Approximate Inference in Generalized Linear Mixed Models
Author(s): N. E. Breslow and D. G. Clayton
Reviewed work(s):
Source: *Journal of the American Statistical Association*, Vol. 88, No. 421 (Mar., 1993), pp. 9-25
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2290687
Accessed: 27/11/2012 12:04

Approximate Inference in Generalized Linear Mixed Models

N. E. BRESLOW and D. G. CLAYTON*

Statistical approaches to overdispersion, correlated errors, shrinkage estimation, and smoothing of regression relationships may be encompassed within the framework of the generalized linear mixed model (GLMM). Given an unobserved vector of random effects, observations are assumed to be conditionally independent with means that depend on the linear predictor through a specified link function and conditional variances that are specified by a variance function, known prior weights and a scale factor. The random effects are assumed to be normally distributed with mean zero and dispersion matrix depending on unknown variance components. For problems involving time series, spatial aggregation and smoothing, the dispersion may be specified in terms of a rank deficient inverse covariance matrix. Approximation of the marginal quasi-likelihood using Laplace's method leads eventually to estimating equations based on penalized quasilikelihood or PQL for the mean parameters and pseudo-likelihood for the variances. Implementation involves repeated calls to normal theory procedures for REML estimation in variance components problems. By means of informal mathematical arguments, simulations and a series of worked examples, we conclude that PQL is of practical value for approximate inference on parameters and realizations of random effects in the hierarchical model. The applications cover overdispersion in binomial proportions of seed germination; longitudinal analysis of attack rates in epilepsy patients; smoothing of birth cohort effects in an age-cohort model of breast cancer incidence; evaluation of curvature of birth cohort effects in a case-control study of childhood cancer and obstetric radiation; spatial aggregation of lip cancer rates in Scottish counties; and the success of salamander matings in a complicated experiment involving crossing of male and female effects. PQL tends to underestimate somewhat the variance components and (in absolute value) fixed effects when applied to clustered binary data, but the situation improves rapidly for binomial observations having denominators greater than one.

KEY WORDS: Longitudinal data; Overdispersion; Penalized quasi-likelihood; Spatial aggregation; Variance components.

The generalized linear model (GLM) (McCullagh and Nelder 1989) neatly synthesizes likelihood-based approaches to regression analysis for a variety of outcome measures. Several recent extensions of this useful theory involve models with random terms in the linear predictor. Such generalized linear mixed models (GLMM's) are useful for accommodating the overdispersion often observed among outcomes that nominally have binomial (Williams 1982) or Poisson (Breslow 1984) distributions; for modeling the dependence among outcome variables inherent in longitudinal or repeated measures designs (Stiratelli, Laird, and Ware 1984; Zeger, Liang, and Albert 1988); and for producing shrinkage estimates in multiparameter problems, such as the construction of maps of small area disease rates (Clayton and Kaldor 1987; Manton et al. 1989).

It is often a reasonable approximation and certainly traditional to assume that the random error terms have a multivariate normal distribution whose variance components are to be estimated from the data. When the outcomes come in the form of proportions or counts, a full maximum likelihood analysis based on their joint marginal distribution requires numerical integration techniques for calculation of the log-likelihood, score equations, and information matrix. This method has been implemented successfully in relatively simple problems involving binomial (Brillinger and Preisler 1986; Crouch and Spiegelman 1990) and Poisson (Hinde 1982) mixtures with a high degree of independence among the observations. To date it has proved intractable for more complicated problems involving irreducibly high-dimensional integrals, however. Recent Bayesian procedures avoid the need for numerical integration by taking repeated samples from the posterior distributions using importance (Ii and Raghunathan 1991) or Gibbs (Besag, York, and Mollié 1991; Zeger and Karim 1991) sampling techniques. An attractive feature of the Bayesian approach is its flexibility for full assessment of the uncertainty in the estimated random effects and functions of model parameters. Potential drawbacks include the intensive computations and questions about when the sampling process has achieved equilibrium (Ripley and Kirkland 1990). There is still room for simple, approximate methods both for exploratory analyses and to provide starting values for use with other, more exact procedures.

This article considers two closely related approximate methods of inference in GLMM's and investigates their suitability for practical work by means of Monte Carlo studies and illustrative applications. Both have been considered previously, although not at the level of generality adopted here. The penalized quasi-likelihood (PQL) method exploited by Green (1987) for semiparametric regression analysis is available for inference in hierarchical models where the focus is on shrinkage estimation of the random effects (Robinson 1991). PQL was proposed as an approximate Bayes procedure for some commonly occurring GLMM's by Laird (1978) and by Stiratelli et al. (1984) and it has been used more recently by Schall (1991) and McGilchrist and Aisbett (1991). Marginal quasi-likelihood (MQL) is the name that we give to the procedure proposed by Goldstein (1991) as an extension to GLM's of his work on multilevel modeling

* N. E. Breslow is Professor and Chairman, Department of Biostatistics, University of Washington, Seattle, WA 98195. D. G. Clayton is Statistician, MRC Biostatistics Unit, Cambridge CB2 2SR, United Kingdom. This work was supported in part by United States Public Health Service Grant CA40644 and Senior International Fellowship TW01617. The authors thank the editors and referees for helpful comments that substantially improved the exposition.

© 1993 American Statistical Association
Journal of the American Statistical Association
March 1993, Vol. 88, No. 421, Applications and Case Studies

(Goldstein 1986, 1988). It is appropriate when interest is focused more on the marginal relationship between covariables and outcome (Liang, Zeger, and Qaqish 1992; Zeger et al. 1988). A key feature of these methods is that they may be implemented by repeated use of standard software for variance components analysis of normally distributed observations, just as GLM's may be fitted by repeated calls to weighted least squares procedures. Both provide for shrinkage estimates of random error terms and for analogs of restricted maximum likelihood (REML) estimates of the variance components (Patterson and Thompson 1971).

The organization of the article is as follows. After a brief statement of the model in Section 1, the PQL criterion is motivated in Section 2 by approximating the integrated quasi-likelihood. In Section 3 an approximate GLM for the marginal distribution of the data is developed and related to the generalized estimating equation approach of Zeger et al. (1988). An outline of the computational procedure is presented in Section 4. A Monte Carlo investigation using binomial-normal observations is discussed in Section 5. Illustrative analyses presented in Section 6 indicate a wide range of applications. In Section 7, PQL and MQL are related to other recent work and some suggestions for further research are made. Readers who are less interested in theoretical and computational aspects may wish to concentrate their attention on Sections 1, 2.3, 3.1, 5, and 6.

1. THE HIERARCHICAL MODEL

Observations on the ith of n units consist of a univariate response variable y_i together with vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ of explanatory variables associated with the fixed and random effects. Units may be blocked in some way, for example when they involve repeated measures on the same subject. We suppose that, given a q-dimensional vector $\mathbf{b}$ of random effects, the y_i are conditionally independent with means $E(y_i \mid \mathbf{b}) = \mu_i^b$ and variances $\operatorname{var}(y_i \mid \mathbf{b}) = \phi a_i v(\mu_i^b)$, where $v(\cdot)$ is a specified variance function, a_i is a known constant (e.g., the reciprocal of a binomial denominator) and ϕ is a dispersion parameter that may or may not be known. This formulation encompasses situations where the random effects are nested within subjects (e.g., Sections 6.1, 6.5) and when they are not (Sections 6.2–6.4, 6.6). The conditional mean is related to the linear predictor $\eta_i^b = \mathbf{x}_i^t \boldsymbol{\alpha} + \mathbf{z}_i^t \mathbf{b}$ by the link function $g(\mu_i^b) = \eta_i^b$, with inverse $h = g^{-1}$, where $\boldsymbol{\alpha}$ is a p vector of fixed effects. Denoting the observation vector by $\mathbf{y} = (y_1, \ldots, y_n)^t$ and the design matrices with rows $\mathbf{x}_i^t$ and $\mathbf{z}_i^t$ by $\mathbf{X}$ and $\mathbf{Z}$, the conditional mean satisfies

$$E(\mathbf{y} \mid \mathbf{b}) = h(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b}). \tag{1}$$

The model is completed by the assumption that $\mathbf{b}$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta})$ depending on an unknown vector $\boldsymbol{\theta}$ of variance components. In all the examples we consider, which involve binomial, Poisson, and hypergeometric specifications for the conditional distribution of y_i, the dispersion parameter ϕ is fixed at unity. In other applications, however, it may be estimated together with $\boldsymbol{\theta}$ as a parameter in the covariance matrix of the marginal distribution of $\mathbf{y}$.

The integrated quasi-likelihood function used to estimate $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ is defined by

$$e^{ql(\boldsymbol{\alpha}, \boldsymbol{\theta})} \propto |\mathbf{D}|^{-1/2} \int \exp\left[-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i; \mu_i^b) - \frac{1}{2} \mathbf{b}^t \mathbf{D}^{-1} \mathbf{b} \right] d\mathbf{b},$$

where (2)

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

denotes the deviance measure of fit. If, conditionally on $\mathbf{b}$, the observations are drawn from a linear exponential family with variance function $v(\cdot)$, then the deviance is well known to equal to the scaled difference $2\phi\{l(y; y, \phi) - l(y; \mu, \phi)\}$, where $l(y; \mu, \phi)$ denotes the conditional likelihood of y given its mean μ (see, for example, McCullagh and Nelder 1989). In this case $ql(\boldsymbol{\alpha}, \boldsymbol{\theta})$ represents the true log-likelihood of the data. As already mentioned, the primary difficulty in implementing full likelihood inference lies in the integrations needed to evaluate ql and its partial derivatives.

2. PENALIZED QUASI-LIKELIHOOD

2.1 Motivation of the PQL Criterion

Writing Equation (2) in the form $c|\mathbf{D}|^{-1/2} \int e^{-\kappa(\mathbf{b})} \, d\mathbf{b}$, we apply Laplace's method for integral approximation (Barndorff-Nielsen and Cox 1989, sec. 3.3; Tierney and Kadane 1986). Let κ' and κ'' denote the q vector and $q \times q$ dimensional matrix of first- and second-order partial derivatives of κ with respect to $\mathbf{b}$. Ignoring the multiplicative constant c, the approximation yields

$$ql(\boldsymbol{\alpha}, \boldsymbol{\theta}) \approx -\frac{1}{2} \log|\mathbf{D}| - \frac{1}{2} \log|\kappa''(\tilde{\mathbf{b}})| - \kappa(\tilde{\mathbf{b}}), \tag{3}$$

where $\tilde{\mathbf{b}} = \tilde{\mathbf{b}}(\boldsymbol{\alpha}, \boldsymbol{\theta})$ denotes the solution to

$$\kappa'(\mathbf{b}) = -\sum_{i=1}^{n} \frac{(y_i - \mu_i^b)\mathbf{z}_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} + \mathbf{D}^{-1}\mathbf{b} = 0$$

that minimizes $\kappa(\mathbf{b})$. Differentiating again with respect to $\mathbf{b}$, we have

$$\kappa''(\mathbf{b}) = \sum_{i=1}^{n} \frac{\mathbf{z}_i \mathbf{z}_i^t}{\phi a_i v(\mu_i^b)[g'(\mu_i^b)]^2} + \mathbf{D}^{-1} + \mathbf{R}$$

$$\approx \mathbf{Z}^t \mathbf{W} \mathbf{Z} + \mathbf{D}^{-1}, \tag{4}$$

where $\mathbf{W}$ is the $n \times n$ diagonal matrix with diagonal terms $w_i = \{\phi a_i v(\mu_i^b)[g'(\mu_i^b)]^2\}^{-1}$ that are recognizable as the GLM iterated weights (Firth 1991, p. 63; McCullagh and Nelder 1989, sec. 2.5). The remainder term

$$\mathbf{R} = -\sum_{i=1}^{n} (y_i - \mu_i^b)\mathbf{z}_i \frac{\partial}{\partial \mathbf{b}} \left[\frac{1}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} \right]$$

has expectation 0 and is thus, in probability as a function of n, of lower order than the two leading terms in Equation (4). $\mathbf{R}$ equals $\mathbf{0}$ for the canonical link functions, for which

$g'(\mu) = v^{-1}(\mu)$ (McCullagh and Nelder 1989, p. 32). Combining (2)–(4) and ignoring $\mathbf{R}$ leads to

$$ql(\alpha, \theta) \approx -\frac{1}{2}\log|\mathbf{I} + \mathbf{Z}'\mathbf{WZD}| - \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i, \mu_i^{\tilde{b}}) - \frac{1}{2}\tilde{\mathbf{b}}'\mathbf{D}^{-1}\tilde{\mathbf{b}}, \quad (5)$$

where $\tilde{\mathbf{b}}$ is chosen to maximize the sum of the last two terms.

