Lifetime Data Anal (2007) 13:497–512
DOI 10.1007/s10985-007-9065-x

# Generalized linear mixed models: a review and some extensions

**C. B. Dean · Jason D. Nielsen**

Received: 2 October 2007 / Accepted: 3 October 2007 / Published online: 14 November 2007

**Abstract** Breslow and Clayton (J Am Stat Assoc 88:9–25,1993) was, and still is, a highly influential paper mobilizing the use of generalized linear mixed models in epidemiology and a wide variety of fields. An important aspect is the feasibility in implementation through the ready availability of related software in SAS (SAS Institute, PROC GLIMMIX, SAS Institute Inc., URL http://www.sas.com, 2007), S-plus (Insightful Corporation, S-PLUS 8, Insightful Corporation, Seattle, WA, URL http://www.insightful.com, 2007), and R (R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, 2006) for example, facilitating its broad usage. This paper reviews background to generalized linear mixed models and the inferential techniques which have been developed for them. To provide the reader with a flavor of the utility and wide applicability of this fundamental methodology we consider a few extensions including additive models, models for zero-heavy data, and models accommodating latent clusters.

**Keywords** Generalized linear mixed model · Random effects · Longitudinal data analysis · Penalized quasi-likelihood

---

C. B. Dean (✉)
Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC,
Canada V5A 1S6
e-mail: dean@stat.sfu.ca

J. D. Nielsen
School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada K1S 5B6
e-mail: jdn@math.carleton.ca

## 1 Introduction

Generalized linear models (GLM; McCullagh and Nelder 1989) provide an extension of linear models which relaxes the assumptions of normality, constant error variance and a linear relationship between the covariate effects and the mean. The introduction of GLMs produced a unified likelihood regression approach for the analysis of a wide range of continuous and discrete outcomes. More recently, there have been a wealth of developments allowing many complicated hierarchical additive and mixture models particularly for the analysis of count and binomial data. Many special cases have also been considered as this area is driven by motivating applications from diverse fields each with some special features. Indeed, the generalized linear model and its extensions are fundamental in epidemiology; for example, Stangle et al. (2007) investigate a lead exposure treatment, Dubin et al. (2007) conduct a study to see if triggered sampling reduces bias in longitudinal studies subject to dropout, Rich-Edwards et al. (2005) analyze pre-term delivery in Boston before and after September 11, 2001 using a mixed logistic model and Kleinman et al. (2004) study small area disease incidents in order to detect biological terrorism.

A GLM is specified by a random component which specifies the probability distribution of the response variable, a systematic component, which specifies a linear function of the explanatory variables used as the predictor, and the link function which relates the systematic component and the mean value of the random component. The random component consists of independent observations $y$ from a distribution in the exponential family

$$f(y) = \exp\left\{ \frac{y\,\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specified functions, $\theta$ is the canonical parameter and $\phi$ is the dispersion parameter. The mean and variance of $Y$ are $\mu = E[Y] = b'(\theta)$ and $Var[Y] = a(\phi)v(\mu)$, where $v(\mu) = b''(\theta)$; dependence of the variance function on $\mu$ arises through the canonical parameter $\theta$. Covariates are introduced in the systematic component, $\eta = g(\mu) = \mathbf{x}^T\boldsymbol{\alpha}$, where $\mathbf{x}$ is a vector of covariates and $\boldsymbol{\alpha}$, a vector of regression parameters; $g(\mu)$ is the link function. If $\theta = \mu$, then the link $g$ is the canonical link.

For many years, these models were standard for the analysis of count and binomial data. Inference using maximum likelihood is straightforward and diagnostic procedures were developed including tests for the link function (Cheng and Wu 1994), residual diagnostics (Pierce and Schafer 1986; Cook and Weisberg 1989), and tests for the common problem of overdispersion (Breslow 1989). Wedderburn's (1974) formulation of inference using quasi-likelihood estimation assumed only independence and a functional relationship between the mean and variance. The quasi-likelihood function is expressed as

$$\int_y^{\mu} \frac{y - t}{a(\phi)v(t)}\,\mathrm{d}t$$

and the estimating equations for $\boldsymbol{\theta}$ based on a sample $\mathbf{y} = [y_1, \ldots, y_n]^T$ are well-known as

$$\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \tag{1}$$

where $\mathbf{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\theta}$, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^T$, $\mu_i = E[Y_i]$ and $\mathbf{V} = \mathrm{diag}\,\{a_i(\phi)v(\mu_i)\}$. Quasi-likelihood extends the scope of GLMs by utilizing only the first two moment assumptions concerning the response variable. Quasi-likelihood estimates are consistent and asymptotically normal (White 1982) and need not correspond to a specific probability distribution for $y$. Nelder and Pregibon (1987) further introduced extended quasi-likelihood where $Var[Y]$ need not be a known function of $\mu$ allowing more general variance structures. McCullagh and Nelder (1989) showed that the extended quasi-likelihood is the unnormalized saddlepoint approximation for the exponential family (Barndorff-Nielsen and Cox 1979).

With the notion that overdispersion is an important problem in the analysis of count data which leads to underestimation of the standard error estimates of the covariate effects, several methods for handling this phenomenon were developed. Likelihood approaches incorporating correlation in the data such as through the negative binomial alternative to the Poisson were commonplace. Generalized estimating equations (GEEs; Liang and Zeger 1986) were developed as a multivariate analogue of quasi-likelihood for longitudinal studies where correlation is likely. With the GEE approach the regression analysis is the prime focus and modeling of correlations less important. The GEEs take the same form as (1) but use a working covariance $\mathbf{V}$ for $Var[\mathbf{Y}]$.

In other situations, special cases which extend quasi-likelihood to include unknown parameters in the variance function were also considered. For example, Breslow (1989), in an analysis of pock counts showing substantial overdispersion, utilized a statistical model with $Var[Y] = \phi\mu^\lambda$ and employed pseudo-likelihood (Davidian and Carroll 1987) for estimating variance components. Pseudo-likelihood estimating equations are derived under the assumption that the residuals are well-approximated by a normal distribution. Inferential methods for means and variance components were developed.

Random effects models have also been popular with development mimicking that for linear mixed models. The random effects typically represent cluster frailty terms that impart correlation within clusters. In some situations these frailties are of prime importance, for example where they may measure hospital cluster effects or spatial small-area risks.

These models led to the development of generalized linear mixed models (GLMM) where several hierarchies and crossings of clustering effects may be considered beyond the random intercept model. The development and utilization of GLMMs was initially hampered by the cumbersome intractable integrals required for estimation. A host of inferential approaches were offered including penalized quasi-likelihood (PQL), Monte Carlo EM, simulated maximum likelihood as well as variants of GEEs.

