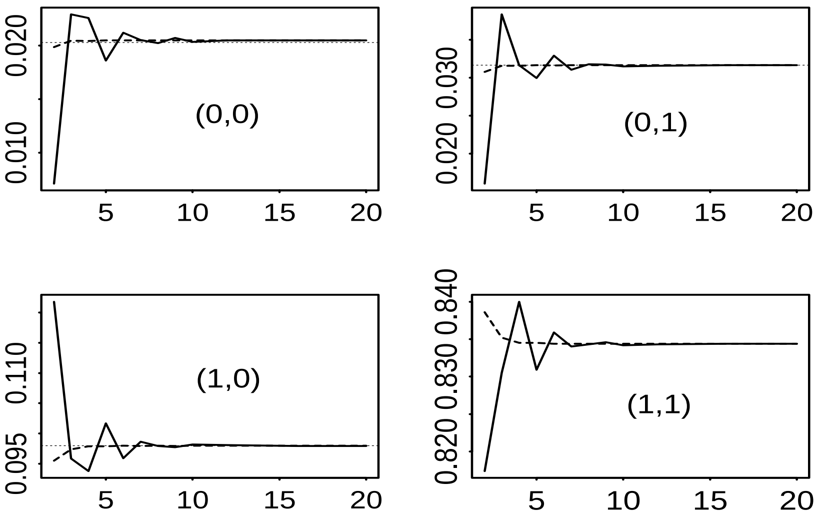


FIGURE 1. Marginal probabilities (ordinates) of each of four possible outcomes with matched pairs of binary outcome data estimated using standard (solid line) and adaptive (dashed line) Gauss-Hermite quadrature. The outcome vector is shown in the center of each panel. The abscissae show the number of quadrature points.

Use of standard and adaptive Gauss-Hermite quadrature to approximate the integrated likelihood is illustrated in Figure 1 for the special case of matched pairs of binary outcome data with a single binary covariate that varies within clusters. Defining expit$(x) = [1 + e^{-x}]^{-1}$, this model has $\mu_{i1}^b$ =expit$(\alpha_0 + b_i)$, $\mu_{i2}^b$ =expit$(\alpha_0 + \alpha_1 + b_i)$ and $v(\mu) = \mu(1-\mu)$. The parameter values used were $\alpha_0 = 4$, $\alpha_1 = -2$ and $\theta = 4$. The four panels of the figure show plots of the approximated marginal probabilities of the four possible outcomes for $(y_{i1}, y_{i2})$, namely (0,0), (0,1), (1,0), and (1,1), as functions of the number $R$ of quadrature points. Using the adaptive procedure, the integral approximations converged to the fourth decimal place with $R = 5$. By contrast, with the standard procedure, they continued to oscillate for $R$ well beyond 5 .

Pinheiro and Bates [36] adapted this methodology for multidimensional integrals and their methods have been incorporated into PROC NLMIXED in SAS.[45] In the sequel we compare results using NLMIXED and GLIMMIX, *i.e.*, using ML and PQL, with two well studied sets of data.

## 4 Meta Analysis of Clinical Trials Data

Our first example involves a series of 2×2 tables of counts of "successes" and "failures" among 293 patients distributed in treatment and control groups in eight clinical centers (Table 1). Introduced to statisticians by Beitler and Landis [2], these data have been widely used to illustrate different methods for mixed effects modeling of categorical data. The developers of GLIMMIX, for example, noted that their macro converged more consistently if one first converted the table of counts to a series of binary outcome variables and covariates.[26, p. 440] The data also featured prominently in a recent review by Agresti and Hartzel [1] of methods for meta analysis of binary outcome data. The object of many of these analyses has been to estimate the clinic specific treatment effect, expressed as an odds ratio and assumed constant over clinics, while adjusting for clinic to clinic variation in baseline success rates via random effects modeling. There has also been interest in deciding whether there is evidence for treatment by center interaction.

TABLE 1. Clinical trial of topical cream for infection

| Center | Treatment | Response | | Total | Success |
|---|---|---|---|---|---|
| | | Success | Failure | patients | Rate (%) |
| 1 | Drug | 11 | 25 | 36 | 30.6 |
| | Control | 10 | 27 | 37 | 27.0 |
| 2 | Drug | 16 | 4 | 20 | 80.0 |
| | Control | 22 | 10 | 32 | 68.8 |
| 3 | Drug | 14 | 5 | 19 | 73.7 |
| | Control | 7 | 12 | 19 | 36.8 |
| 4 | Drug | 2 | 14 | 16 | 12.5 |
| | Control | 1 | 16 | 17 | 5.9 |
| 5 | Drug | 6 | 11 | 17 | 35.3 |
| | Control | 0 | 12 | 12 | 0.0 |
| 6 | Drug | 1 | 10 | 11 | 9.1 |
| | Control | 0 | 10 | 10 | 0.0 |
| 7 | Drug | 1 | 4 | 5 | 20.0 |
| | Control | 1 | 8 | 9 | 11.1 |
| 8 | Drug | 4 | 2 | 6 | 66.7 |
| | Control | 6 | 1 | 7 | 85.7 |
| Total | Drug | 55 | 75 | 130 | 42.3 |
| | Control | 47 | 96 | 143 | 32.9 |