Assuming that the GLM iterative weights vary slowly (or not at all) as a function of the mean, we ignore the first term in this expression and choose α to maximize the second. Thus $(\hat{\alpha}, \hat{\mathbf{b}}) = (\hat{\alpha}(\theta), \hat{\mathbf{b}}(\theta))$, where $\hat{\mathbf{b}}(\theta) = \tilde{\mathbf{b}}(\hat{\alpha}(\theta))$, jointly maximize Green's (1987) PQL

$$-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i, \mu_i^b) - \frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}. \quad (6)$$

Differentiation with respect to α and $\mathbf{b}$ leads to the score equations for the mean parameters:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i^b)\mathbf{x}_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} = 0 \quad (7)$$

and

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i^b)\mathbf{z}_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} = \mathbf{D}^{-1}\mathbf{b}. \quad (8)$$

Stiratelli et al. (1984) derived these equations for logistic regression of binary data by maximizing the Bayes posterior distribution for $(\alpha, \mathbf{b})$ under a diffuse prior for α (cf. Schall 1991).

2.2 Fisher Scoring

Green (1987) developed the Fisher scoring algorithm for solution of Equations (7) and (8) as an iterated weighted least squares (IWLS) problem involving a working dependent variable and a weight matrix that are updated at each iteration. His development is modified slightly here to exploit the close correspondence with the normal theory calculations of Harville (1977). Defining the working vector $\mathbf{Y}$ to have components $Y_i = \eta_i^b + (y_i - \mu_i^b)g'(\mu_i^b)$, the solution to (7) and (8) via Fisher scoring may be expressed as the iterative solution to the system

$$\begin{bmatrix} \mathbf{X}'\mathbf{WX} & \mathbf{X}'\mathbf{WZD} \\ \mathbf{Z}'\mathbf{WX} & \mathbf{I} + \mathbf{Z}'\mathbf{WZD} \end{bmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\nu} \end{pmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{WY} \\ \mathbf{Z}'\mathbf{WY} \end{bmatrix}, \quad (9)$$

where $\mathbf{b} = \mathbf{D}\boldsymbol{\nu}$. Harville (1977) derived (9) for the best linear unbiased estimation (BLUE) of α and $\mathbf{b}$ in the associated normal theory model $\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Zb} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, ε and $\mathbf{b}$ independent. Equivalently, one may first solve for α in

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\alpha = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \quad (10)$$

where $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{ZDZ}'$, and then set

$$\hat{\mathbf{b}} = \mathbf{D}\hat{\boldsymbol{\nu}} = \mathbf{DZ}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\alpha}). \quad (11)$$

This suggests that one take as an approximate covariance for $\hat{\alpha}$ the matrix $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. This is exact for the normal theory linear model, provided that θ is known. Standard errors for $\hat{\mathbf{b}}$ may be calculated from (11). But both sets of standard errors ignore the additional variability stemming from the need to estimate θ. Normal theory approximations that account for this additional variability have been proposed (Kackar and Harville 1984), and it would be of interest to explore their suitability for the more general models considered here.

2.3 Singular Variance Matrices

As seen in the examples in Sections 6.3–6.5, which involve autoregressive models for smoothing and spatial aggregation, it is sometimes useful to specify the random effects model in terms of the *inverse* dispersion matrix $\mathbf{R}$ rather than $\mathbf{D}$. Reciprocals of the diagonal terms in $\mathbf{R}$ equal conditional variances given the remaining variables, whereas the off-diagonal terms determine the conditional regression and correlation relationships (Dempster 1972; Whittaker 1990). In stochastic smoothing models of this type, the term $\mathbf{b}'\mathbf{Rb}$ in Equation (6) represents a "roughness penalty" (Good and Gaskins 1971). Typically, however, no penalty is associated with the overall average or level of the components of $\mathbf{b}$, nor sometimes even with their linear trend. Then $\mathbf{R}$ is singular and the probability distribution of the random effects is not fully specified.

Suppose that r linearly independent combinations of $\mathbf{b}$, the "aliased" components, have no associated probability distribution. Typically one constrains these components to 0, say by $\mathbf{Gb} = \mathbf{0}$ where $\mathbf{G}$ is a $r \times q$ matrix of rank r, and includes them instead in the fixed part of a model. There are then two ways to proceed. Either one can reduce the problem to that of estimating a reduced set of $q - r$ random effects that have a specified, full rank probability distribution (e.g., McGilchrist and Aisbett 1991). Or equivalently, one can use for the dispersion matrix $\mathbf{D}$ the Moore–Penrose generalized inverse, $\mathbf{R}^-$ (Graybill 1983, sec. 6.2.1). This sets to 0 the variances for the aliased linear combinations of $\mathbf{b}$, so that they are effectively constrained to take on the value 0. The latter approach offers the greater flexibility and is the one chosen here.

2.4 Variance Component Estimation

Substitution of the maximized value of (6) into (5) and evaluation of $\mathbf{W}$ at $(\hat{\alpha}(\theta), \hat{\mathbf{b}}(\theta))$ generates an approximate profile quasi-likelihood function for inference on θ. We make some further approximations to motivate standard estimating equations in terms of the working vector $\mathbf{Y}$, the iterated weights $\mathbf{W}$, and the design matrices $\mathbf{X}$ and $\mathbf{Z}$. Ignoring throughout the dependence of $\mathbf{W}$ on θ and replacing the deviance $\sum d_i(y_i, \mu_i^b)$ by the Pearson chi-squared statistic $\sum (y_i - \mu_i^b)^2/[a_i v(\mu_i^b)]$, we have up to the usual additive constant

$$ql(\hat{\alpha}(\theta), \theta) \approx -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\hat{\alpha})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\alpha}). \quad (12)$$

This quantity, whose derivation used Harville's (1977) Equations 5.1 and 5.2, may be recognized as the profile likelihood based on the associated normal theory model for $\mathbf{Y}$. To make degrees-of-freedom adjustments that account for the fact that $\hat{\alpha}$ rather than α appears in the quadratic form in (12), we use in practice the REML version (Patterson and Thompson 1971):

$$ql_1(\hat{\alpha}(\theta), \theta) \approx -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\hat{\alpha})^t\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\alpha}). \quad (13)$$

For the normal theory linear model this adjustment corresponds to the profile likelihood correction of Cox and Reid (1987). Full justification requires (a) that α and θ be orthogonal parameters and (b) that the information matrix for $\hat{\alpha}(\theta)$ be $\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}$. Neither requirement holds exactly for the general GLMM developed in this section. Both do hold, however, for the marginal model considered in the next section.

Following Harville (1977), we define $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X} \times (\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}$ and differentiate (13) with respect to the components of θ to obtain estimating equations for the variance parameters:

$$-\frac{1}{2}\left[(\mathbf{Y} - \mathbf{X}\alpha)^t\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\alpha) - \operatorname{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)\right] = 0. \quad (14)$$

The corresponding Fisher information matrix $\mathcal{J}$ has components

$$\mathcal{J}_{jk} = -\frac{1}{2}\operatorname{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_k}\right). \quad (15)$$

Note that because the dependence of $\mathbf{W}$ on θ is ignored in calculating $\partial\mathbf{V}/\partial\theta_j$, in (14) and (15) ql_1 cannot be used as an objective function to help solve the equations.

2.5 Remarks on Asymptotic Theory

Our "derivation" of the penalized quasi-likelihood (6) and modified profile quasi-likelihood (13) involved several ad hoc adjustments and approximations for which no formal justification was given. It is best viewed as providing heuristic motivation for the estimating Equations (7), (8), and (14) that may be studied in their own right. Informal considerations, however, do suggest circumstances in which the approximations should perform well. First, the equations are the REML equations under the normal theory linear model, for which the working and observation vectors coincide and $\mathbf{W}$ is the identity matrix. Unless $\mathbf{y}$ can be partitioned into K independent components where K increases with n, of course, even this does not guarantee that the standard asymptotic theory applies (Harville 1977, sec. 4.2). Second, key portions of the argument involved approximating the deviance increments by the normed, squared residuals or the penalized deviance by a quadratic function of $\mathbf{b}$. Both approximations are likely to improve as the individual y_i become more normally distributed. Such "small dispersion asymptotics" occur, for example, as the denominators of binomial proportions or the means of Poisson observations increase (Jorgensen 1987).

3. MARGINAL QUASI-LIKELIHOOD

3.1 The Marginal Model

A key feature of the hierarchical model is that the regression structure in Equation (1) is conditional on the values of the random effects $\mathbf{b}$. When separate random effects are estimated for each person in a medical study, for example, this means that α represents covariable effects at the level of the individual subject. Because one generally desires estimates of covariable effects on population averages (e.g., on the survival rates of specific subgroups), it is often more appropriate to specify the GLM in terms of the *marginal* mean as

$$E(y_i) = \mu_i = h(\mathbf{x}_i^t\alpha). \quad (16)$$

Unless the link function is the identity, however, the marginal mean so defined does not generally coincide with the marginal mean calculated from Equation (1).