Section 2 reviews approximate and marginal methods for inference in generalized linear mixed models. In Sect. 3 we discuss some specific extensions to provide a few examples of recent developments to this methodology. The extensions considered include additive models, discrete mixtures, including mixtures of nonhomogeneous

Poisson processes, and component mixtures where one component is a degenerate distribution, such as zero-inflated count data models. Section 4 considers the important topic of model checking and diagnostics. Section 5 discusses software for implementation of GLMMs.

## 2 Generalized linear mixed models

Generalized linear mixed models (GLMM) incorporate a wide variety of random effects mimicking the familiar framework for linear mixed models, and thus permitting the flexibility associated with classical linear models with multiple levels of random effects incorporated in various hierarchies. We adopt the structure and notation here as in Breslow and Clayton (1993). Let $\mathbf{Y}$ be a response vector of length $n$ and let $\mathbf{X}$ and $\mathbf{Z}$ define $n \times p$ and $n \times q$ matrices of explanatory variables associated with fixed and random effects, respectively. Conditional on the $q$-dimensional vector of random effects, $\mathbf{b}$, the observations $Y_i$, $i = 1, \ldots, n$ are independent with $E[Y_i|\mathbf{b}] = \mu_i^{\mathbf{b}}$ and variance $Var[Y_i|\mathbf{b}] = \phi\, a_i^{-1} v(\mu_i^{\mathbf{b}})$, where $v(\cdot)$ is a known variance function, and $a_i$ is a known constant. The conditional mean is related to the linear predictor through the link function: $\boldsymbol{\eta}^{\mathbf{b}} = [g(\mu_1^{\mathbf{b}}), \ldots, g(\mu_n^{\mathbf{b}})]^T = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b}$ where $\boldsymbol{\mu}^{\mathbf{b}} = [\mu_1^{\mathbf{b}}, \ldots, \mu_n^{\mathbf{b}}]^T$. The random effects may introduce clustering, spatial correlation and other forms of dependence among outcomes and are assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\mathbf{R}(\boldsymbol{\theta})$, $\boldsymbol{\theta}$ being the variance components. The integrated quasi-likelihood is proportional to

$$|\mathbf{R}|^{-1/2} \int \exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i; \mu_i^{\mathbf{b}}) - \frac{1}{2}\mathbf{b}^T\mathbf{R}^{-1}\mathbf{b}\right\} d\mathbf{b} \tag{2}$$

where $d_i(y; \mu) = -2\int_y^{\mu} a_i(y-u)/v(u)\, du$ is the conditional deviance (Breslow and Clayton 1993). If, conditional on $\mathbf{b}$, $Y_i$ is a member of the exponential family, then $-d_i(y_i; \mu_i^{\mathbf{b}})/(2\phi)$ is the conditional log-likelihood of $Y_i$ given $\mathbf{b}$, and $E_{\mathbf{b}}\big[\sum_{i=1}^{n} d_i(y_i; \mu_i^{\mathbf{b}})/(2\phi)\big]$ is the log-likelihood function.

Random effects are useful for accommodating the heterogeneity often seen in longitudinal data. Such heterogeneity may arise from spatial correlation, subject-specific frailties or clustering effects, for example. Though GLMMs provide considerable opportunities for advancement and flexibility in modeling count and binary data, inference for these models is complicated by the integrals in the marginal maximum likelihood or quasi-likelihood estimating equations. With some simple special cases, such as the random intercept Poisson model, a variety of estimation techniques may be adapted leading to straightforward approaches. In this case, if the exponential of the random intercept term is gamma distributed, the marginal likelihood is negative binomial and maximum likelihood estimating equations for the parameters are available in closed form (Lawless 1987). Alternatively, if no distributional assumptions on the random effects are preferred, Lawless and Zhan (1998) suggest the use of Poisson estimating equations for the parameters in the mean but use a robust standard error estimator to account for any overdispersion. McCullagh and Nelder (1989) also use the Poisson estimating equations and inflate the standard errors by a factor determined by the ratio of the Pearson goodness-of-fit statistic to its degrees of freedom. With more complicated models, neither marginal nor quasi-likelihood estimating equations are available in closed form. The simplest method for fitting such models uses the Laplace approximation (Tierney et al. 1989) and is called penalized quasi-likelihood (PQL; Breslow and Clayton 1993). If we apply the Laplace approximation to the integrated quasi-likelihood (2), estimates of $[\boldsymbol{\alpha}^T, \mathbf{b}^T]^T$ for fixed $\boldsymbol{\theta}$ are obtained by maximizing the penalized quasi log-likelihood

$$-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i;\mu_i^{\mathbf{b}}) - \frac{1}{2}\mathbf{b}^T\mathbf{R}^{-1}\mathbf{b}$$

or equivalently by solving the following estimating equations

$$\mathbf{U}_{\boldsymbol{\alpha}} = \mathbf{X}^T\mathbf{W}^{-1}(\mathbf{Y}-\boldsymbol{\mu}^{\mathbf{b}}) = \mathbf{0} \tag{3}$$

$$\mathbf{U}_{\mathbf{b}} = \mathbf{Z}^T\mathbf{W}^{-1}(\mathbf{Y}-\boldsymbol{\mu}^{\mathbf{b}}) - \mathbf{R}^{-1}\mathbf{b} = \mathbf{0} \tag{4}$$

where $\mathbf{W} = \operatorname{diag}\left\{\phi\, a_i^{-1} v(\mu_i^{\mathbf{b}})[g'(\mu_i^{\mathbf{b}})]^2\right\}$. The solution to (3) and (4) is equivalently obtained by iteratively solving the linear system:

$$\begin{bmatrix} \mathbf{X}^T\mathbf{W}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{W}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W}^{-1}\mathbf{X} & \mathbf{R}^{-1}+\mathbf{Z}^T\mathbf{W}^{-1}\mathbf{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{W}^{-1}\mathbf{Y}^* \\ \mathbf{Z}^T\mathbf{W}^{-1}\mathbf{Y}^* \end{bmatrix} \tag{5}$$