Source: Beitler and Landis [26]

Let $y_{ij}$ denote the binary outcome (1 for success, 0 for failure) for the $j^{\text{th}}$ subject in the $i^{\text{th}}$ clinic. Suppose the covariate $x_{ij}$ takes values $-\frac{1}{2}$ for control and $+\frac{1}{2}$ for treatment. This coding helps to orthogonalize the design matrix and render more plausible the implicit assumption of independence between random intercept and random slope (interaction) terms in what follows. Two models of interest are I: logit $\mathrm{E}(y_{ij}|b_i) = \alpha_0 + \alpha_1 x_{ij} + b_i^0$; and II: logit $\mathrm{E}(y_{ij}|b_i) = \alpha_0 + \alpha_1 x_{ij} + b_i^0 + b_i^1 x_{ij}$, the first corresponding to the hypothesis of constant odds ratio. The parameter of interest $\alpha_1$ represents the *within clinic* log odds ratio comparing treatment and control groups. This is assumed constant across clinics in Model I but may vary by clinic in Model II. Tables 2 and 3 compare results obtained using four procedures for fitting GLMMs, including the PQL2 procedure mentioned at the end of §2. Also shown for Model I are results for the "exact" conditional maximum likelihood (CML) analysis, based on convolutions of the non-central hypergeometric distributions that arise when one conditions on all four marginal totals in each table. [10, §2.5] The analog for Model II is the GLMM that adds a random effect to the log odds ratio parameter in each non-central hypergeometric distribution. This may be fitted by PQL using methods previously described.[5, §6.4] Some notable features of this comparison include: (*i*) the lack of any suggestion for a treatment by clinic interaction; (*ii*) the excellent agreement between the estimates and standard errors obtained by ML (adaptive quadrature) and PQL, especially for the variance component of the random intercept; and (*iii*) the fact that the PQL2 results are substantially different from the others. Note that the standard errors of the variance components estimated by the GLIMMIX and MLwiN implementations of PQL differ slightly. Otherwise the results were identical.

TABLE 2. Estimates ± standard errors for Model I

| Method | $\alpha_0$ | $\alpha_1$ | $\theta_0$ |
|---|---|---|---|
| NLMIXED (ML) | -0.828±0.533 | 0.739±0.300 | 1.96±1.19 |
| GLIMMIX (PQL) | -0.784±0.537 | 0.724±0.296 | 2.03±1.26 |
| MLwiN (PQL) | -0.784±0.537 | 0.724±0.296 | 2.03±1.19 |
| MLwiN (PQL2) | -0.789±0.606 | 0.859±0.310 | 2.56±1.46 |
| Hypergeometric | (CML) | 0.756±0.303 | |

TABLE 3. Estimates ± standard errors for Model II

| Method | $\alpha_0$ | $\alpha_1$ | $\theta_0$ | $\theta_1$ |
|---|---|---|---|---|
| NLMIXED (ML) | -0.830±0.535 | 0.746±0.323 | 1.97±1.20 | 0.02±0.32 |
| GLIMMIX (PQL) | -0.791±0.538 | 0.749±0.333 | 2.04±1.27 | 0.12±0.41 |
| MLwiN (PQL) | -0.791±0.538 | 0.749±0.333 | 2.04±1.15 | 0.12±0.37 |
| MLwiN (PQL2) | -0.870±0.614 | 0.830±0.367 | 2.61±1.46 | 0.20±0.45 |
| Hypergeometric | (PQL) | 0.793±0.352 | | 0.16±0.48 |

Table 4 reports results of a small simulation study designed to evaluate more systematically the performance of PQL in this setting.[4] For each of 10,000 simulations, 8 pairs of independent binomial observations $r_{ij} \sim \text{binom}(p_{ij}, n_{ij})$, $i = 1, \ldots, 8$, $j = 1, 2$ were drawn with denominators $n_{ij}$ chosen equal to those in the penultimate column of Table 1. The GLMM was specified by logit $p_{ij} = \alpha_0 + \alpha_1(2x_{ij} - 1) + b_i^0 + b_{ij}^1$ where $b_i^0 \sim \mathcal{N}(0, \theta_0)$ and $b_{ij}^1 \sim \mathcal{N}(0, \theta_1/2)$ were mutually independent sets of random effects. Thus the $b_i^0$ were random clinic effects, with roughly the same amount of clinic-to-clinic variation as for the data in Table 1, while the differences between $b_{ij}^1$ for $j = 1$ and $j = 2$ represented the variation in treatment effects (log odds ratios). Parameter settings were $\alpha_0 = 0$, $\theta_0 = 2$, $\alpha_1 = 0, 1, 2$ and $\theta_1 = \text{Var}(b_{i1}^1 - b_{i2}^1) = 0, 0.5, 1, 2$. $\bar{\alpha}_1$ and $\bar{\theta}_1$ refer to the averages of the estimates of these two parameters over the 10,000 replications. The error rates refer to the proportion of replicates for which the 95% confidence interval for $\alpha_1$ excluded the true value on the left or the right side.