One may, nonetheless, think of (16) as derived from a rather crude, first-order approximation to the hierarchical model that is valid in the limit as the components of dispersion approach 0. Writing the model in the form $y_i = \mu_i^b + \varepsilon_i$ with $\operatorname{var}(\varepsilon_i) = \phi a_i v(\mu_i^b)$ and $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, one has $y_i \approx h(\mathbf{x}_i^t\alpha) + h'(\mathbf{x}_i^t\alpha)\mathbf{z}_i^t\mathbf{b} + \varepsilon_i$ (Goldstein 1991). Defining $\mathbf{V}_0$ and Δ to be the diagonal matrices with diagonal elements $\phi a_i v(\mu_i)$ and $g'(\mu_i)$, the corresponding first-order variance approximation is

$$\operatorname{var}(\mathbf{y}) = \mathbf{V}_0 + \Delta^{-1}\mathbf{Z}\mathbf{D}\mathbf{Z}^t\Delta^{-1}. \quad (17)$$

Marginal models of this form were investigated for longitudinal designs by Zeger et al. (1988). They showed that the true marginal mean for the hierarchical model with normally distributed random effects often could be expressed in the form of (16), at least approximately, but with *altered values* for the regression variables or regression coefficients. With the log link, for example, one finds $E(y_i) = \exp(\mathbf{x}_i^t\alpha + \mathbf{z}_i^t\mathbf{D}\mathbf{z}_i/2)$, and thus that the random effects add an *offset* to the equation for the marginal mean. (Nonnormal random effects lead to other offsets.) With the logit link, the α coefficients are *attenuated:*

$$E(y_i) \approx \frac{\exp(c_i\mathbf{x}_i^t\alpha)}{1 + \exp(c_i\mathbf{x}_i^t\alpha)}, \quad (18)$$

where $c_i = |c^2\mathbf{D}\mathbf{z}_i\mathbf{z}_i^t + \mathbf{I}|^{-1/2} = (1 + c^2\mathbf{z}_i^t\mathbf{D}\mathbf{z}_i)^{-1/2}$ and $c = 16\sqrt{3}/(15\pi)$. They considered the approximation (17) sufficiently accurate for use as a "working covariance" matrix in their iterative estimation procedure but based their inferences on an empirical covariance matrix derived from the estimating equations.

3.2 Estimation of Fixed and Random Effects

For fixed θ, we estimate the regression coefficients α in the marginal model using the quasi-likelihood equations appropriate for dependent outcomes (McCullagh and Nelder

1989, sec. 9.3). Denoting the marginal mean vector by $\mu = (\mu_1, \ldots, \mu_n)^t$, the estimating equations

$$\mathbf{U}(\alpha, \theta) = \frac{\partial\mu}{\partial\alpha^t} \operatorname{var}^{-1}(\mathbf{y})(\mathbf{y} - \mu) = 0$$

take the form

$$\mathbf{X}^t(\Delta\mathbf{V}_0\Delta + \mathbf{ZDZ}^t)^{-1}\Delta(\mathbf{y} - \mu) = 0. \qquad (19)$$

With $\eta = (\eta_1, \ldots, \eta_n)^t$ denoting the vector of linear predictors $\eta_i = \mathbf{x}_i^t\alpha$, Fisher scoring leads to IWLS regression of the working vector $\mathbf{Y} = \eta + \Delta(\mathbf{y} - \mu)$ on $\mathbf{X}$ with weight matrix

$$\mathbf{V}^{-1} = (\Delta\mathbf{V}_0\Delta + \mathbf{ZDZ}^t)^{-1} = (\mathbf{W}^{-1} + \mathbf{ZDZ}^t)^{-1}, \qquad (20)$$

where $\mathbf{W}$ is once again the diagonal matrix with the GLM iterated weights

$$w_i = \{\phi a_i v(\mu_i)[g'(\mu_i)]^2\}^{-1}$$

as diagonal elements. At each step in the iteration, the problem is formally equivalent to that of estimating α in the associated normal theory model $\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Zb} + \varepsilon$, with $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$. Shrinkage estimates for the random effects are again obtained as $\hat{\mathbf{b}} = \mathbf{DZ}^t\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\alpha})$.

The essential difference between the MQL estimating equations for the marginal model and the PQL equations for the hierarchical model is that the latter incorporate the random effect terms $\mathbf{z}_i^t\mathbf{b}$ in the linear predictor.

3.3 Estimation of Variance Components

Variance parameters may be estimated using Carroll and Ruppert's (1982) method of pseudolikelihood. Assuming that $E(\mathbf{y})$ is known, we consider the normal theory likelihood based on the variance approximation (17) and take logarithmic derivatives with respect to θ. This leads to precisely the same REML equations for the variance parameters as were derived previously for PQL. Instead of iterating back and forth between (7)–(8) and (14), obtaining new values of $\mathbf{b}$ at each iteration for use in $\mathbf{Y}$ and $\mathbf{V}$, we iterate between (19) for α and (14) for θ, delaying the estimation of $\mathbf{b}$ from (11) until convergence. Goldstein (1991) and colleagues have implemented this procedure in the context of multilevel models involving nested random effects. Approximate estimation of the random effects is still a by-product of this approach, although their interpretation is not as clear as it is for the hierarchical model.

3.4 Asymptotic Justification

If (16) in fact correctly specifies the marginal mean of $\mathbf{y}$, then the estimating equations (19) are unbiased so that, under suitable regularity conditions, their root $\hat{\alpha}$ is consistent for α and asymptotically normally distributed (McCullagh and Nelder 1989, sec. 9.3). Provided also that (17) correctly specifies the variance, the asymptotic covariance matrix may be estimated by $\operatorname{cov}(\hat{\alpha}) = (\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}$. Because $E[\partial\mathbf{U}/\partial\theta] = \mathbf{0}$, furthermore, these asymptotic properties will continue to hold even if the parameters in $\mathbf{V}$ are replaced by consistent estimates. Substituting $\mathbf{V}^{-1}$ for $\mathbf{P}$ in (14), we see that these also are unbiased estimating equations and hence may be expected to yield consistent estimates for θ. Using $\mathbf{P}$ rather than $\mathbf{V}^{-1}$ is intended to alleviate the small sample bias that arises when $\hat{\alpha}$ is substituted for α, and it does not affect the asymptotic argument.

The most rigorous demonstrations to date of these properties are those of Liang and Zeger (1986) and Prentice (1988) for the case of block diagonal $\mathbf{V}$. Their arguments lead to an alternative, empirical estimate of $\operatorname{cov}(\hat{\alpha})$ that requires only that the regression model (16) for the marginal means be correctly specified.

4. COMPUTATIONAL ASPECTS

4.1 Initial Estimates of the Variance Parameters

The computations associated with the two procedures are nearly identical and may be described in broad outline as follows. Standard GLM techniques (McCullagh and Nelder 1989, sec. 2.5) lead to an initial estimate of α under the assumption that the n observations are independent ($\mathbf{D} = \mathbf{0}$). Residuals from this initial fit may be used to compute initial values for the parameters in $\mathbf{V}$, namely, θ and also ϕ if it is to be estimated. The exact procedure will depend on the particular problem. Any reasonable method is likely to be satisfactory for simple overdispersion models involving independent observations and a single variance component. We used moment equations (Moore 1986). In more complicated problems, the method of generalized least squares may be tried. Here the upper triangular elements of $\mathbf{Y}^* = (\mathbf{Y} - \mathbf{X}\hat{\alpha})(\mathbf{Y} - \mathbf{X}\hat{\alpha})^t + \mathbf{X}^t(\mathbf{X}^t\mathbf{WX})^{-1}\mathbf{X}$ are regressed on the appropriate design matrix for the variance components using ordinary least squares. Goldstein (1986, 1988) has shown that iterated generalized least squares, whereby the covariance matrix of $\mathbf{Y}^*$ is updated at each cycle and used in a weighted analysis, produces REML estimates under normal theory. In our experience with sparse discrete data, however, the initial estimates obtained with ordinary least squares often were not satisfactory. In such cases we set the parameters corresponding to covariances to 0 and started the variance parameters from small positive values.

4.2 Alternate Scoring for Means and Variances

The initial estimate θ^0 permits evaluation of the covariance matrix $\mathbf{D}$ that occurs in $\mathbf{V}$. Updated values for $(\hat{\alpha}, \hat{\mathbf{b}})$ under PQL are calculated from (7) and (8), or for $\hat{\alpha}$ alone from (19) under MQL, with the fitted values from the GLM fit being used to initialize the iterative solution of these equations by IWLS, as outlined in Section 2.2. Once the equations for the mean parameters have been solved, the variance scores (14) and expected Hessian (15) are used to take a Newton step towards a new value θ^1, after which one returns to (7)–(8) or (19) to reestimate the mean parameters. The Newton step is halved if the resulting $\mathbf{D}(\theta^1)$ is not positive definite. This simple procedure led to joint solutions of the mean and variance equations at an interior point of the parameter space in most of the examples considered. In general, however, one must anticipate all the problems that attend REML variance component estimation under standard normal theory. In the simulations we found that an adaptation of Marquardt's (1963) compromise between Fisher scoring and the method of steepest descent, as described by Harville (1977), was useful in the initial stages of the search.

Evaluation of $\mathbf{V}^{-1}$ and of the products of the $n \times n$ matrices occurring in the gradient and expected Hessian for variances is the most cumbersome aspect of the computation. Harville's Equation (3.6), corresponding to Green's (5.1) and (5.2), is useful for calculating $\mathbf{V}^{-1}$. Special structure in the covariance matrices must be exploited, however, if these procedures are to be used to estimate large numbers of random effects.

4.3 Scoring for REML Estimation in the Two-Level Problem

Longford (1987), using formulas of Lamotte (1972), developed a "fast scoring algorithm" for maximum likelihood (ML) estimation of variance components in normal theory models involving nested random effects. We extended the development in his Appendix for the two-level problem to obtain REML rather than ML estimates. Further extensions to accommodate more levels of nesting would be desirable. Prosser, Rasbash, and Goldstein (1990) have implemented the normal theory procedures for two and three levels using iterated generalized least squares. Because the procedures we describe involve repeated application of normal theory calculations to the working vector $\mathbf{Y}$ with associated covariance matrix $\mathbf{V}$, such algorithms may be employed for approximate inference in GLMM's (Goldstein 1991). The results reported in Sections 5 and 6 were obtained using the matrix programming language GAUSS (Aptech Systems 1988).

5. A SIMULATION STUDY

The simulation study followed the design specifications of Zeger and Karim (1991) to compare results with theirs obtained using a Bayesian approach and the Gibbs sampler. Each data set involved $K = 100$ clusters of size $n_k = 7$. Conditionally independent binary observations y_{kl} were generated within each cluster with conditional response probabilities given for $k = 1, \ldots, 100$ by

$$\text{logit } E(y_{kl} | b_k) = \alpha_0 + \alpha_1 t_l + \alpha_2 x_k + \alpha_3 x_k t_l + b_k^0 + b_k^1 t_l,$$

where $x_k = 1$ for half the sample and 0 for the other half and $t_l = l - 4$ for $l = 1, \ldots, 7$. The regression coefficients were fixed at $\alpha^t = (-2.5, 1, -1, -.5)$ while the random effects (b_k^0, b_k^1) were generated as a series of 100 iid normal variables with mean 0 and covariance structure

$$\mathbf{D}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{or} \quad \mathbf{D}_2 = \begin{pmatrix} .50 & 0 \\ 0 & .25 \end{pmatrix}.$$

(Zeger and Karim in fact used .49 rather than .50 for σ_{00} in $\mathbf{D}_2$ and set $\alpha_3 = .5$ rather than $-.5$, but this has negligible effect on the comparative results.) Two hundred data sets of 700 observations each were generated with $\mathbf{D}_1$ and 100 each were generated with $\mathbf{D}_2$. The entire experiment was replicated with binomial observations y_{kl} whose denominators were $m = 1, 2, 4, 8$.