where $\mathbf{Y}^* = \boldsymbol{\eta}^{\mathbf{b}} + (\mathbf{Y}-\boldsymbol{\mu}^{\mathbf{b}})[g'(\mu_i^{\mathbf{b}})]_{n\times 1}$, the so-called working vector (McCullagh and Nelder 1989). The linear system (5) is readily recognizable as the mixed model equations of Harville (1977) for the normal theory model $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \mathbf{W})$, $\mathbf{b} \sim N(0, \mathbf{R})$, $\boldsymbol{\epsilon}$ and $\mathbf{b}$ are independent so that $\mathbf{Y}^* \sim N(\mathbf{X}\boldsymbol{\alpha}, \mathbf{V})$, with $\mathbf{V} = \mathbf{W} + \mathbf{Z}\mathbf{R}\mathbf{Z}^T$. Due to this relationship it is simple to show that for fixed values of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, estimates for the random effects are given by $\hat{\mathbf{b}} = \mathbf{R}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\alpha})$. Similarly, restricted maximum likelihood (REML) estimating equations for $\boldsymbol{\theta}$ are given by $\mathbf{U}_{\boldsymbol{\theta}} = [U_{\theta_1}, \ldots, U_{\theta_r}]^T = \mathbf{0}$ where

$$U_{\theta_s} = (\mathbf{Y}^* - \mathbf{X}\hat{\boldsymbol{\alpha}})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_s}\mathbf{V}^{-1}(\mathbf{Y}^* - \mathbf{X}\hat{\boldsymbol{\alpha}}) - \operatorname{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_s}\right) = 0 \tag{6}$$

with $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. A robust and efficient algorithm for finding estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\mathbf{b}}$ and $\hat{\boldsymbol{\theta}}$ is carried out by solving (3) and (4) for an initial starting value $\boldsymbol{\theta}_{\text{old}}$ of $\boldsymbol{\theta}$. The value of $\boldsymbol{\theta}$ is then updated to $\boldsymbol{\theta}_{\text{new}}$ by a single Newton step (for example, Breslow and Clayton 1993)

$$\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} + \mathbf{H}^{-1}\mathbf{U}_{\boldsymbol{\theta}_{\text{old}}}$$

where $\mathbf{H} = \{h_{ij}\}_{r\times r}$ with $h_{ij} = \operatorname{tr}\left(\mathbf{P}\partial\mathbf{V}/\partial\theta_i\ \mathbf{P}\partial\mathbf{V}/\partial\theta_j\right)$ evaluated at $\boldsymbol{\theta}_{\text{old}}$. This process is repeated until convergence. Numerical stability of the algorithm can be improved by working with the non-zero elements of the Cholesky factor, $\mathbf{L}$, a lower triangular matrix, obtained from the decomposition $\mathbf{R} = \mathbf{L}\mathbf{L}^T$, which ensures that $\mathbf{R}$ remains positive definite (Lindstrom and Bates 1988). Alternatively one could also work with the reciprocal of the non-zero eigenvalues and their associated eigenvectors, the so-called Moore-Penrose inverse (Harville 1977), but this approach is less computationally efficient.

Breslow and Clayton (1993) was a highly influential paper allowing the adoption of several variance components in mixed Poisson and binomial models. Application areas for these models are widespread and the PQL approach affords the scientific community the ability to handle complicated analyses with relative ease. Science indicators of innovation and excellence typically rate the most cited papers in a 10-year period. In 2004, the ISI Essential Science Indicator identified Breslow and Clayton (1993) as the most cited paper in mathematics over the previous decade.

There are many approximations made in deriving the PQL estimators but even so they generally work very well for discrete data analysis with moderate to large cell frequencies. When data are sparse, Breslow and Clayton (1993) demonstrate that PQL has the potential to seriously underestimate parameters, especially variance components. For binary regression models with nested random effects, for example, the approximation of the conditional deviance by the Pearson chi-square statistic is unsatisfactory, and the ignored higher order terms in the Taylor expansion of the Laplace approximation are not negligible even as the sample size increases. So caution when using PQL for inference needs to be taken, particularly with binary outcomes. Breslow and Lin (1995) and Lin and Breslow (1996) develop bias-corrected terms for sparse data scenarios. The correction is derived by taking a Taylor expansion of the PQL variance estimating equations (6) about $\boldsymbol{\theta} = \mathbf{0}$. For the case of a single variance component, the bias-correction is based on the fact that asymptotic bias is approximately a linear function of $\theta$ in the neighborhood of the origin. The proposed bias corrections often inflate variances, especially when both the variance component and sample size are small. However, bias corrections are quite advantageous for the difficult binary data context when the sample size is even moderately large. Raudenbush et al. (2000) also investigate bias reduction in the GLMM context by retaining the higher order terms of the Taylor expansion in the Laplace approximation (Shun and McCullagh 1995). These remainder terms are functions of multivariate normal random variables and so closed form expressions of the expectations can be obtained, though these expressions are complicated functions of the higher order derivatives of the quasi log-likelihood. Their results are promising and in simulations compare favorably to full maximum likelihood.

Other proposals for estimation include Monte Carlo EM; see, for example, Chen et al. (2002) who also propose to relax the distributional assumptions of the random effects and use a rejection sampling scheme to estimate parameters in the mean and variance components. Zeger and Karim (1991) adopt Gibbs sampling (Casella and George 1992), a Markovian updating scheme for estimating a posterior distribution. Gibbs sampling is particularly useful when the conditional distribution of $\boldsymbol{\alpha}$ given $\mathbf{b}$ is approximated by a multivariate Gaussian distribution while the more difficult conditional distribution of $\mathbf{b}$ given $\boldsymbol{\alpha}$, $\mathbf{R}$ and $\mathbf{Y}$ is also Gaussian approximated. Maximum likelihood with Monte Carlo EM, simulated method of moments and simulated maximum likelihood approaches have also been considered (McCulloch 1997; Jiang 1998). If the dimension of $\mathbf{b}$ is not too large, say less than 5, directly computing the marginal likelihood by adaptive Gaussian quadrature can be done efficiently (Pinheiro and Bates 2000).

When the regression objective is primary and the nature of the correlation and variance components secondary, generalized estimating equations (Liang and Zeger 1986) provides a marginal estimation approach which is very popular and simple to implement. GEEs are robust with respect to variance misspecification and offer consistent estimates of $\boldsymbol{\alpha}$, under standard regularity conditions, provided that the model for $E[\mathbf{Y}] = \boldsymbol{\mu}$ is correctly specified. An extension to GEEs and PQL using estimating equations involving first and second order conditional moments is discussed by Vonesh et al. (2002). The variance-covariance parameters of the random effects are estimated using a Gaussian posterior approximation of Laird and Louis (1982). Note that few approaches are available which incorporate covariate effects into the variance components. Lin et al. (1997) consider such a scenario with heterogeneous within-cluster variances in a cluster analysis, the heterogeneity explained through regression modeling.