TABLE 4. Results of the simulation study of PQL

| True values | | Estimates | | Error rates | |
|---|---|---|---|---|---|
| $\theta_1$ | $\alpha_1$ | $\bar{\theta}_1$ | $\bar{\alpha}_1 - \alpha_1$ | Left | Right |
| 0.0 | 0 | 0.15 | 0.000 | 0.015 | 0.016 |
| | 1 | 0.16 | 0.015 | 0.012 | 0.017 |
| | 2 | 0.18 | 0.030 | 0.013 | 0.018 |
| 0.5 | 0 | 0.58 | 0.002 | 0.032 | 0.027 |
| | 1 | 0.58 | 0.013 | 0.029 | 0.033 |
| | 2 | 0.60 | 0.023 | 0.018 | 0.034 |
| 1.0 | 0 | 1.05 | -0.003 | 0.030 | 0.032 |
| | 1 | 0.96 | -0.012 | 0.027 | 0.035 |
| | 2 | 1.04 | 0.002 | 0.024 | 0.038 |
| 2.0 | 0 | 2.00 | -0.016 | 0.026 | 0.031 |
| | 1 | 1.98 | 0.000 | 0.030 | 0.032 |
| | 2 | 1.99 | -0.000 | 0.025 | 0.029 |

Source: Breslow, Leroux and Platt [4]

The simulated data were analyzed using PQL as described above for the log odds ratio GLMM based on the non-central hypergeometric distribution. As with any mixed model, there was a tendency to over-estimate slightly the small (or null) values of the variance component since negative estimates were not allowed. The systematic underestimation of variance components often observed with clustered binary data (see §6 below) was not a problem here, probably because of the relatively large denominators and mid-range values for many of the binomial observations. PQL estimates of the regression coefficient $\alpha_1$ and of the larger values of the variance component were remarkably unbiased. Error rates for interval estimation were quite satisfactory. Not shown here are corresponding results for the empirical transform (ET) method, which consisted of applying ordinary LMM methods to derived outcome variables. The derived variable was the logarithm of the observed odds ratio in each table, with 0.5 added to both cells whenever any marginal total of success or failure was zero, so as to avoid infinities. Conditional on the random effects, this outcome variable was treated as normally distributed with variance equal to the inverse of the sum of reciprocals of the cell frequencies. The ET estimates of both the variance component and the regression coefficient were seriously biased towards zero, so that the random effect predictors were similarly misbehaved.[4, pp. 57-58] A similar tendency of ET to underestimate the variance component was observed for simulated Poisson observations representing spatially correlated rates when the mean rates were very small.[4, pp. 58-59] Thus the recent recommendation that ET methods be used in preference to PQL in such situations appears to be unfounded.[30, p. 283]

# 5 Longitudinal Series of Counts

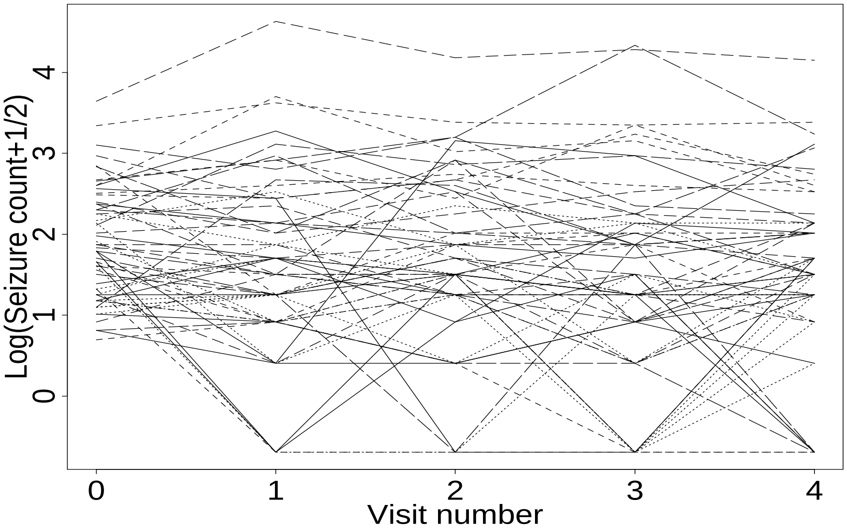


FIGURE 2. Epilepsy seizure counts at baseline and four follow-up periods.

Our second example involves a series of counts of seizures recorded by 59 patients with epilepsy for each of four two-week periods that preceded clinic visits. Introduced by Thall and Vail [47], these data also have been used by numerous statisticians to illustrate methods for analysis of longitudinal data with discrete outcomes. Figure 2 plots the patient trajectories of the log counts, augmented by 0.5 to avoid infinities. Each trajectory starts with the log of the baseline count over the eight-week period before the study, which was divided by four for comparability. Other fixed covariates of interest included a binary treatment indicator, the logarithm of age in years, and either a binary indicator for the fourth visit or the visit number $j$ after division by ten.