Average values of the PQL regression coefficients were, in absolute value, less than the true values but approached the latter as the denominators of the individual binomial observations increased (see Table 1). Positive estimates of $\sigma_{00} = \text{var}(b_j^0)$ were obtained for all data sets sampled with $\mathbf{D} = \mathbf{D}_1$. This single variance component was seriously underestimated when $m = 1$ but was only moderately underestimated when $m = 8$. With binary data generated under $\mathbf{D}_2$, PQL frequently converged toward a non-positive definite covariance matrix, in which case we set the smaller variance and the covariance term, as well as the corresponding elements of their dispersion matrix, to 0. We estimated σ_{00} (resp. σ_{11}) to be 0 in 40% (resp. 12%) of samples with $m = 1$, 5% (8%) with $m = 2$, 1% (0%) with $m = 4$, and 0% (0%) with $m = 8$. Moreover, PQL tended to underestimate both variance components. Rather large discrepancies were observed between the simulated and estimated standard errors of the variance components for small m (see Table 2). The standard errors estimated for the regression coefficients, on other hand, agreed reasonably well with the simulated standard errors.

Even though the simulated data were generated under the hierarchical model, we also analyzed them under the marginal model using MQL to study the accuracy of the ap-

Table 1. Mean Values of Parameter Estimates in the Simulation Study

		Parameter						
Method	*m*	σ_{00}	σ_{01}	σ_{11}	α_0	α_1	α_2	α_3
					D = D₁ (200 replications)			
	True value:	1.00	—	—	−2.50	1.00	−1.00	−.50
PQL	1	.68	—	—	−2.31	.93	−.94	−.42
	2	.79	—	—	−2.36	.95	−.90	−.46
	4	.82	—	—	−2.38	.96	−.93	−.46
	8	.90	—	—	−2.46	.98	−.94	−.48
Gibbs sampler[a]	1	1.21	—	—	−2.67	1.07	−.96	.49[b]
					D = D₂ (100 replications)			
	True value:	.50	.00	.25	−2.50	1.00	−1.00	−.50
PQL	1	.35	−.05	.15	−2.30	.91	−.80	−.41
	2	.43	−.04	.17	−2.32	.93	−.84	−.43
	4	.36	−.00	.20	−2.35	.94	−.85	−.46
	8	.41	−.01	.22	−2.38	.95	−.85	−.48
Gibbs sampler[a]	1	.83	−.04	.37	−2.74	1.10	−1.14	.57[b]

[a] From Zeger and Karim (1991).
[b] See the text.

Table 2. Comparison of Simulated and Estimated Standard Errors

Method	m		Parameter						
			σ_{00}	σ_{01}	σ_{11}	α_0	α_1	α_2	α_3
			$D = D_1$ (200 replications)						
PQL	1	Sim[a]	.29	—	—	.30	.11	.50	.20
		Est[b]	.35	—	—	.28	.12	.46	.20
	2	Sim	.24	—	—	.22	.08	.31	.13
		Est	.25	—	—	.22	.08	.34	.14
	4	Sim	.20	—	—	.19	.06	.28	.10
		Est	.19	—	—	.18	.06	.28	.10
	8	Sim	.18	—	—	.17	.05	.24	.07
		Est	.17	—	—	.17	.05	.25	.07
Gibbs sampler[c]	1	Sim	.63	—	—	.40	.16	.61	.25
		Est	.60	—	—	.36	.15	.56	.24
			$D = D_2$ (100 replications)						
PQL	1	Sim[a]	.30	.10	.10	.28	.12	.41	.19
		Est[b]	.41	.16	.10	.26	.13	.42	.20
	2	Sim	.23	.07	.09	.21	.11	.37	.18
		Est	.30	.12	.08	.20	.10	.31	.16
	4	Sim	.15	.06	.06	.16	.09	.24	.14
		Est	.18	.08	.06	.15	.09	.23	.13
	8	Sim	.14	.05	.04	.14	.07	.22	.10
		Est	.13	.06	.05	.13	.08	.20	.12
Gibbs sampler[c]	1	Sim	.50	.12	.14	.38	.17	.54	.31
		Est	.59	.21	.22	.37	.19	.57	.28

[a] Simulated standard error.
[b] Root mean estimated variance.
[c] From Zeger and Karim (1991).

proximation (18) and to evaluate the distortions in the estimated covariance matrix caused by the failure to model the mean correctly. For this balanced design the MQL estimates of the fixed effects were identical to the standard logistic regression estimates there were used as starting values (McCullagh and Nelder 1989, ex. 14.8 and 14.9). Thus, regardless of the binomial denominator m, $\hat{\alpha}$ was estimating the same quantity. With $m = 8$ and $\mathbf{D}_1$, the average value of $\hat{\alpha}$ was $(-2.19, .87, -.92, -.40)^t$, which may be compared with the approximate value of $.862 \times (-2.5, 1, -1, -.5) = (-2.16, .86, -.86, -.43)$ based on (18). The estimates of σ_{00} and their standard errors were quite similar to those obtained under PQL. With $\mathbf{D}_2$, in contrast, MQL tended to overestimate σ_{00} and to give a negative value (average of $-.15$) to σ_{01}. Such behavior would be anticipated from the discussion in Section 3. According to (18), the "true" linear predictor for this problem is approximately equal to the assumed linear predictor plus

$$c^*_{kl}(\alpha_0 + \alpha_1 t_l + \alpha_2 x_k + \alpha_3 x_k t_l),$$

where $c^*_{kl} = c_{kl} - \bar{c}$ and $\bar{c}$ denotes the average of the c_{kl}. This equals $c^*_{kl}(-2.5 + 1.0t_l)$ when $x_k = 0$ and $c^*_{kl}(-1.5 + 1.0t_l)$ when $x_k = 1$. Because in both cases the constant and t coefficient are of opposite signs, the extra variability translates into a negative covariance when modeled via the random term $b^0_k + b^1_k t_l$.

6. ILLUSTRATIVE EXAMPLES

6.1 Overdispersion

Crowder (1978, table 3) presented data on the proportion of seeds that germinated on each of 21 plates arranged according to a 2 × 2 factorial layout by seed variety and type of root extract. He noted that the within-group variation exceeded that predicted by binomial sampling theory, and he was concerned that his logistic regression analysis of treatment and interaction effects should account appropriately for such overdispersion. One rather natural way of accounting for the extraneous plate-to-plate variability in this situation is by means of a GLMM that has linear predictor

$$\text{logit } \Pr[y_i = 1 \mid x_i, b_i] = \eta^b_i = x^t_i \alpha + b_i,$$

$i = 1, \ldots, 21$, where α represents the fixed effects associated with seed and extract and the b_i, assumed iid $\mathcal{N}(0, \sigma^2)$, represent random effects associated with each plate. For the agriculturalist interested in the effects of seed variety and root extract treatment on germination rates, it is more appropriate to model the marginal probabilities of germination (averaged over plates). In contrast, the hierarchical model is of interest in selecting plates containing subgroups of seeds that may have particularly high germination rates.

Table 3 presents the regression coefficients in linear logistic models fitted to the 21 binomial proportions of seed ger-

Table 3. Model Fits to Crowder's Seed Data

	Method of analysis			
	LR[a]	PQL	MQL	ML
Variable	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$
	Main effects model			
Constant	−.430 ± .114	−.375 ± .182	−.369 ± .180	−.389 ± .166
Seed (2)	−.270 ± .155	−.363 ± .228	−.357 ± .227	−.347 ± .215
Extract (2)	1.065 ± .144	1.012 ± .224	.998 ± .222	1.029 ± .205
σ	—	.352 ± .118	.349 ± .117	.295 ± .112
	Interaction model			
Constant	−.558 ± .126	−.542 ± .190	−.536 ± .190	−.548 ± .167
Seed (2)	.146 ± .223	.077 ± .308	.074 ± .308	.097 ± .278
Extract (2)	1.318 ± .177	1.339 ± .270	1.326 ± .269	1.337 ± .237
Interaction	−.778 ± .306	−.825 ± .430	−.816 ± .429	−.811 ± .385
σ	—	.313 ± .121	.313 ± .120	.236 ± .110

[a] LR = ordinary logistic regression.

mination. The results for ML analysis under the hierarchical model were obtained using the program EGRET (SERC 1989), which evaluates the integrated likelihood (2) and its logarithmic derivatives using Gaussian quadrature. This gave smaller estimates of the overdispersion variance component than did PQL and MQL, and the interaction between seed type and root extract appeared slightly more significant as a consequence. The parameter estimates using MQL were noticeably attenuated in comparison with those for PQL, as would be anticipated from Sections 3 and 5. From a practical viewpoint, however, there is little to choose between the two analyses. Figure 1 presents the observed proportions and corresponding fitted values from PQL under the main effects model. There was substantial shrinkage toward the estimated group means when the random effects were incorporated in the linear predictor, especially for proportions with small denominators.

Rotnitzky and Jewell (1990, sec. 4.4) considered robust Wald and score tests for these data based on the empirical variance. Their Wald tests using a "working" exchangeable correlation structure agreed quite well with those calculated from the ML estimates and standard errors shown in Table 3. The corresponding tests for PQL and MQL were slightly conservative in comparison.

6.2 Longitudinal Data

Thall and Vail (1990, table 2) presented data from a clinical trial of 59 epileptics who were randomized to a new drug (Trt = 1) or a placebo (Trt = 0) as an adjuvant to the standard chemotherapy. Baseline data available at entry into the trial included the number of epileptic seizures recorded in the preceding 8-week period and age in years. The logarithm of $\frac{1}{4}$ the number of baseline seizures (Base) and the logarithm of age (Age) were treated as covariables in the analysis. A multivariate response variable consisted of the counts of seizures during the 2-weeks before each of four clinic visits (Visit, coded $\text{Visit}_1 = -3, -1, 1, \text{Visit}_4 = 3$). Preliminary analysis indicated that the counts were substantially lower during the fourth visit and a binary variable (V4 = 1 for fourth visit, 0 otherwise) was constructed to model such effects.

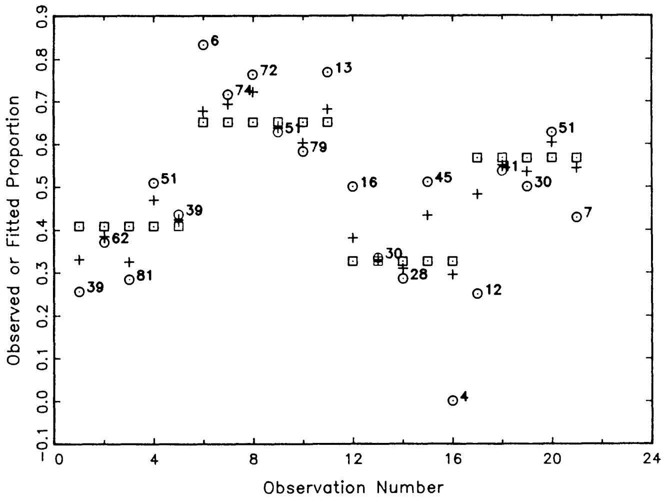

Figure 1. Proportion of Seeds That Germinated. The observed proportions are plotted as circles (⊙), followed by their denominators. Fitted proportions under the fixed and random effects models are plotted as squares (⊡) and plusses (+).