Conditional approaches may also be used when random effects are not of prime importance. Let $\mathbf{s}$ be a sufficient statistics for $\mathbf{b}$ for fixed and known $\boldsymbol{\alpha}$. Then $f(\mathbf{y}|\mathbf{b}) = f(\mathbf{y}|\mathbf{s})f(\mathbf{s}|\mathbf{b})$ and

$$\mathcal{L}(\boldsymbol{\alpha}) = f(\mathbf{y}|\mathbf{s}) \int f(\mathbf{s}|\mathbf{b})\, \mathrm{d}F(\mathbf{b}).$$

The first term is called the conditional likelihood and has been used for inference on $\boldsymbol{\alpha}$. For example, consider a mixed Poisson model where, given the random effects $\nu_i = \exp(b_i)$, the distribution of $Y_{ij}$ is Poisson with mean $\nu_i \mu(\mathbf{x}_{ij}; \boldsymbol{\alpha}) = \nu_i \mu_{ij}$ where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of fixed covariates including a constant intercept term. Here $i$ might index individuals and $j$ the times at which observations are recorded; or $i$ might represent clusters and $j$ individuals within these clusters, $j = 1, \ldots, e_i$, $i = 1, \ldots, n$. Given $\nu_i$, the $Y_{ij}$'s, $j = 1, \ldots, e_i$ are independent variates. Assume that the $\nu_i$'s are i.i.d. random variables with probability density function $m(\nu; \tau)$, depending on a parameter $\tau$ with mean 1 and variance $\tau$. Unconditionally, $E[\mathbf{Y}_i] = \boldsymbol{\mu}_i$ and $Var[\mathbf{Y}_i] = \text{diag}\left\{\boldsymbol{\mu}_i\right\} + \tau \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T$ where here $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{ie_i}]^T$ and $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{ie_i}]^T$. When $\nu_i, i = 1, \ldots, n$ is held fixed, $Y_{i+} = \sum_j Y_{ij}, i = 1, \ldots, n$ is sufficient for $\mathbf{b}$ and the conditional likelihood is given by

$$\frac{\exp\left(\sum_{i,j} y_{ij}\ \mathbf{x}_{ij}^T \boldsymbol{\alpha}\right)}{\prod_i \left\{\sum_j \exp(\mathbf{x}_{ij}^T \boldsymbol{\alpha})\right\}^{y_{i+}}}.$$

Sartori and Severini (2004) discuss the use of conditional likelihoods and their relationship to marginal likelihoods. They discuss a few important and simple special cases to develop ideas in depth, including issues of identifiability. The conditional likelihood approach has been successfully implemented in simple random intercept models involving binary outcomes (Laird 1991; Tjur 1982).

Recent work by Song et al. (2005) on algorithms for finding maximum likelihood estimates when the structure of the log-likelihood yields a decomposition may facilitate new inferential techniques for GLMMs. The decomposition is such that one part of the log-likelihood is more simply analyzed while the second, more complicated part, is used to update estimates from the first. This also has implications for efficiency studies where simpler components of the likelihood may provide highly efficient estimators (Dean and Balshaw 1997).

In the following section we illustrate the utility of GLMMs and extensions of GLMMs through several motivating examples.

## 3 Some special cases and extensions

### 3.1 Generalized mixed Poisson models

Let $Y_1, \ldots, Y_n$, conditional on subject specific random effects $\nu_i$, be independent Poisson random variables with means $\nu_i \mu_i$ where $\log(\mu_i) = \exp\{\mathbf{x}_i\, \boldsymbol{\alpha}\}$; a conditional log-linear Poisson regression model. Further assume that the $\mathcal{V}_i \overset{i.i.d}{\sim} GIG(\tau, \omega)$ where $GIG(\tau, \omega)$ denotes the generalized inverse Gaussian distribution (Jørgensen 1982) with density,

$$m(y; \tau, \omega) = \frac{1}{2K_\tau(\omega)} y^{\tau-1} \exp\left\{-\frac{1}{2}\omega(y^{-1} + y)\right\} \mathbf{1}(y > 0) \quad \text{on}$$

$$\Theta = \begin{cases} \omega \geq 0, & \tau > 0 \\ \omega > 0, & \tau = 0 \\ \omega \geq 0, & \tau < 0 \end{cases}$$

and $K_\tau(\cdot)$ is the modified Bessel function of the second kind (Abramowitz and Stegun 1984). This is a very flexible class of distributions that includes as special cases: the gamma ($\omega \to 0, \tau > 0$), the reciprocal gamma ($\omega \to 0, \tau < 0$), the inverse Gaussian ($\tau = -1/2$), the reciprocal inverse Gaussian ($\tau = 1/2$) and the hyperbola distribution ($\tau = 0$). The unconditional distribution of each $Y_i$ is then given by

$$\begin{aligned} f_{Y_i}(y_i; \tau, \omega) &= \int_0^\infty \frac{(\nu_i \mu_i)^{y_i} \exp\{-\nu_i \mu_i\}}{y_i!} m(\nu_i; \tau, \omega)\, \mathrm{d}\nu_i \\ &= \frac{\mu_i^{y_i}}{y_i!} \left(\frac{\omega}{\sqrt{\omega^2 + 2\omega\mu_i}}\right)^{\tau + y_i} \frac{K_{\tau+y_i}\left(\sqrt{\omega^2 + 2\omega\mu_i}\right)}{K_\tau(\omega)} \end{aligned} \tag{7}$$

for $y_i = 0, 1, \ldots$. This is the so-called Sichel distribution (Sichel 1974) with mean $E[Y_i] = R_\tau(\omega)\mu_i$ where $R_\tau(\omega) = K_{\tau+1}(\omega)/K_\tau(\omega)$ and variance

$$V[Y_i] = R_\tau(\omega)\mu_i + \left(K_{\tau+2}(\omega)/K_\tau(\omega) - R_\tau(\omega)^2\right) \mu_i^2.$$

Special cases of this mixed Poisson regression model of interest are the negative binomial (Lawless 1987)

$$\lim_{\omega \to 0} f_{Y_i}(y_i; \tau, \omega) = \frac{\Gamma(y_i + \tau)}{y_i!\, \Gamma(\tau)} \left( \frac{\mu_i}{\tau + \mu_i} \right)^{y_i} \left( \frac{\tau}{\tau + \mu_i} \right)^{\tau}, \quad \tau > 0$$

and Poisson inverse Gaussian distributions (Dean et al. 1989) when $\tau = -1/2$.