TABLE 5. Estimates ± standard errors for Model III

| Parameter | NLMIXED<br>(ML) | GLIMMIX<br>(PQL) | MLwiN<br>(PQL) | MLwiN<br>(PQL2) |
|---|---|---|---|---|
| | | Regression coefficients | | |
| Constant | -1.117±1.182 | -1.256±1.220 | -1.256±1.220 | -1.335±1.239 |
| Baseline* | 0.884±0.131 | 0.872±0.136 | 0.872±0.136 | 0.881±0.138 |
| Treatment | -0.933±0.401 | -0.917±0.413 | -0.917±0.413 | -0.929±0.420 |
| Bas*×Trt | 0.338±0.203 | 0.331±0.210 | 0.331±0.210 | 0.336±0.213 |
| Age* | 0.484±0.347 | 0.472±0.358 | 0.472±0.359 | 0.481±0.364 |
| Visit 4 | -0.161±0.055 | -0.161±0.055 | -0.161±0.055 | -0.161±0.055 |
| | | Variance component | | |
| $\sqrt{\theta_0}$ | 0.503±0.059 | 0.524±0.062 | 0.524±0.059 | 0.529±0.060 |

* log transform

With $y_{ij}$ now denoting the seizure count reported at the $j^{\text{th}}$ visit by the $i^{\text{th}}$ patient, assumed to have a Poisson distribution after conditioning on the random effects, two models of interest were Model III: log $\text{E}(y_{ij}|b_i) = x_i^T\alpha + b_i^0$ and Model IV: log $\text{E}(y_{ij}|b_i) = x_i^T\alpha + b_i^0 + b_i^1 j/10$. Model IV was the more interesting in that it provided for a patient specific random slope and intercept, assumed to follow a bivariate normal distribution, to model the trends in the trajectories. Results of fitting these models using the NLMIXED and GLIMMIX procedures in SAS, and the PQL and PQL2 methods in MLwiN, are shown in Tables 5 and 6. There was remarkably good agreement in estimation of the regression coefficients and their standard errors. By contrast to the previous example, PQL2 produced regression coefficients slightly closer to those of ML than did PQL. The PQL2 estimates of the variance components, however, were slightly further from the ML estimates. The high (0.4 or so) within cluster (patient) correlation in the log epilepsy counts is reflected in the large, and highly statistically significant, estimates of variance components.

TABLE 6. Estimates ± standard errors for Model IV

| Parameter | NLMIXED<br>(ML) | GLIMMIX<br>(PQL) | MLwiN<br>(PQL) | MLwiN<br>(PQL2) |
|---|---|---|---|---|
| | | Regression coefficients | | |
| Constant | -1.368±1.201 | -1.267±1.215 | -1.268±1.215 | -1.361±1.241 |
| Baseline* | 0.885±0.131 | 0.870±0.135 | 0.870±0.135 | 0.882±0.138 |
| Treatment | -0.929±0.402 | -0.910±0.411 | -0.910±0.411 | -0.922±0.421 |
| Bas*×Trt | 0.338±0.204 | 0.330±0.209 | 0.330±0.209 | 0.335±0.214 |
| Age* | 0.477±0.354 | 0.463±0.357 | 0.463±0.357 | 0.472±0.364 |
| Visit/10 | -0.266±0.165 | -0.264±0.157 | -0.264±0.157 | - 0.267±0.160 |
| | | Variance components | | |
| $\sqrt{\theta_{00}}$ | 0.502±0.059 | 0.521±0.062 | 0.521±0.061 | 0.527±0.063 |
| $\theta_{01}$ | 0.003±0.089 | 0.002±0.090 | 0.002±0.088 | 0.005±0.091 |
| $\sqrt{\theta_{11}}$ | 0.729±0.157 | 0.737±0.157 | 0.737±0.162 | 0.756±0.165 |

* log transform

## 6 Further Simulations with Binary Outcome Data

To further evaluate the bias of PQL estimates of variance components with binary outcome data, and assess the degree of correction afforded by CPQL, a new series of simulation experiments was run using a variant of a model originally proposed by Zeger and Karim [50] for clustered data. Each experiment involved $K$ clusters of constant size $n$. Binary outcome variables $y_{ij}$ for $i = 1, \ldots, K$ and $j = 1, \ldots, n$ were generated according to the hierarchical model