Thall and Vail's analysis focused on estimating the regression coefficients in a log-linear model for the marginal event rates and the covariance parameters in various patterned covariance matrices. Most of the latter were derived from a hierarchical, conditionally Poisson model that included one set of independent random effects associated with each subject and another set associated with each visit. By carefully examining residuals comparing the observed and fitted counts of each subject at each visit, they identified as "outliers" a number of patients who either had particularly large counts relative to the fitted model or had marked changes over time in their counts, or both.

Our reanalysis of these data using a GLMM was oriented primarily toward more systematic identification of patients who had extreme levels, or extreme degrees of change over time, in their event rates. We assumed that y_{jk}, the seizure count for patient j on the kth visit, was (conditionally) Poisson-distributed with mean μ^b_{jk}. The most general model considered was

$$\log \mu^b_{jk} = x^t_{jk}\alpha + b^1_j + b^2_j \text{Visit}_k / 10 + b^0_{jk},$$

where $\mathbf{x}_{jk}$ denotes the vector of treatment, visit, and covariable main effects and interactions; (b^1_j, b^2_j) are bivariate normal random effects that represent the residual level and rate of change in the event rate for the jth subject; and the b^0_{jk} are additional random error terms that represent nonspecific overdispersion beyond that introduced by the subject-to-subject variation. In view of the discussion in Section 3.1, the regression coefficients α for all but the constant term may be interpreted also as log (relative) event rates in the corresponding marginal model.

Table 4 presents the results obtained with PQL. Model I involved a standard Poisson regression analysis, without accounting for the intrasubject correlation and overdispersion; as a result, the standard errors of the subject level variables were seriously in error. There was a marked attenuation in the regression coefficients, especially the constant, and a sharp increase in the standard errors when random subject effects were introduced (Model II) and a further reduction in the fixed effect associated with the fourth visit with the addition of unit level random variation (Model III). (Models II and III correspond to Models 52 and 42 of Thall and Vail; our numerical results were similar to those shown in their table 4 for the closely related Model 22.)

Model IV was the most interesting. There was substantial heterogeneity among subjects, even after accounting for the treatment and baseline variables, both in the overall level and in the trend in mean seizure counts. We estimated the correlation between these two random effects to be effectively 0. Figure 2 graphs the random effects estimated for the 59 subjects and identifies by number those with particularly extreme levels or changes in attack rates, even after covariable adjustment. We identified patient 135 as having marked improvement over time after an initially high seizure rate, and identified patients 227, 225, and 112 as having the highest

Table 4. PQL Model Fits to Thall and Vail's Epilepsy Data

	Model			
	I	*II*	*III*	*IV*
Variable	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$
Fixed effects	−2.76 ± .41	−1.25 ± 1.2	−1.27 ± 1.2	−1.27 ± 1.2
Constant				
Base	.95 ± .04	.87 ± .14	.86 ± .13	.87 ± .14
Trt	−1.34 ± .16	−.91 ± .41	−.93 ± .40	−.91 ± .41
Base × Trt	.56 ± .06	.33 ± .21	.34 ± .21	.33 ± .21
Age	.90 ± .12	.47 ± .36	.47 ± .35	.46 ± .36
V4	−.16 ± .05	−.16 ± .05	−.10 ± .09	—
Visit/10	—	—	—	−.26 ± .16
Subject level random effects				
Constant ($\sqrt{\sigma_{11}}$)	—	.53 ± .06	.48 ± .06	.52 ± .06
Visit/10 ($\sqrt{\sigma_{22}}$)	—	—	—	.74 ± .16
Covariance (σ_{12})	—	—	—	−.01 ± .03
Unit level random effects				
Constant ($\sqrt{\sigma_{00}}$)	—	—	.36 ± .04	—

overall count levels relative to expectation based on the covariables. Because we expressed the random effects on the same log scale as the fixed effects, moreover, we also identified patients with especially low or zero counts (e.g., patient 232) that were not so apparent in Thall and Vail's analysis.

6.3 Smoothing of Birth Cohort Effects in an Age-Cohort Model

Breslow and Day (1975, table 2) analyzed breast cancer rates in Iceland according to year of birth in $K = 11$ cohorts from 1840–1849 to 1940–1949 and age in $J = 13$ groups from 20–24 years to 80–84 years. They fitted a log-linear model with fixed effects for age and cohort, both treated as factors, to the two-way table of rates using Poisson regression with the logarithm of the person-years denominators as an offset in the regression equation. This yielded an excellent fit, with a deviance of 49.7 on 54 degrees of freedom. Because case ascertainment was limited to the period 1910–1971, no data were recorded for the younger age groups in the older cohorts nor for the older age groups in the recent cohorts; 63 of the 11 × 13 = 143 cells in the two-way layout were empty. Moreover, the 1840 (11 cases) and 1940 (7 cases) cohorts were so small that the fixed effects estimated for them were particularly unstable. To try to reduce the random error in the relative risks estimated for these extreme cohorts, we fitted a log-linear model with a single term for the logarithm of birth cohort number in place of the cohort factor. This increased the deviance by only 10.8, with an increase of 9 degrees of freedom. Figure 3 plots the two sets of fixed cohort effects, with the fifth (1880) cohort selected as baseline.

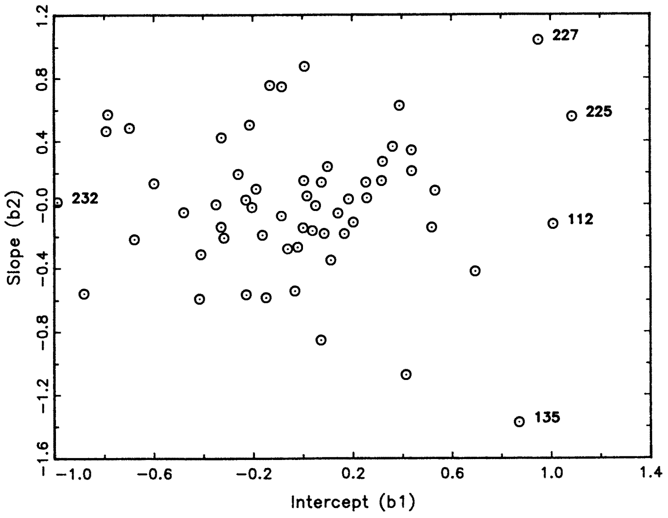

Figure 2. Random Intercepts and Slopes for Epilepsy Patients. Individual patients are identified according to the ID number in Table 2 of Thall and Vail (1990).

Despite the excellent fit of the linear model, however, there was concern that the strong parametric assumption was unwarranted. To explore this possibility, we used a nonparametric smoother based on a GLMM with an autoregressive error component. The logarithm of the mean number of breast cancer cases in the jth age group and kth birth cohort was assumed to satisfy

$$\log(\mu_{jk}) = \log n_{jk} + \alpha_j + \beta k + \sigma_0 u_k + \sigma_1 v_k,$$

where n_{jk} denotes the person-years denominator and α_j denotes the fixed effect of age. The fixed effect trend β and the random effect vectors $\mathbf{u}$ and $\mathbf{v}$ modeled three aspects of the variation of rates with date of birth.

The effects v_k were assumed to be independent $\mathcal{N}(0, 1)$, modeling unstructured heterogeneity of risk over birth cohorts. The model for $\mathbf{u}$, intended to represent smooth variation over time, was specified in the forward direction as a Gaussian autoregressive model. We initially considered a simple random walk model, as is used commonly in Bayesian forecasting (Harrison and Stevens 1976). But to improve performance at the endpoints, we decided instead on the model in which each point is predicted by linear extrapolation from its two immediate predecessors rather than from

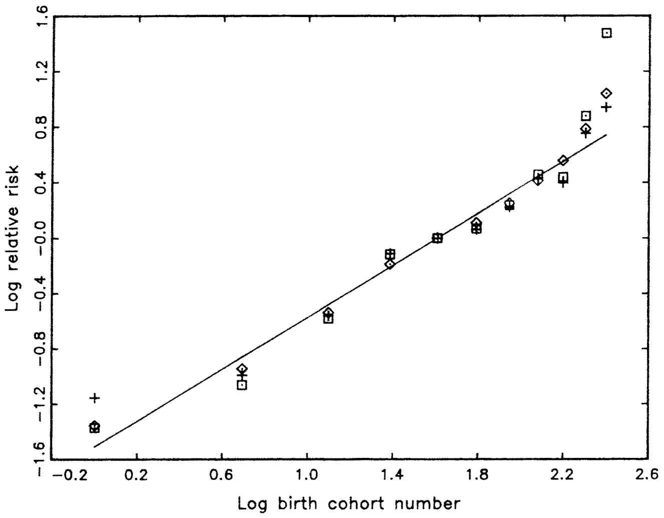

Figure 3. Relative Rates of Breast Cancer Incidence in Iceland, by Coded Year of Birth. The logarithms of fitted rate ratios, using birth cohort number 5 as standard, are plotted for the fixed-effects model as squares (□), for the independent random effects model as plusses (+), and for the autoregressive random effects model as diamonds (◇). The straight line (——) represents the fitted rate ratios for the fixed-effects model with a linear term for the logarithm of the coded birth cohort.

just the latest one. Thus for $k > 2$ we assumed

$$E(u_k | u_j, j < k) = 2u_{k-1} - u_{k-2}$$

and

$$\operatorname{var}(u_k | u_j, j < k) = 1, \qquad (21)$$

which treats the *second differences* of the series as independent $\mathcal{N}(0, 1)$ variates. The inverse dispersion matrix for $\mathbf{u}$ is given by

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ -2 & 1 & 0 & \cdots \\ 1 & -2 & 1 & \cdots \\ 0 & 1 & -2 & \cdots \\ 0 & 0 & 1 & \cdots \\ \cdots & \cdots & \cdots & \text{etc.} \end{pmatrix} \times \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots \\ 0 & 1 & -2 & 1 & \cdots \\ 0 & 0 & 1 & -2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \text{etc.} \end{pmatrix} \qquad (22)$$

$$= \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ -2 & 5 & -4 & 1 & 0 & 0 & \cdots \\ 1 & -4 & 6 & -4 & 1 & 0 & \cdots \\ 0 & 1 & -4 & 6 & -4 & 1 & \cdots \\ \cdots & \cdots & \cdots & \text{etc.} & \cdots & \cdots & \cdots \end{pmatrix}. \qquad (23)$$

Equation (23) gives an alternative specification of the model as an *undirected* conditional regression model. For $2 < k < K - 1$,

$$E(u_k | u_j, j \neq k) = \frac{1}{6}(4u_{k-1} + 4u_{k+1} - u_{k-2} - u_{k+2})$$

and

$$\operatorname{var}(u_k | u_j, j \neq k) = \frac{1}{6}, \qquad (24)$$

so that the conditional expectation is obtained by cubic interpolation from the two points on either side. The first (and last) point in the series may be seen to have conditional expectations given by linear extrapolation from the adjacent two points and conditional variance 1. The second (and penultimate) point has conditional expectation intermediate between quadratic and linear interpolation, which would correspond to rows of $\mathbf{R}$ of $(-1, 2, -1, 0, \ldots)$ and $(-1, 3, -3, 1, \ldots)$. $\mathbf{R}$ has rank $K - 2$ in this model, reflecting the fact that both *level* and *trend* are aliased. The former is taken up by the period parameters α_j. The latter is included as a fixed effect, βk, for technical reasons described in Section 2.3.