### 3.2 Zero-inflated count models

Although a modest number of excess zeros can be accommodated in usual mixed Poisson regression models, these models are generally unable to handle zero-heaviness beyond a certain threshold, or situations where zero-heaviness leads to a bi-modal distribution. In such situations, or when there is scientific reasoning to postulate that there is a subset of the population which generates only zero counts, mixed Poisson models are not appropriate. An alternative approach utilizes a finite mixture model whereby one component is a degenerate distribution with mass at zero and the other a non-degenerate distribution such as a mixed Poisson model. A simple zero-inflated model for count data was introduced by Lambert (1992) as a mixture of degenerate and Poisson distributions, the so-called “zero-inflated Poisson” (ZIP) model.

Let $Y_1, \ldots, Y_n$ be independent random variables such that

$$\mathbf{Y}_i \sim \begin{cases} 0, & \text{with probability} \, p_i \\ f_{Y_i}(y_i; \tau, \omega), & \text{with probability} \, 1 - p_i \end{cases}$$

where $f_{Y_i}(y_i; \tau, \omega)$ is Sichel defined in (7) with $\log(\mu_i) = \mathbf{x}_{1i}\boldsymbol{\alpha}_1$ and $\log\left(p_i/(1 - p_i)\right) = \mathbf{x}_{2i}\boldsymbol{\alpha}_2$; $\mathbf{x}_{ji}$ and $\boldsymbol{\alpha}_{ji}$, $j = 1, 2$ being vectors of covariate effects and their associated parameters. This defines a class of zero-inflated count regression models including the popular zero-inflated negative binomial (ZINB) model (Nodtvedt et al. 2002; Simons et al. 2006; Martin et al. 2005) and ZIP models. Zero-inflated models for geo-referenced data are considered by Ainsworth and Dean (2007). This class of models is interesting as it can be viewed as a mixture model where the mixing distribution has both discrete and continuous components.

### 3.3 Generalized additive mixed spline models

Lin and Zhang (1999) proposed the use of generalized additive mixed models to provide greater flexibility in the manner in which covariate effects modulate the mean. The very general formulation they use models the linear predictor as

$$g(\mu_i^{\mathbf{b}}) = \alpha_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \mathbf{z}_i^T \mathbf{b}$$

where $f_1, \ldots, f_p$ are assumed to be twice differentiable functions modeled as smoothing splines Gree and Silverman 1994, $\alpha_0$ is a fixed intercept and all other quantities are as previously defined. In matrix notation this can then be written as

$$\boldsymbol{\eta}^{\mathbf{b}} = \mathbf{1}\alpha_0 + \mathbf{Q}_1\mathbf{f}_1 + \cdots + \mathbf{Q}_p\mathbf{f}_p + \mathbf{Zb}$$

where $\mathbf{f}_j = [f_j(x_{1j}), \ldots, f_j(x_{r_j,j})]^T$, a vector of values of $f_j$ evaluated at the $r_j$ distinct ordered values of $x_{ij}/\max_j\{x_{ij}\}$, $i = 1, \ldots, r_j$, $\mathbf{1}$ is a vector of ones and $\mathbf{Q}_j$ is an $n \times r_j$ matrix such that the $i$th component of $\mathbf{Q}_j\mathbf{f}_j$ is $f_j(x_{ij})$. The $f_j$ are assumed centered so the restriction $\mathbf{1}^T\mathbf{f}_j = 0$ is enforced, $j = 1, \ldots, p$ and $\mathbf{f}_j$ can be expressed as

$$\mathbf{f}_j = \mathbf{T}_j\gamma_j + \mathbf{E}_j\mathbf{c}_j$$

where $\mathbf{T}_j$ is a $r_j \times 1$ vector of the centered, ordered distinct values of $x_{ij}/\max_j\{x_{ij}\}$, $\mathbf{E}_j = \mathbf{L}_j(\mathbf{L}_j^T\mathbf{L}_j)^{-1}$ where $L_j$ is the full rank $r_j \times (r_j - 2)$ matrix satisfying $\mathbf{K}_j = \mathbf{L}_j\mathbf{L}_j^T$ with $\mathbf{K}_j = \{k_{ij}\}_{r \times r}$ being the usual penalty matrix for a smoothing spline with this parameterization (Eq. 2.3; Green and Silverman 1994) and derived using the penalty operator $\int(f'')^2$. Given this specification, estimates of $\alpha_0$, $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_p]^T$, $\mathbf{b}$ and $\mathbf{c}_j$, $j = 1, \ldots, p$ are found by maximizing

$$-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i; \mu_i^{\mathbf{b}}) - \frac{1}{2}\mathbf{b}^T\mathbf{R}^{-1}\mathbf{b} - \mathbf{c}^T\mathbf{P}\mathbf{c} \tag{8}$$

where $\mathbf{P} = \text{diag}\left\{\tau_1\mathbf{I}_{r1}, \ldots, \tau_p\mathbf{I}_{r_p}\right\}$ and $\mathbf{c} = [\mathbf{c}_1^T, \ldots, \mathbf{c}_p^T]^T$. From (8) it is clear why Lin and Zhang (1999) refer to this method as double PQL (DPQL) as there are penalties for both the spline terms and the random effects. This extends the work of Hastie and Tibshirani (1999) by incorporating random effects to deal with clustering or spatial correlations, for example, and also results in automatic selection of the amount of smoothing by estimating the smoothing parameters $\tau_j$, $j = 1, \ldots, p$ via REML.

Henderson and Shimakura (2003) describe a study of patient controlled analgesia which permits patients to administer pain relief according to individual needs after surgery. The 30 patients in the control and the 35 in the treatment group receive morphine with those in the treatment group receiving half the dosage (1 mg) at any treatment request. The patients were followed for 48 h with the number of requests per 4-h period recorded. Note there are restrictions to prevent overdose but no patient seemed constrained by these with the maximum number of requests substantially lower than constraints would provide.

Lin and Zhang's (1999) additive mixed spline model was fit to this data with control $f_1(t)$ and treatment effects $f_2(t)$ modeled by temporal splines and incorporating individual-specific subject effects. Estimates of $f_1(t)$ and $f_2(t)$ with 95% confidence intervals are provided in Fig. 1. These plots indicate that the overall intake of morphine is generally decreasing for both treatments with the (low dose) treatment group requesting the drug at a slightly higher rate but much less than twice as often.