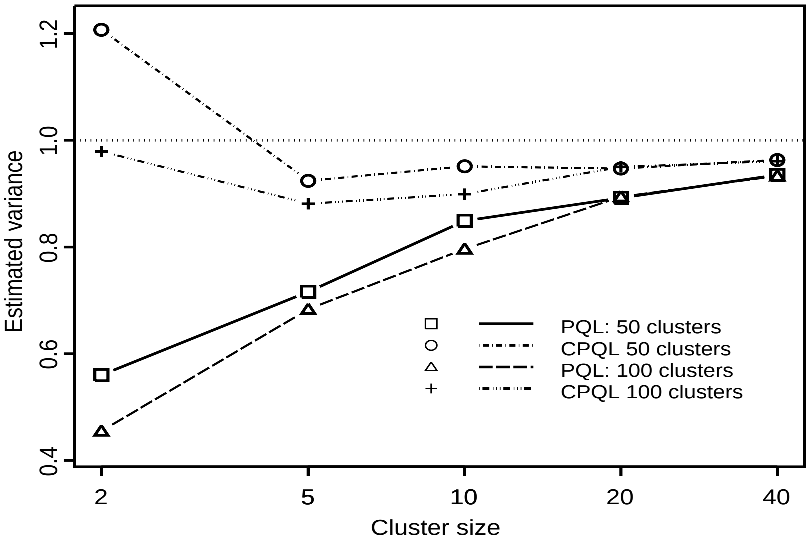
$$\text{logitE}(y_{ij}|b_i) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 x_i + \alpha_3 t_{ij} x_i + b_i,$$

where the $t_{ij}$ were unit level covariates that were randomly generated from the uniform distribution on the interval [-$\frac{1}{2}$, $\frac{1}{2}$], the $x_i$ were subject level covariates of which the first half took the value 0 and the remainder the value 1, and the $b_i$ were independent, normally distributed random effects with mean 0 and variance $\theta$. The parameter values were $\alpha_0$ = -0.5, $\alpha_1 = 1$, $\alpha_2$ = -1, $\alpha_3 = 0.5$ and $\theta = 1$. The number $K$ of clusters was 50 or 100 and the sample size $n$ per cluster ranged between 2 and 40. Each experiment was replicated 200 times at each parameter setting.

The results in Figure 3 demonstrate the substantial bias in the PQL estimates. With matched pairs of binary outcome data, the true variance of the subject specific effects was underestimated by about a half.[6] Even with as many as 40 observations per cluster, the variance was still underestimated by 6%. The degree of bias was affected more by cluster size than by the number of clusters. Indeed, it was worse for $K = 100$ than for $K = 50$. When $\alpha_0$ = -2.5, the bias in $\hat{\theta}_{\text{p}}$ with $n = 40$ was closer to 10%.

CPQL substantially reduced the bias, overcorrecting with 50 clusters

FIGURE 3. Mean values of estimated variance component



of size 2. However, the slow rates at which the averages of the PQL and CPQL estimates approached the true value 1 suggests that cluster sizes might need to be quite large to eliminate entirely the bias in the variance component. As noted previously [6, 24], the bias in the regression coefficients is unimportant once the variance components have been estimated correctly.

## 7 Higher Order Laplace Expansions

The integral in the expression (1) for the likelihood has dimensionality equal to the number of random effects and hence, for many problems of interest, increases with the sample size $n$. Shun and McCullagh [42] and Shun [41] noted that the standard Laplace approximation failed to have an asymptotic ($n \uparrow \infty$) justification in such circumstances, and derived a remainder term that improved its performance. Raudenbush, Yang and Yosef [39] developed a systematic approach to higher order Laplace expansions, and provided details for two-level models involving a series of clusters of independently distributed observations. Here we consider the simplest case, the GLMM with canonical link function where each cluster has a single random effect, in order to illustrate the potential of this approach.

Suppose then that the likelihood may be written

$$\mathrm{L}(\alpha,\theta)=\prod_{i=1}^{K}\mathrm{L}_i(\alpha,\theta)=\prod_{i=1}^{K}\frac{1}{\sqrt{2\pi}}\int\exp\left[\ell_i(\alpha,b_i)-\frac{1}{2\theta}b_i^2\right]db_i,$$

where $K$ is the number of clusters, the $b_i$ are $K$ independently distributed random effects from a normal distribution with mean 0 and variance $\theta$ and the conditional log-likelihoods are

$$\ell_i(\alpha,b_i)=\frac{\phi}{a_i}\sum_{j=1}^{n_i}\left\{[y_{ij}\eta_{ij}^b-\delta(\eta_{ij}^b)]+\gamma(y_{ij};\phi)\right\}$$

with $\eta_{ij}^b=x_{ij}^T\alpha+b_i$ denoting the linear predictor and $n_i$ the number of observations in the $i^{th}$ cluster. Let $\tilde{b}_i=\tilde{b}_i(\alpha,\theta)=\operatorname{argmax}[\ell_i(\alpha,b)-b^2/(2\theta)]$ denote the PQL estimate of the $i^{th}$ random effect and define

$$\tilde{\ell}_i^{(k)}=\left(\frac{\partial}{\partial\beta}\right)^k\ell_i\Big|_{b=\tilde{b}_i}$$

and

$$v_i=-\left[\tilde{\ell}_i^{(2)}-\frac{1}{\theta}\right]^{-1}=\frac{\theta}{1-\theta\tilde{\ell}_i^{(2)}}.$$