Not surprisingly, in view of the goodness of fit of the model with a strong linear trend in cohort effects, the estimate for σ_1 converged to 0 when both $\mathbf{u}$ and $\mathbf{v}$ terms were included in the model. When the autoregressive component alone was included, we estimated $\hat{\sigma}_0 = .12 \pm .06$ using PQL. With independent random effects alone, excluding also βk, we found $\hat{\sigma}_1 = .69 \pm .17$. Figure 3 shows the cohort effects estimated under these two mixed models as well as those for the fixed effects models. The greatest differences between the estimated random effects occur at the endpoints, where the autoregressive effects are fitted by linear extrapolation but the independent effects are pulled in towards the common mean. Differences also are evident at interior points, where the log-relative risk is pulled towards the local cubic fit by the autoregressive model and towards the overall mean by the independence model. Both mixed models yield more reasonable estimates than does the highly variable fixed effect at the upper endpoint.

6.4 A Mixed Model for the Log Odds Ratio

Kneale (1971) classified deaths from childhood cancer and matched controls in the Oxford region by age at death, year of birth, and whether or not the mother reported having received pelvic radiation during pregnancy. These data were arranged by Breslow (1976) into a series of 2 × 2 tables of cases versus controls and x-ray versus no x-ray, one table for each of 120 combinations of age 0–9 and birth year 1944–1964. Conditioning on the marginal totals in each table, he fitted log-odds-ratio regression models using maximum likelihood techniques based on the noncentral hypergeometric distribution (Zelen 1971). Although the relative risk for radiation was reasonably constant across age groups, there was a marked decline with year of birth, which was interpreted as an effect of temporal decline in radiation dose. The fixed-effects analysis seemed to suggest some curvature, as measured by a quadratic term in the regression of the log-relative risk on year of birth (Fig. 4). But doubts about the significance of such curvature were raised by the excess scatter of the individually estimated relative risks (shown as "fixed effects" in Fig. 4) about the regression line. This source of variation had not been formally considered in the earlier analysis.

Mixed models for log-odds-ratio regression are easily accommodated within this article's general framework. Although the exact conditional means and variances of the noncentral hypergeometric required for (quasi)likelihood analysis are prohibitively complicated when the marginal totals are large, simple approximations suggested by Mc-

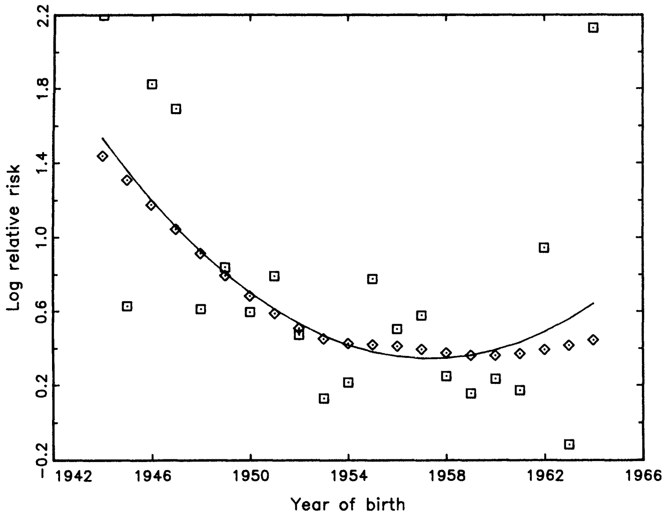

Figure 4. Relative Risks of Childhood Cancer in the Oxford Region for Children Exposed to In Utero Irradiation Versus the Nonexposed, by Year of Birth. The logarithms of fitted relative risks are plotted for the fixed-effects model as squares (□) and for the autoregressive random effects model as diamonds (◇). The curved line (——) represents the fitted relative risks in a fixed-effects model with linear and quadratic terms for year of birth.

Cullagh (1984) are accurate enough for most practical purposes (Breslow and Cologne 1986). Denote the odds ratio of the expected values in the ith table by ψ_i and denote the two sets of marginal totals by (n_{i1}, n_{i2}) and (m_{i1}, m_{i2}), where $n_{i1} + n_{i2} = m_{i1} + m_{i2} = N_i$. Rather than parameterize the model in terms of the mean $\mu_i = E(y_i)$, where y_i is the observed frequency in the upper left cell, it is more convenient here to parameterize it directly in terms of the odds ratio. Thus we suppose that $\psi_i^b = \exp(\mathbf{x}_i^t\alpha + \mathbf{z}_i^t\mathbf{b})$ under PQL or $\psi_i = \exp(\mathbf{x}_i^t\alpha)$ under MQL. Using McCullagh's approximation, the mean μ_i and variance v_i jointly satisfy the equations

$$\psi_i = \frac{\mu_i(\mu_i + n_{i1} - m_{i1}) + v_i}{(m_{i1} - \mu_i)(n_{i2} - \mu_i) + v_i}$$

and

$$v_i = \frac{N_i}{(N_i - 1)}\left\{\frac{1}{\mu_i} + \frac{1}{(\mu_i + n_{i1} - m_{i1})} + \frac{1}{(m_{i1} - \mu_i)} + \frac{1}{(n_{i2} - \mu_i)}\right\}^{-1}.$$

Because log ψ is the canonical parameter for the noncentral hypergeometric distribution (McCullagh and Nelder 1989, sec. 7.3.2), the link derivative $g'(\mu)$ equals the inverse variance $v^{-1}(\mu)$, so that the denominator terms drop out of the PQL score equations (7) and (8). The link derivative so defined is needed, however, to calculate the GLM iterative weights that enter into the PQL and MQL estimation procedures.

Table 5 shows the results of several PQL model fits to the Oxford childhood cancer data using the previously described approximation. The general form of the fitted models was

$$\log \psi_{jk} = \alpha + \beta_1 \mathrm{Year}_k + \beta_2(\mathrm{Year}_k^2 - 22) + \sigma\mu_k,$$

where ψ_{jk} represents the log-relative risk of radiation in the 2×2 table formed for the jth age group and the kth year of birth. Year$_k$ is coded −10 for 1944, −9 for 1945, . . . , 10 for 1964, and μ_k is an iid $\mathcal{N}(0, 1)$ error term representing extraneous year-to-year variation. For $\sigma = 0$ the results were identical to those reported by Breslow and Cologne (1986, table 3) for the McCullagh approximation. Note the marked decrease in the estimated variance component as the linear birth cohort effect is added to the fixed part of the model. There is no clear evidence for extraneous variation of the log-relative risk about the regression line. Even if such variation is accounted for in the model, however, the coefficient of the quadratic term remains statistically significant.

We also fitted an autoregressive model, without the quadratic term, to estimate "nonparametrically" the evolution of risk with time. Using the specifications of Section 6.3 for μ_k, this model clearly showed the flattening of risk as radiation dosage was controlled during the mid-1950s (Fig. 4). The fixed-effects quadratic fit was less satisfactory, because it suggested a sharper increase in risk in the early 1960s, which seemed more a consequence of the parametric formulation than of any clear feature of the data.

6.5 Spatial Aggregation in Scottish Lip Cancer Rates

Clayton and Kaldor (1987) analyzed observed and expected numbers of lip cancer cases in the 56 counties of Scotland with a view toward producing a map that would display regional variations in cancer incidence yet avoid the presentation of unstable rates for the smaller counties. The expected numbers had been calculated accounting for the different age distributions in the counties using a fixed-effects multiplicative model but were regarded for purposes of analysis as constants based on an external set of standard rates. Thus, conditional on a set of values b_i representing county-specific log-relative risks (i.e., standardized morbidity ratios, or SMR's), the observed numbers of cases y_i, $i = 1, \ldots, 56$ were assumed to have independent Poisson distributions with means $\mu_i^b = n_i\exp(\alpha + b_i)$. Here the n_i denote the expected numbers and α, the grand mean, plays the role of the logarithm of the overall SMR, which one would expect to be near unity in view of the way the expected numbers were derived.

Clayton and Kaldor (1987) proposed empirical Bayes estimation of the county-specific SMR's using several alternative assumptions about the distribution of the random effects. Specification of these as a random sample from a

Table 5. PQL Model Fits to the Oxford Childhood Cancer Data

Regression coefficients ± standard error			
Constant	*YEAR*	*YEAR² − 22*	*σ*
.505 ± .056	—	—	—
.531 ± .076	—	—	.23 ± .07
.516 ± .056	−.0385 ± .0144	—	—
.536 ± .070	−.0406 ± .0162	—	.16 ± .09
.565 ± .061	−.0445 ± .0149	.0067 ± .0030	—
.566 ± .070	−.0469 ± .0167	.0071 ± .0033	.15 ± .10

log-gamma distribution led to an exact analysis based on the resulting independent negative binomial distributions for the y_i. To account for spatial aggregation, they also considered a model with normally distributed random effects. Approximation of the conditionally Poisson log-likelihood by a quadratic expansion centered at the empirical log SMR, $\tilde{b}_i = \log[(d_i + \frac{1}{2})/n_i]$, led to a solution based on the EM algorithm (Dempster, Laird, and Rubin 1977). There was, however, an inconsistency in their specification of the spatial process (Clayton and Bernadelli 1991).

Presumably the spatial aggregation is due in large part to the effects of environmental risk factors. Data were available on the percentage of the work force in each county employed in agriculture, fishing, or forestry. The authors who compiled the data (Kemp, Boyle, Smans, and Muir 1985) noted that this covariable, x_i, exhibited spatial aggregation paralleling that for lip cancer itself. Because all three occupations involve outdoor work, the authors suggested that exposure to sunlight, the principal known risk factor for lip cancer, might be the explanation. We analyzed the augmented data (Table