### 3.4 Clustered mixed nonhomogeneous Poisson process spline models for panel data

Let $\{Y_i(t), t \in [T_{i0}, T_{ie_i})\}$ be a counting process governing the number of events experienced by subject $i$ over the period of observation, $[T_{i0}, T_{ie_i})$. Suppose that the

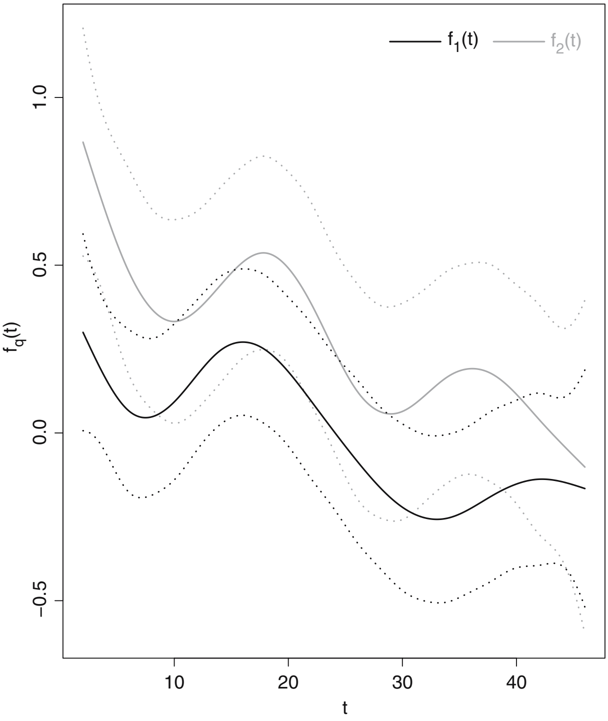

**Fig. 1** Plots of estimates of $f_1(t)$ (control—black line) overlayed with that for $f_2(t)$ (treatment—gray line) along with their associated 95% point-wise confidence intervals

process is not followed continuously and only panel data are available, i.e. $Y_{ij} = Y_i[T_{i(j-1)}, T_{ij})$, representing the number of events occurring between follow-up times $T_{i0} < T_{i1} < T_{i2} < \cdots < T_{ie_i}$, $i = 1, \ldots, n$, $j = 1, \ldots, e_i$. Where counts arise from latent classes Nielsen and Dean (2007) describe such processes by assuming that the counts for subject $i$ are generated by a counting process $Y_i(t) = \sum_{g=1}^{G} z_{gi} C_{gi}(t)$, $t \in [0, T_{ie_i})$ where events for each subject are assumed to be generated from one of $G$ sub-processes $C_{gi}(t)$ and $z_{gi}$ denotes an unknown indicator of sub-group membership. Conditional on group-specific individual frailty terms, $\nu_{gi}$, each $C_{gi}(t)$ is assumed to follow a nonhomogeneous Poisson process with intensity function $\nu_{gi}\ \lambda_{gi}(t)$. Group frailties $\nu_{gi}$, are used to account for within cluster extra-Poisson variation and are assumed to be independent and identically distributed from a density $m_g$ with mean 1 and variance $\tau_g$. To allow for the flexible modeling of temporal trends and time-varying covariate effects the intensity function of each sub-process is assumed to have the form

$$\lambda_{gi}(t) = \exp\left\{\gamma_{g0}(t) + \sum_{q=1}^{p} x_{iq}\gamma_{gq}(t)\right\}$$

where the $\gamma$'s are modeled as penalized cubic B-splines, $\gamma_{gq}(t) = \sum_{r=1}^{k+4} \psi_{gqr} B_r(t)$, $\psi_{gqr}$ being the spline coefficients and $B_r(t)$'s, the B-spline basis functions. Knots are placed at all unique panel midpoints, $\left\{\frac{T_{i(j-1)}+T_{ij}}{2}\right\}$ and the penalties for each spline are chosen so that as the magnitude of the penalization increases the functions $\gamma_{gi}(t)$ tend to constants. A special case is then the proportional intensity model which occurs when $\gamma_{gq}(t) = \gamma_{gq}$, $q = 1, \ldots, p$. Under these modeling assumptions the expected number of events generated during $[T_{i(j-1)}, T_{ij})$ by the $g$th sub-group is given by $\mu_{gij} = \int_{T_{i(j-1)}}^{T_{ij}} \lambda_{gi}(t)\mathrm{d}t$. The unconditional expected number of counts in $[T_{i(j-1)}, T_{ij})$ is thus given by

$$\mu_{ij} = \sum_{g=1}^{G} p_g \mu_{gij} = \sum_{g=1}^{G} p_g \int_{T_{i(j-1)}}^{T_{ij}} \lambda_{gi}(t)\, \mathrm{d}t \tag{9}$$

where $p_g = E[z_{gi}]$, the probability of group membership. Estimation for the model is carried out using an adaptation of the expectation-solution (ES) algorithm (Rosen et al. 2000) where the sub-group process parameters are estimated by PQL given the individual-specific group membership probabilities (solution step); these probabilities are then updated with empirical Bayes posterior expectations (expectation step) given the other parameters. This procedure is repeated until convergence. This sort of clustering model is helpful in situations where hidden sub-groups may behave differently to treatments and it is of interest to model and isolate such trends. It is most useful where there are scientific bases for postulating the hidden sub-groups as discussed in the example considered in Nielsen and Dean(2007). The model may also be useful as an exploratory tool as illustrated through the analysis of the data considered in the previous section using a two-component clustered mixed NHPP.

Figure 2 presents the estimates of the overall control arm (high dose) intensity $\sum_{g=1}^{2} p_g \lambda_{g0}(t)$ with 95% point-wise coverage probabilities in the top left panel and the corresponding estimated group specific intensities $\lambda_{g0}(t)$ in the top right panel. The bottom two panels of Fig. 2 reflect similar quantities for the treatment arm (low dose) $\lambda_1(t) = \sum_{g=1}^{2} p_g \lambda_{g1}(t)$ with $\lambda_{g1}(t) = \exp\left\{\gamma_{g0}(t) + \gamma_{g1}(t)\right\}$. Estimates of the overall intensities agree with the trends observed from the analysis in Sect. 3.3. In addition, the group-specific intensities convey interesting information as they indicate that there are two underlying sub-populations across both arms; those that decrease their dosage over time, around 0.381 (s.e. 0.067) of the population, and a group, about 0.619 (s.e. 0.067) of the patients that seems to medicate regularly. It should be noted that these results are simply exploratory and interaction with the scientific investigators is required before any conclusions can be drawn. However, these findings merit further scientific exploration.