Then, using a Taylor expansion,

$$\mathrm{L}_i=\frac{e^{\tilde{\ell}_i-\tilde{b}_i^2/(2\theta)}}{\sqrt{2\pi\theta}}\int\exp\left[-\frac{1}{2v_i}(b-\tilde{b}_i)^2+R_i(b)\right]db,$$

where $R_i(b)=\sum_{k=3}^{\infty}T_{ki}(b)$ with $T_{ki}(b)=\tilde{\ell}_i^{(k)}(b-\tilde{b}_i)^k/k!$. It follows that

$$\begin{aligned}L_i &= \sqrt{\frac{v_i}{\theta}}e^{\tilde{\ell}_i-\tilde{b}_i^2/(2\theta)}\mathrm{E}_i\left[1+R_i(b)+\frac{1}{2}R_i^2(b)+\frac{1}{3}R_i^3(b)+\cdots\right]\\ &= \sqrt{\frac{v_i}{\theta}}e^{\tilde{\ell}_i-\tilde{b}_i^2/(2\theta)}\left[1+\mathrm{E}_i(T_{4i})+\mathrm{E}_i(T_{6i})+\frac{1}{2}\mathrm{E}_i(T_{3i}^2)+\cdots\right]\end{aligned}$$

where $\mathrm{E}_i=\mathrm{E}_{\mathcal{N}(\tilde{b}_i,v_i)}$ denotes expectation with respect to a normal distribution with mean $\tilde{b}_i$ and variance $v_i$.

We evaluate the higher terms in this expansion, and note their asymptotic order in terms of $\theta\downarrow 0$ and the cluster-specific sample size $n\equiv n_i\uparrow\infty$:

$$\begin{aligned}\mathrm{E}_i(T_{4i}) &= \frac{\theta^2\tilde{\ell}_i^{(4)}}{8[1-\theta\tilde{\ell}_i^{(2)}]^2} = O\left(\theta^2\right)\times O\left(\frac{1}{n}\right)\\ \mathrm{E}_i(T_{6i}) &= \frac{\theta^3\tilde{\ell}_i^{(6)}}{48[1-\theta\tilde{\ell}_i^{(2)}]^3} = O\left(\theta^3\right)\times O\left(\frac{1}{n^2}\right)\\ \frac{1}{2}\mathrm{E}_i(T_{3i}^2) &= \frac{15\theta^3[\tilde{\ell}_i^{(3)}]^2}{72[1-\theta\tilde{\ell}_i^{(2)}]^3} = O\left(\theta^3\right)\times O\left(\frac{1}{n}\right).\end{aligned}$$

Quartic expansions, *i.e.*, those involving terms up to $\mathrm{E}_i(T_{4i})$, were considered by several groups interested in approximations valid for small variance components. [43, 27, 6, 24] However, these are inadequate for larger values of $\theta$ no matter what the sample size in each cluster. Approximate inference based on terms up to sixth order, as shown, has been implemented for the logistic-normal model in version 5 of HLM.[38] Terms of all orders, for both univariate and multivariate normal random effects distributions, are available in principle.[39]

TABLE 7. Results of a simulation study of integral approximations

| Parameter | True value | PQL | GH-20 | L-6 |
|---|---|---|---|---|
| | | Averages of estimates | | |
| $\alpha_0$ | -1.200 | -1.090 | -1.205 | -1.201 |
| $\alpha_1$ | 1.000 | 0.900 | 1.015 | 1.003 |
| $\alpha_2$ | 1.000 | 0.911 | 0.998 | 0.998 |
| $\theta_{00}$ | 1.625 | 1.275 | 1.655 | 1.635 |
| $\theta_{01}$ | 0.100 | 0.054 | 0.100 | 0.096 |
| $\theta_{11}$ | 0.250 | 0.161 | 0.256 | 0.267 |
| | | Mean squared errors | | |
| $\alpha_0$ | -1.200 | 0.027 | 0.020 | 0.019 |
| $\alpha_1$ | 1.000 | 0.024 | 0.018 | 0.016 |
| $\alpha_2$ | 1.000 | 0.116 | 0.005 | 0.005 |
| $\theta_{00}$ | 1.625 | 0.152 | 0.063 | 0.056 |
| $\theta_{01}$ | 0.100 | 0.008 | 0.012 | 0.011 |
| $\theta_{11}$ | 0.250 | 0.113 | 0.007 | 0.008 |

Source: Raudenbush, Ying and Yosef [39]

Table 7 reports results of a simulation study of three estimators of paramaters in a logistic-normal random intercepts and slopes model with binary outcomes.[39] Here the $K$=200 clusters were intended to represent communities, in each of which $n = 20$ observations representing children were sampled. Both child and community level covariates were generated from normal distributions. The intercept (mean $\alpha_0$) and slope (mean $\alpha_1$) of the regression on the child level covariate were allowed to vary from community to community with variances $\theta_{00}$ and $\theta_{01}$, respectively, and covariance $\theta_{01}$. In addition to PQL, the estimation methods included 20 point Gauss-Hermite quadrature using software developed by Hedeker and Gibbons [18] (GH-20) and the sixth order Laplace expansion as described above (L-6). Although the average estimation time for quadrature was substantially higher (720 seconds) than for the Laplace approximation (35 seconds), the latter method proved to be at least as accurate.