Table 6. Observed and Fitted SMR's for Lip Cancer in 56 Scottish Counties

					Fitted SMR's			
Co	*Obs*	*Exp*	*Cov*	*Obs SMR*	*I*	*II*	*III*	*Adjacent Counties*
1	9	1.4	16	652.2	495.5	473.9	446.3	5, 9, 11, 19
2	39	8.7	16	450.3	424.5	424.2	438.3	7, 10
3	11	3.0	10	361.8	310.6	305.9	352.1	6, 12
4	9	2.5	24	355.7	298.1	291.8	248.7	18, 20, 28
5	15	4.3	10	352.1	313.9	311.5	332.4	1, 11, 12, 13, 19
6	8	2.4	24	333.3	277.9	271.2	332.4	3, 8
7	26	8.1	10	320.6	300.6	300.0	308.2	2, 10, 13, 16, 17
8	7	2.3	7	304.3	253.3	246.7	308.7	6
9	6	2.0	7	303.0	247.1	238.9	240.3	1, 11, 17, 19, 23, 29
10	20	6.6	16	301.7	279.5	278.5	288.8	2, 7, 16, 22
11	13	4.4	7	295.5	265.0	262.8	302.9	1, 5, 9, 12
12	5	1.8	16	279.3	226.2	217.5	300.0	3, 5, 11
13	3	1.1	10	277.8	208.8	193.8	263.7	5, 7, 17, 19
14	8	3.3	24	241.7	213.3	210.0	171.4	31, 32, 35
15	17	7.8	7	216.8	204.8	204.1	181.1	25, 29, 50
16	9	4.6	16	197.8	182.3	180.6	192.9	7, 10, 17, 21, 22, 29
17	2	1.1	10	186.9	156.2	147.0	206.5	7, 9, 13, 16, 19, 29
18	7	4.2	7	167.5	156.7	155.0	131.3	4, 20, 28, 33, 55, 56
19	9	5.5	7	162.7	154.8	153.6	204.4	1, 5, 9, 13, 17
20	7	4.4	10	157.7	149.3	147.8	146.7	4, 18, 55
21	16	10.5	7	153.0	149.2	148.7	141.5	16, 29, 50
22	31	22.7	16	136.7	135.7	135.5	142.8	10, 16
23	11	8.8	10	125.4	124.5	123.9	115.3	9, 29, 34, 36, 37, 39
24	7	5.6	7	124.6	123.6	122.6	82.4	27, 30, 31, 44 ,47, 48, 55, 56
25	19	15.5	1	122.8	122.5	122.1	124.6	15, 26, 29
26	15	12.5	1	120.1	120.0	119.5	110.8	25, 29, 42, 43
27	7	6.0	7	115.9	116.6	115.7	101.4	24, 31, 32, 55
28	10	9.0	7	111.6	112.7	112.1	110.1	4, 18, 33, 45
29	16	14.4	10	111.3	112.1	111.7	117.4	9, 15, 16, 17, 21, 23, 25, 26, 34, 43, 50
30	11	10.2	10	107.8	109.3	108.8	82.7	24, 38, 42, 44, 45, 56
31	5	4.8	7	105.3	108.8	107.6	92.7	14, 24, 27, 32, 35, 46, 47
32	3	2.9	24	104.2	109.7	107.9	108.0	14, 27, 31, 35
33	7	7.0	10	99.6	103.2	102.4	91.0	18, 28, 45, 56
34	8	8.5	7	93.8	97.9	97.2	76.0	23, 29, 39, 40, 42, 43, 51, 52, 54
35	11	12.3	7	89.3	92.9	92.4	91.4	14, 31, 32, 37, 46
36	9	10.1	0	89.1	93.4	92.8	85.9	23, 37, 39, 41
37	11	12.7	10	86.8	90.6	90.1	83.0	23, 35, 36, 41, 46
38	8	9.4	1	85.6	90.8	90.1	65.0	30, 42, 44, 49, 51, 54
39	6	7.2	16	83.3	90.3	89.4	79.6	23, 34, 36, 40, 41
40	4	5.3	0	75.9	86.7	85.6	62.3	34, 39, 41, 49, 52
41	10	18.8	1	53.3	60.1	59.4	56.2	36, 37, 39, 40, 46, 49, 53
42	8	15.8	16	50.7	59.0	58.2	60.9	26, 30, 34, 38, 43, 51
43	2	4.3	16	46.3	69.8	67.6	71.9	26, 29, 34, 42
44	6	14.6	0	41.0	51.9	50.7	49.1	24, 30, 38, 48, 49
45	19	50.7	1	37.5	41.4	41.0	44.6	28, 30, 33, 56
46	3	8.2	7	36.6	55.2	53.0	54.3	31, 35, 37, 41, 47, 53
47	2	5.6	1	35.8	60.2	57.3	49.2	24, 31, 46, 48, 49, 53
48	3	9.3	1	32.1	50.7	48.2	43.0	24, 44, 47, 49
49	28	88.7	0	31.6	34.2	34.0	35.9	38, 40, 41, 44, 47, 48, 52, 53, 54
50	6	19.6	1	30.6	41.3	39.9	51.9	15, 21, 29
51	1	3.4	1	29.1	65.0	61.0	51.3	34, 38, 42, 54
52	1	3.6	0	27.6	63.5	59.3	45.5	34, 40, 49, 54
53	1	5.7	1	17.4	51.6	45.5	38.6	41, 46, 47, 49
54	1	7.0	1	14.2	47.1	40.2	40.4	34, 38, 49, 51, 52
55	0	4.2	16	.0	58.5	42.0	67.7	18, 20, 24, 27, 56
56	0	1.8	10	.0	70.4	61.5	68.3	18, 24, 30, 33, 45, 55

NOTE: I, Clayton and Kaldor; II, PQL, independence; and III, PQL, spatial correlation.

6) using a conditionally independent Poisson model with the conditional means for county i specified by

$$\log \mu_i^b = \log n_i + \alpha_0 + \alpha_1 x_i/10 + b_i.$$

Two separate assumptions were made regarding the distribution of the random effects b_i: (a) iid $\mathcal{N}(0, \sigma^2)$ and (b) Gaussian intrinsic autoregression. The latter specified the inverse variance matrix $\mathbf{R}$ according to Besag et al. (1991) and gave $\mathbf{b}$ an improper density proportional to $\exp\{-\Sigma_{i\sim j}(b_i - b_j)^2/(2\sigma^2)\}$, where $i \sim j$ denotes adjacent counties. Here the conditional expectation of b_i given b_j, $j \neq i$ equals the mean of the b_j in contiguous counties, whereas the conditional variance equals σ^2 divided by their number.

Table 7 presents regression coefficients estimated by PQL for models with and without the covariable x_i and with both independence and autoregressive structures for the random effects. Note the slight attenuation in the coefficient of the covariable when independent random effects are incorporated and the much more marked attenuation under the spatial structure. Similarly, there is a reduction in the variance components when the covariable is included. This confirms that much of the spatial aggregation is explained by the percentage of the work force engaged in outdoor occupations. Caution should be exercised in interpreting this result, however, because the model assumed that the random effects were unrelated to the covariables. If the other factors that contributed to the spatial correlation were positively correlated with the work force covariable, which seems likely, then the regression coefficient estimated for this variable would be attenuated, with some of the work force effect being attributed instead to location.

Table 6 presents the SMRs estimated for each county under Clayton and Kaldor's (1987) approximation to the normal theory independence model (Model I), under PQL with independent random effects (Model II), and under PQL with spatial autocorrelation (Model III). Because Clayton and Kaldor had not considered the covariable, it was omitted from the PQL fits to achieve comparability. All models realized the fundamental goal of pulling in the extreme SMRs based on small numbers of observed cases. Estimated SMRs for Models II and III, however, were markedly different for those counties for which the mean SMR of the neighbors differed notably from the overall mean. For example, the SMR for county 19 (note that the counties are ordered according to the observed SMR) was pulled down from 162.7 to 153.6 by the independence model, but upwards to 204.2 by the spatial model. Its SMR was higher than average, but those of the neighboring counties were higher still. For counties with 20 or more cases, the PQL estimates under independence (Model II) agreed closely with those of Clayton and Kaldor (Model I). For smaller counties, the PQL estimates were generally lower. The difference can be attributed largely to their use of the .5 correction in the empirical log SMR used to center the approximation, which acts to increase this quantity above the level actually observed.

Table 7. PQL Model Fits to the Scottish Lip Cancer Data

Regression coefficients ± standard error			
Constant	*x/10*	*σ(spatial)*	*σ(independence)*
.00 ± .04	—	—	—
.14 ± .11	—	—	.76 ± .09
.13 ± .05	—	.86 ± .14	—
−.54 ± .07	.74 ± .06	—	—
−.44 ± .16	.68 ± .14	—	.60 ± .08
−.18 ± .12	.35 ± .12	.73 ± .13	—

6.6 Crossed Random Effects: The Salamander Data

McCullagh and Nelder (1989, sec. 14.5) published an interesting set of data on the success of matings between male and female salamanders drawn from two populations, the roughbutts (RB) and the whitesides (WS), that had been geographically isolated from each other. In the first of three experiments, conducted during the summer of 1986, 10 RB females and 10 WS females were mated with three RB males and three WS males, for a total of six matings each over 24 days. Each of 10 RB males and 10 WS males likewise served as mates for three females of each type. These same 40 salamanders were used in a repeat experiment conducted in the fall that involved no repetitions of the earlier male–female pairs. A third experiment, also conducted in the fall, used a new set of 40 animals. Each experiment involved 30 matings of each of the four gender–population combinations. Simple inspection of the data revealed that three of the crosses had success rates of about 70%, whereas the mating of WS females with RB males was successful only 25% of the time. Evaluating the statistical significance of these differences was complicated by the fact that the 360 binary responses were not independent.

McCullagh and Nelder considered a linear logistic model for the marginal probabilities of success, using a linearization as in Section 3.1 to derive an approximate covariance matrix for the vector of 360 binary outcomes. Because of the balanced design, their quasi-likelihood estimates of the regression parameters were identical to those obtained from standard logistic regression under independence (McCullagh and Nelder 1989, ex. 14.8 and 14.9). They used a method of moments procedure based on the observed covariance matrix of the residuals (observed binary minus fitted values) to estimate the variance components.

Karim and Zeger (1992) reanalyzed these data under a hierarchical model, using the Gibbs sampler with a noninformative prior on the dispersion matrix to approximate the posterior distributions of the parameters and random effects of interest. Denoting by y_{ijk} the response for female i and male j in experiment k, their model B may be written

$$\operatorname{logit} \Pr[y_{ijk} = 1 \,|\, \mathbf{x}_{ijk}; \mathbf{b}_i^f, \mathbf{b}_j^m] = \mathbf{x}_{ijk}^t\alpha + \mathbf{z}_{ik}^t\mathbf{b}_i^f + \mathbf{z}_{jk}^t\mathbf{b}_j^m,$$

where $\mathbf{b}_i^f$ and $\mathbf{b}_j^m$ each have two components representing the random effects associated with the indicated animal in summer and fall. The fixed effects consisted of a constant, an indicator WS_F of whiteside females, an indicator WS_M of whiteside males, their interaction, and an indicator of fall season. They also considered Model A, in which the gender specific random effects from each experiment were assumed independent with equal variances, and Model C, in which separate fixed and random effects were estimated for each experiment. Models A and C were also considered by McCullagh and Nelder. Neither contained the fixed effect for fall season.

We fitted these same models with PQL. For Model B the variance matrix of Equation (20) took the form

$$\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}^{\mathrm{f}}\mathbf{D}^{\mathrm{f}}(\mathbf{Z}^{\mathrm{f}})^{t} + \mathbf{Z}^{\mathrm{m}}\mathbf{D}^{\mathrm{m}}(\mathbf{Z}^{\mathrm{m}})^{t},$$

where $\mathbf{Z}^{\mathrm{f}}$ and $\mathbf{Z}^{\mathrm{m}}$ identified the female and male animals involved in each mating. The covariance matrix $\mathbf{D}^{\mathrm{f}}$ of female random effects was

$$\mathbf{D}^{\mathrm{f}} = \begin{pmatrix} \sigma_{11}^{\mathrm{f}}\mathbf{I} & \sigma_{12}^{\mathrm{f}}\mathbf{I} & 0 \\ \sigma_{12}^{\mathrm{f}}\mathbf{I} & \sigma_{22}^{\mathrm{f}}\mathbf{I} & 0 \\ 0 & 0 & \sigma_{22}^{\mathrm{f}}\mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ is the identity matrix of dimension 20, σ_{11} is the variance of the summer effects, σ_{22} is the variance of the fall effects, and σ_{12} is the covariance. $\mathbf{D}^{\mathrm{m}}$ had a parallel structure. Using Harville's (1977) Equation (3.6) and standard formulas for the inversion of partitioned matrices, we calculated $\mathbf{V}^{-1}$ by inversion of matrices no larger than 60 × 60. But the need to consider 360 × 360 matrices in the variance equations (14) and (15) meant that this problem was about as large as could be handled reasonably by a microcomputer.