## 4 Diagnostics and testing

Though many complicated hierarchical models have been developed, methods for model assessment are sparse. Ideally diagnostics would consider each layer of the hierarchy and each assumption at the specific layers. This would help target model refinement strategies to those assumptions which seem least appropriate. In many cases where diagnostics have been developed the model is assessed based on the

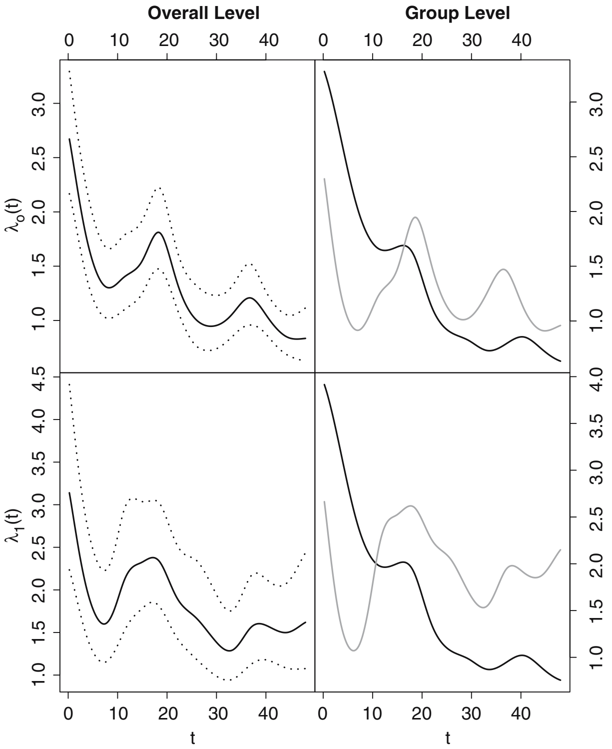

**Fig. 2** (Top row) Plot of the estimated overall intensity for the control (high dose) arm $\lambda_0(t)$ along with 95% point-wise confidence intervals in the left panel and plots of the corresponding estimated group intensities $\lambda_{g0}(t)$ on the right (g = 1, black line; g = 2, gray line). (Bottom row) Plot of the estimated overall intensity for the treatment (low dose) arm $\lambda_1(t)$ along with 95% point-wise confidence intervals in the left panel and plots of the corresponding estimated group intensities $\lambda_{g1}(t)$

whole ensemble of assumptions and no sub-models nested in such complex hierarchy are verified. This black-box approach to goodness-of-fit can be very unsatisfactory, and certainly this is an important area for future research. Here we highlight a few techniques including relatively recent and novel procedures.

For GLMMs, Pearson, Anscombe and deviance residuals (McCullagh and Nelder 1989) may be created for the marginal means $\boldsymbol{\mu}$ or the conditional means $\boldsymbol{\mu}^{\mathbf{b}}$. Both require that the assumptions on the random effects, $\mathbf{b}$, are valid. A visual check on such assumptions could be performed using a normal q–q plot based on $\hat{\mathbf{L}}^{-1}\hat{\mathbf{b}}$. Here $\mathbf{R} = Var[\mathbf{b}] = \mathbf{L}\mathbf{L}^T$, so that $\mathbf{b} = \mathbf{L}\mathbf{z}$ where $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. Tchetgen and Coull (2006) derive a test for a misspecification of the random effects distribution of a GLMM. The test statistic is based on the difference between the marginal and conditional maximum likelihood estimates of the fixed effects parameters $\boldsymbol{\alpha}$, with the idea being that if the random effects are properly specified, then both conditional and marginal methods are consistent so $\boldsymbol{\alpha}_m = \boldsymbol{\alpha}_c$ under such a scenario.

Resampling methods may also be used for goodness-of-fit with respect to certain gross quantities. For example, with the zero-heavy model resampling methods may compare observed and simulated estimates of the expected number of zeros.

Two typical approaches to formal testing of goodness-of-fit rely on (i) nesting the model under consideration in an alternative more complex model and (ii) the development of omnibus tests via the probability integral transformation. For example, the generalized inverse Gaussian family in Sect. 3 may be used to test the appropriateness of the negative binomial model. Chen et al. (2002) relax the assumption of multivariate normal random effects and assume that $\mathbf{b} = \mathbf{Lz}$ where $\mathbf{z}$ has seminonparametric density $h_K(\mathbf{z}) = P_k(\mathbf{z})^2\psi(\mathbf{z})$, $\psi(\mathbf{z})$ being the density of a $d$ dimensional standard normal and $P_k(\mathbf{z}) = \sum_{0\le\gamma_1+\cdots+\gamma_d\le k} a_{\boldsymbol{\gamma}}\, z_1^{\gamma_1} z_2^{\gamma_2} \ldots z_d^{\gamma_d}$ with $a_{00\ldots0} = 1$. A test that all $a_{\boldsymbol{\gamma}}$'s are 0 (excluding $a_{00\ldots0}$ defined as 1) would assess the appropriateness of the assumption of multivariate normal random effects. A recent example of approach (ii) above is Waagepetersen (2006) who uses the empirical process to test the normality assumption of the random effects via the Andersen-Darling statistic; $p$-values are computed by simulation.

Note that there is scope for much further work in the arena of diagnostics for GLMMs. What is required in the toolkit for the analysis of count data are mechanisms for graphical exploratory analyses and model checks which fine-tune specific model components. For example, in spatial analyses of zero-heavy data, zeros may arise as spatially correlated indicating regions where there is less susceptibility for the event, or isolated zeros in regions with larger counts, indicating small clusters of individual-specific resistance. Diagnostics which help in discriminating which of these predominate would assist in the building of covariance structures of the zero-mixing component. Where there are many random effects in a model, it is often difficult to isolate goodness-of-fit techniques which consider each separately.

## 5 Some remarks on software

Generalized linear mixed and additive models are now fundamental tools in the analysis of longitudinal data. Software for their analysis is readily available: the GLIMMIX macro in SAS (SAS Institute 2007), SAS macro *spmm* (available at www.hsph.harvard.edu/xlin/software.html) and several packages in R (R Development Core Team 2006), for example, *glmmPQL()* in the MASS package (Venables and Ripley 2002) or *lmer()* in the lme4 package (Bates and Sarkar 2007) and *mgcv()* for generalized additive models in the mgcv package (Wood 2006).

**Acknowledgements** Research was supported by the Natural Sciences and Engineering Research Council of Canada, the National Program on Complex Data Structures and Geomatics for Informed Decisions, a Networks of Centres of Excellence program.