# 8 Conclusions

The implementation of the PQL algorithm in several commercial software packages has stimulated interest in the procedure and encouraged the use of GLMMs in a wide variety of scientific fields. This review was motivated by the desire to further evaluate the accuracy of the inferences that result from its use in settings that are typical of current practice.

PQL yields the standard REML estimates for normal theory, linear mixed models. In view of the correspondence between mixed models and smoothing splines, it therefore also provides correct inferences for normal theory, semiparametric linear mixed models where certain fixed covariate terms are replaced by nonparametric smooths.[25]

The illustrative analyses presented in §4 and §5, and simulations based on those analyses, suggest that PQL performs adequately for GLMMs with categorical outcomes provided that the nominally Poisson or binomial observations have distributions that are approximately Gaussian. Experience suggests that the algorithm provides reasonable approximations for Poisson outcomes provided that their means are generally greater than 5. An even lower cutoff may be adequate for many problems. With binomial outcomes, a rule of thumb might be that the expected numbers of "successes" and "failures" for each observation should also generally exceed 5. This means that the binomial denominators should be at least 10 for reponse probabilities in the midrange, with larger denominators needed if many of the probabilities were near 0 or 1. This is not all that different from standard guidelines for the practical adequacy of approximate inference procedures such as chi-squared tests for the analysis of contingency tables. However, one cannot hope to have a simple rule cover all contigencies.

For many of these situations where PQL performs well, application of the corresponding linear mixed model (LLM) to transformed outcome data likely will also be adequate. However, results [4] quoted at the end of §4 suggest that, for some sparse data situations, PQL will perform adequately whereas the empirical transform approach may not.

Where PQL has the greatest difficulty is for the analysis of binomial outcomes based on small denominators, especially binary outcomes. With clustered data, the critical feature is the number of conditionally independent binary observations per cluster. As the within cluster sample size increases, so does the information for prediction of the unobserved random effect. The within cluster sum of conditional deviances is then well approximated by a quadratic function of the random effect $b$, and the cluster's contribution to the likelihood (1) is well represented by its Laplace approximation. Lee and Nelder [22, 23] state asymptotic results that formalize this intuition.

As the simulations in §6 and similar evaluations by other authors demonstrate, however, the within cluster sample sizes may need to be quite large indeed before the asympotic results hold and the variance components are correctly estimated. Several groups have developed refinements or modifications to the PQL algorithm in an attempt to improve its performance in this setting. As shown in §6, multiplicative correction of the estimated variance components using CPQL can substantially improve performance in some situations. The PQL2 procedure available with MLwiN likewise improved performance in simulations conducted by its developers.[16, 37] An improved methodology for variance components has also been proposed for "h-likelihood" estimation, the generalization of PQL for HGLMs.[23]

One or more of these modifications should definitely be implemented when analyzing small clusters of binary outcome data using PQL. Since some uncertainty may remain as to whether or not significant bias persists, however, recourse should likely be made also to one of the presumptively more accurate methods developed during the past several years. The adaptive Gauss-Hermite quadrature procedure now available with SAS PROC NLMIXED should suffice for many multi-level, clustered data problems. It is still restricted, however, to situations where a very small number of correlated random effects is observed within each cluster. As illustrated in §7, the higher order Laplace methods developed by Raudenbush, Ying and Yosef [39] hold promise for the analysis of clustered data with higher dimensional random effects. Further commercial implementation of this approach would be desirable, as would implementation of the automated Monte Carlo EM algorithm [3] mentioned briefly in §2.

Finally, for time series, spatial statistics and crossed designs, where it is not possible to reduce the dimensionality of the integrations, approximation based on MCMC simulation is at present the only viable general approach. It is to be hoped that programs for maximum likelihood estimation using this approach will soon become available, as they have for Bayesian inference.

## Acknowledgements:

This work was supported in part by USPHS grant 5-R01-CA40644. I am grateful to Kerrie Nelson for an extensive literature search that identified hundreds of papers on GLMM, PQL and their application.

## 9 References

[1] A. Agresti and J. Hartzel. Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine*, 19:1115–1139, 2000.

[2] P. J. Beitler and J. R. Landis. A mixed-effects model for categorical data. *Biometrics*, 41:991–1000, 1985.

[3] J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, 61:265–285, 1999.

[4] N. Breslow, B. Leroux, and R. Platt. Approximate hierarchical modelling of discrete data in epidemiology. *Statistical Methods in Medical Research*, 7:49–62, 1998.

[5] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[6] N. E. Breslow and X. H. Lin. Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91, 1995.