## References

Abramowitz M, Stegun IA (eds) (1984) Handbook of mathematical functions with formulas, graphs, and mathematical tables. A Wiley-Interscience Publication, John Wiley, New York

Ainsworth L, Dean CB (2007) Detection of local and global outliers in mapping studies. Environmetrics. doi:10.1002/env.851

Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications (with discussion). J R Stat Soc Ser B: Methodol 41:279–299

Bates D, Sarkar D (2007) lme4: Linear mixed-effects models using S4 classes. R package version 0.99875–1

Breslow N (1989) Score tests in overdispersed GLM's. In: Decarli A, Francis BJ, Gilchrist R, Seeber GUH, (eds) Statistical modelling. Springer-Verlag Inc., pp 64–74

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25

Breslow NE, Lin X (1995) Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika 82:81–91

Casella G, George EI (1992) Explaining the Gibbs sampler. Am Stat 46:167–174

Chen J, Zhang D, Davidian M (2002) A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. Biostatistics (Oxford) 3:347–360

Cheng KF, Wu JW (1994) Testing goodness of fit for a parametric family of link functions. J Am Stat Assoc 89:657–664

Cook RD, Weisberg S (1989) Regression diagnostics with dynamic graphics (C/R: P293–311). Technometrics 31:277–291

Davidian M, Carroll RJ (1987) Variance function estimation. J Am Stat Assoc 82:1079–1091

Dean CB, Balshaw R (1997) Efficiency lost by analyzing counts rather than event times in Poisson and overdispersed Poisson regression models. J Am Stat Assoc 92:1387–1398

Dean C, Lawless JF, Willmot GE (1989) A mixed Poisson-inverse–Gaussian regression model. Can J Stat 17:171–181

Dubin JA, Han L, Fried TR (2007) Triggered sampling could help improve longitudinal studies of persons with elevated mortality risk. J Clin Epidemiol 60:288–93

Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman & Hall Ltd.

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–338

Hastie T, Tibshirani R (1999) Generalized additive models. Chapman & Hall Ltd.

Henderson R, Shimakura S (2003) A serially correlated gamma frailty model for longitudinal count data. Biometrika 90:355–366

Insightful Corporation (2007) S-PLUS 8. Insightful Corporation, Seattle, WA. URL http://www.insightful.com, accessed on 25 October 2007

Jiang J (1998) Consistent estimators in generalized linear mixed models. J Am Stat Assoc 93:720–729

Jørgensen B (1982) Statistical properties of the generalized inverse Gaussian distribution. Lecture Notes in Statistics, vol 9. Springer-Verlag, New York

Kleinman K, Lazarus R, Platt R (2004) A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. Am J Epidemiol 159:217–24

Laird NM (1991) Topics in likelihood-based methods for longitudinal data analysis. Statistica Sinica 1:33–50

Laird NM, Louis TA (1982) Approximate posterior distributions for incomplete data problems. J R Stat Soc Ser B: Methodol 44:190–200

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34:1–14

Lawless JF (1987) Negative binomial and mixed Poisson regression. Can J Stat 15:209–225

Lawless JF, Zhan M (1998) Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. Can J Stat 26:549–565

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Lin X, Breslow NE (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. J Am Stat Assoc 91:1007–1016

Lin X, Zhang D (1999) Inference in generalized additive mixed models by using smoothing splines. J R Stat Soc Ser B: Stat Methodol 61:381–400

Lin X, Harlow SD, Raz J, Harlow SD (1997) Linear mixed models with heterogeneous within-cluster variances. Biometrics 53:910–923

Lindstrom MJ, Bates DM (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. J Am Stat Assoc 83:1014–1022, Corr: 94V89, p 1572

Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett 8:1235–1246

McCullagh P, Nelder JA (1989) Generalized linear models. Chapman & Hall Ltd.

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. J Am Stat Assoc 92:162–170

Nelder JA, Pregibon D (1987) An extended quasi-likelihood function. Biometrika 74:221–232

Nielsen JD, Dean CB (2007) Clustered mixed nonhomogeneous Poisson process spline models for the analysis of recurrent event panel data. Biometrics. doi:10.1111/j.1541-0420.2007.00940.x

Nodtvedt A, Dohoo I, Sanchez J, Conboy G, DesCjteaux L, Keefe G, Leslie K, Campbell J (2002) The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows. Can J Vet Res 66:249–257

Pierce DA, Schafer DW (1986) Residuals in generalized linear models. J Am Stat Assoc 81:977–986

Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS. Springer-Verlag Inc.

R Development Core Team (2006) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org, accessed on 25 October 2007

Raudenbush SW, Yang M-L, Yosef M (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. J Comput Graph Stat 9:141–157

Rich-Edwards JW, Kleinman KP, Strong EF, Oken E, Gillman MW (2005) Preterm delivery in Boston before and after September 11th, 2001. Epidemiology 16:323–327

Rosen O, Jiang W, Tanner MA (2000) Mixtures of marginal models. Biometrika 87:391–404

Sartori N, Severini TA (2004) Conditional likelihood inference in generalized linear mixed models. Statistica Sinica 14:349–360

SAS Institute (2007) PROC GLIMMIX. SAS Institute Inc. URL http://www.sas.com, accessed on 25 October 2007

Shun Z, McCullagh P (1995) Laplace approximation of high dimensional integrals. J R Stat Soc Ser B: Methodol 57:749–760

Sichel HS (1974) On a distribution representing sentence-length in written prose. J R Stat Soc Ser A 137:25–34

Simons JS, Neal DJ, Gaher RM (2006) Risk for marijuana-related problems among college students: an application of zero-inflated negative binomial regression. Am J Drug Alcohol Abuse 32:41–53

Song PX-K, Fan Y, Kalbfleisch JD (2005) Maximization by parts in likelihood inference. J Am Stat Assoc 100:1145–1158

Stangle DE, Smith DR, Beaudin SA, Strawderman MS, Levitsky DA, Strupp BJ (2007) Succimer chelation improves learning, attention, and arousal regulation in lead-exposed rats but produces lasting cognitive impairment in the absence of lead exposure. Environ Health Perspect 115:201–209

Tchetgen EJ, Coull BA (2006) A diagnostic test for the mixing distribution in a generalised linear mixed model. Biometrika 93:1003–1010

Tierney L, Kass RE, Kadane JB (1989) Approximate marginal densities of nonlinear functions. Biometrika 76:425–433, Corr: V78, p233–234

Tjur T (1982) A connection between Rasch's item analysis model and a multiplicative Poisson model. Scand J Stat 9:23–30

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. URL http://www.stats.ox.ac.uk/pub/MASS4, accessed on 25 October 2007

Vonesh EF, Wang H, Nie L, Majumdar D (2002) Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. J Am Stat Assoc 97:271–283

Waagepetersen R (2006) A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. Scand J Stat 33:721–731

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61:439–447

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50:1–26

Wood S (2006) mgcv: GAMs with GCV smoothness estimation and GAMMs by REML/PQL. R package version 1.3–24

Zeger SL, Karim MR (1991) Generalized linear models with random effects: a Gibbs sampling approach. J Am Stat Assoc 86:79–86