[7] Z. Chen and L. Kuo. A note on the estimation of the multinomial logit model with random effects. *American Statistician*, 55:89–95, 2001.

[8] D. G. Clayton. Generalized linear mixed models. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 16, pages 275–301. Chapman and Hall, London, 1999.

[9] CYTEL Software Corporation. *EGRET for Windows.* CYTEL Software Corporation, Cambridge, MA, 1999.

[10] D. R. Cox and E. J. Snell. *Analysis of Binary Data, Second Edition.* Chapman and Hall, London, 1989.

[11] P. J. Davis and I. Polonsky. Numerical interpolation, differentiation and integration. In M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions*, chapter 25, pages 875–924. U.S. Government Printing Office, Washington, D.C., 1964.

[12] B. Engel and A. Keen. A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, 48:1–22, 1994.

[13] C. J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–511, 1992.

[14] H. Goldstein. Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78:45–51, 1991.

[15] H. Goldstein. *Multilevel Statistical Models.* Edward Arnold, London, 1995.

[16] H. Goldstein and J. Rasbash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, 159:505–513, 1996.

[17] S. Greenland. Hierarchical regression for epidemiologic analyses of multiple exposures. *Environmental Health Perspectives*, 102:33–39, 1994.

[18] D. Hedeker and R. D. Gibbons. MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49:157–176, 1996.

[19] C. R. Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447, 1975.

[20] A. Y. C. Kuk and Y. W. Cheng. The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computation and Simulation*, 59:233–250, 1997.

[21] A. Y. C. Kuk and Y. W. Cheng. Pointwise and functional approximations in Monte Carlo maximum likelihood estimation. *Statistics and Computing*, 9:91–99, 1999.

[22] Y. Lee and J. A. Nelder. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:619–678, 1996.

[23] Y. Lee and J. A. Nelder. Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88:987–1006, 2001.

[24] X. Lin and N. E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016, 1996.

[25] X. Lin and D. Zhang. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, 61:381–400, 1999.

[26] R. C. Littell, G. A. Milliken, W. W Stroup, and R. D. Wolfinger. *SAS System for Mixed Models*. SAS Institute Inc., Cary, N.C., 1996.

[27] Q. Liu and D. A. Pierce. A note on gauss-hermite quadrature. *Biometrika*, 81:624–629, 1994.

[28] P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall, London, 1989.

[29] C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170, 1997.

[30] C. E. McCulloch and S. R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, New York, 2001.

[31] C. A. McGilchrist. Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B*, 56:61–69, 1994.

[32] K. J. McKonway, M. C. Jones, and P. C. Taylor. *Statistical Modelling using GENSTAT*. Arnold, London, 1999.

[33] R. B. Millar and T. J. Willis. Estimating the relative density of snapper in and around a marine reserve using a log-linear mixed-effects model. *Australian and New Zealand Journal of Statistics*, 41:383–394, 1999.

[34] J. Myles and D. Clayton. *GLMMGibbs: An R Package for Estimating Bayesian Generalised Linear Mixed Models by Gibbs Sampling*. Imperial Cancer Research Fund, London, 2001.

[35] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.

[36] J. Pinheiro and D. M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35, 1995.

[37] J. Rasbash, W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G Woodhouse, D. Draper, I Langford, and T. Lewis. *A User's Guide to MLwiN*. Institute of Education, London, 2000.

[38] S. W. Raudenbush, A. S. Byrke, Y. F. Cheong, and R Congdon. *HLM 5: Hierarchical Linear and Nonlinear Modeling*. Scientific Software International, Lincolnwood, IL, 2000.

[39] S. W. Raudenbush, M. L. Yang, and M. Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9:141–157, 2000.

[40] R. Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727, 1991.

[41] Z. M. Shun. Another look at the salamander mating data: A modified Laplace approximation approach. *Journal of the American Statistical Association*, 92:341–349, 1997.

[42] Z. M. Shun and P. McCullagh. Laplace approximation of high-dimensional integrals. *Journal of the Royal Statistical Society, Series B*, 57:749–760, 1995.

[43] P. J. Solomon and D. R. Cox. Nonlinear component of variance models. *Biometrika*, 79:1–11, 1992.

[44] D. J. Spiegelhalter, A. Thomas, N. G. Best, and W. R. Gilks. *BUGS: Bayesian Inference using Gibbs Sampling, Version 0.30.* Medical Research Council Biostatistics Unit, Cambridge, 1994.

[45] SAS Institute Inc. Staff. The NLMIXED procedure. In *SAS/STAT User's Guide Version 8*, chapter 46, pages 2421–2504. SAS Publishing, Cary, NC, 2000.

[46] R. Stiratelli, N. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971, 1984.

[47] P. F. Thall and S. C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46:657–671, 1990.

[48] R. Wolfinger. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80:791–795, 1993.

[49] R. Wolfinger and M. O'Connell. Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243, 1993.

[50] S. L. Zeger and M. R. Karim. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86, 1991.