Table 8 presents variance components found with the three procedures. Those estimated by PQL and by McCullagh and Nelder's moments method were in reasonable agreement, with the exception of the summer male variance. Variances estimated by the Gibbs sampler were substantially larger, in accordance with our simulation results. The covariance matrix for males under Model B converged towards singularity with PQL; we thus constrained the summer and fall effects for individual males to be equal, reducing the number of estimated variance parameters from six to four. A high correlation between the summer and fall male effects was also evident with the Gibbs sampler. Absolute values of the regression coefficients for Models A and B (Table 9) were 25–30% larger under the Gibbs sampler than under PQL, which also accords with our simulations (see Sec. 5). The key interaction effect was highly significant regardless of the method of analysis. MQL actually would have been more appropriate in this example, because one is primarily concerned with estimating the marginal success rates associated with different types of matings, not in identifying individual salamanders with particularly high success rates.

Table 8. Variance Components for the Salamander Data

	Method of estimation					
	Moments[a]		Gibbs sampler[b]		PQL	
Season	σ_F^2	σ_M^2	σ_F^2	σ_M^2	σ_F^2	σ_M^2
Model A: Pooled over independent experiments						
Total	.91	.88	1.50	1.36	.72	.63
Model B: Correlated random effects						
Summer	—	—	1.92	1.25	1.09	.90[c]
Fall	—	—	1.37	2.02	.62	.90[c]
Covariance	—	—	−.25	1.52	−.12	.90[c]
Model C: Separate effects each experiment						
Summer '86	1.37	.70	2.35	.14	1.41	.09
Fall '86 rerun	.98	.60	2.99	1.42	1.26	.62
Fall '86	.40	1.34	.33	2.89	.26	1.50

[a] McCullagh and Nelder (1989), table 14.10.
[b] Karim and Zeger (1992), table 3, medians.
[c] Constrained to be equal.

Table 9. Regression Coefficients for the Salamander Data

	Model A		Model B	
	Gibbs sampler*	PQL	Gibbs sampler*	PQL
Regression parameter	$\hat{\beta} \pm SE(\hat{\beta})$	$\hat{\beta} \pm SE(\hat{\beta})$	$\hat{\beta} \pm SE(\hat{\beta})$	$\hat{\beta} \pm SE(\hat{\beta})$
Constant	1.03 ± .43	.79 ± .32	1.48 ± .64	1.18 ± .49
Fall season	—	—	−.62 ± .54	−.50 ± .41
WS_F	−3.01 ± .60	−2.29 ± .43	−3.13 ± .62	−2.43 ± .44
WS_M	−.69 ± .50	−.54 ± .39	−.76 ± .62	−.62 ± .46
$WS_F \times SW_M$	3.74 ± .68	2.82 ± .50	3.90 ± .72	3.01 ± .52

* Zeger and Karim (1991), table 3. *SE* computed as range of 90% *CI* ÷ 3.3.

7. DISCUSSION AND CONCLUSIONS

The limited simulation results and illustrative analyses presented in this article suggest that reasonably simple approximate methods are available for inference on random effects in the context of GLM's. These methods may be implemented by repeated application of normal theory mixed model procedures, and they are subject to the same computational limitations as regards sample size and variance structure. When the random effects are nested, even at three or more levels, algorithms developed by Goldstein (1986, 1988) and Longford (1987) suffice to treat large complex problems. The procedures discussed herein are still rather limited in their ability to handle crossed designs, however. The salamander problem, involving two sets of 60 random effects each, probably represents an approximate break even point, after which the Bayesian treatment based on Gibbs sampling becomes progressively easier to implement in comparison with PQL or MQL, due to their requirement for manipulation of large matrices. Of course, such comparisons are highly dependent on the available computing resources.

Our simulation results were encouraging as regards the ability of PQL to render approximately correct inferences on regression coefficients in hierarchical models. As anticipated by the development of Section 2, accuracy improved as the binomial denominators increased. Nonetheless, even with binary data the results were sufficiently accurate for many practical purposes. Inference on variance parameters was less satisfactory under PQL, due largely to the tendency of the procedure to converge to a nonpositive definite variance matrix when the binomial denominator was 1 or 2. When the response probabilities are small and the data are highly discrete, only limited information is present for estimating random effects and their associated variances and covariances. The Bayesian formulation enjoys an advantage here because of the information on variance components contributed by the prior distribution. It makes full use of the normal theory distributional assumptions, however, whereas PQL requires specification only of the first and second moments.

Another limitation of PQL, in common with empirical Bayes methods more generally, is the failure to account for the contribution of the estimated variance components when assessing the uncertainty in both random and fixed effects. One major distinction between PQL and MQL is the fact that the regression coefficients of the former, but not of the latter, depend strongly on the estimated variance components when the link function is not the identity, even in large samples. Development of an approximate covariance estimate for $(\hat{\alpha}, \hat{\theta})$ and refinements in the expressions given for the variance estimates of $\hat{\alpha}$ and $\hat{\theta}$ individually remain important problems for further research. (In the meantime, some encouragement can perhaps be taken from the results for regression coefficients in Table 2.) Bootstrap methods (Laird and Louis 1987) are used increasingly as a means of constructing confidence intervals for the estimated random effects, but these are controversial as regards their ability to fully reflect the relevant uncertainties. Such problems are handled more easily within the Bayesian context by examining posterior distributions.

Our simulation results for MQL, summarized briefly in the text, confirm the anticipated attenuation in the estimated regression coefficients when the binary data are generated under the hierarchical model yet analyzed under the marginal one. A notable feature is the spurious correlation induced in the estimated random effects by the failure to correctly model the heterogeneity in the fixed effects; see Equation (18). Much of the bias in the estimation of both mean and variance parameters might well be alleviated for the logistic model by treating the terms c_i of (18) as a *multiplicative offset* whose values depended on the current $\hat{\theta}$. We plan further research to investigate this question. Results of another simulation study of Liang and Zeger's (1986) generalized estimating equation (GEE) approach to clustered binary data (Sharples and Breslow 1992), for which the marginal means of the simulated data actually satisfied the assumed linear logistic equation, suggested that the regression parameters and their standard errors were well estimated, but the correlation parameters less so. Because the estimating equations for MQL and GEE are identical as regards the mean parameters, one might anticipate similar results for MQL. MQL (or GEE) is the method of choice when interest is focused on the marginal relationship between covariables and response, and the random effects model serves mainly to suggest a plausible covariance structure, as expressed in $\mathbf{V}$, that enables one to get reasonably efficient estimating equations for the mean value parameters. By contrast, PQL is the method of choice for estimating parameters in the hierarchical model, especially when attention is focused on the random effects.

Our work is closely related to that of several other research teams. The equivalence of MQL and Goldstein's (1991) procedure for GLMM's with nested random effects has been mentioned already. Earlier work on this topic was reported by Morton (1987, 1988). When applied to normally distributed data with nonlinear link functions, PQL is equivalent to the nonlinear mixed model procedure of Lindstrom and Bates (1990). Note in particular the correspondence between their Equation (3.3) and our Equation (6) and between their (4.3) and our (12). Their development extends ours to general nonlinear structures for the mean, whereas ours extends theirs to nonnormal observations having defined mean–variance relationships. A synthesis is clearly possible. Since this article was first submitted for publication, papers by Schall (1991) and McGilchrist and Aisbett (1991) have appeared, both of which discussed PQL for GLMM's that involve sums of independent random effects. Schall (1991) obtained estimates of regression coefficients and variance components for Model C of the salamander data that were identical to ours; he did not consider the correlated random effects Model B.

Liang and Waclawiw (1990) and Waclawiw and Liang (1991) proposed an interesting estimating equation approach to the random effects in GLMM's, in which the fixed effects were estimated from the approximating marginal model as in MQL and were then corrected for attenuation using Equation (18). Optimal estimating equations in the sense of Godambe (1960) were suggested for the random effects. When applied to the simple overdispersion problem involving a log-linear, conditionally Poisson model, their smoothed estimate incorporating the random effect is a weighted average of the actual observation and the fitted mean. With PQL and MQL, in contrast, such averaging is carried out with the working vector $\mathbf{Y}$ on the scale of the linear predictor. This approach seems more natural. Detailed comparisons of the properties of the random effects predicted by these two approaches would be a worthy subject for further research.

Other proposals have been made to introduce autoregressive variance structures into GLMM's. Zeger (1988) considered a time series model for the random effects in a simple overdispersion model for count data. The conditionally log-linear mean structure was preserved in the marginal model. He approximated the marginal variance matrix by a band diagonal matrix having correlation structure appropriate for the assumed autoregressive model, which was not precisely preserved. This enabled him to avoid the large-scale matrix inversion required by our approach to autoregressive structures. Development of similar approximations for other GLMM's is a high priority. Goldstein, Healy, and Rasbash (1991) also considered an autoregressive structure for the unit level random effects in the context of multilevel mixed models for continuous responses and estimated the nonlinear correlation parameters by the method of constructed variables. There are close connections here with the Bayesian approach to dynamic generalized linear models (West, Harrison, and Migon 1985).

The preceding citations give some indication of the current high level of interest in GLMM's. Our goal has been to provide a unified framework in which much of this work may be discussed and compared and to demonstrate by means of simulations and worked examples its potential for applications.

[*Received March 1991. Revised April 1992.*]

REFERENCES

Aptech Systems (1988), *The GAUSS System, Version 2.0,* Kent, WA: Author.

Barndorff-Nielsen, O. E., and Cox, D. R. (1989), *Asymptotic Techniques for Use in Statistics,* London: Chapman and Hall.

Besag, J., York, J., and Mollié, A. (1991), "Bayesian Image Restoration, with Two Applications in Spatial Statistics" (with discussion), *Annals of the Institute of Statistical Mathematics,* 43, 1-59.

Breslow, N. E. (1976), "Regression Analysis of the Log Odds Ratio: A Method for Retrospective Studies," *Biometrics,* 32, 409-411.

——— (1984), "Extra-Poisson Variation in Log-Linear Models," *Applied Statistics,* 33, 38-44.

Breslow, N. E., and Cologne, J. (1986), "Methods of Estimation in Log Odds Ratio Regression Models," *Biometrics,* 42, 949-954.

Breslow, N. E., and Day, N. E. (1975), "Indirect Standardization and Multiplicative Models for Rates, With Reference to the Age Adjustment of Cancer Incidence and Relative Frequency Data," *Journal of Chronic Diseases,* 28, 289-303.

Brillinger, D., and Preisler, H. K. (1986), "Two Examples of Quantal Data Analysis," in *Proceedings of the 13th International Biometrics Conference,* Seattle: The Biometric Society. pp. 94-113.

Carroll, R. J., and Ruppert, D. (1982), "Robust Estimation in Heteroscedastic Linear Models," *The Annals of Statistics,* 10, 429-441.

Clayton, D., and Bernardinelli, L. (1991), "Bayesian Methods for Mapping Disease Risk," in *Small Area Studies in Geographical and Environmental Epidemiology,* eds. J. Cuzick and P. Elliot, Oxford: Oxford University Press.

Clayton, D., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics,* 43, 671-681.

Cox, D. R., and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society,* Ser. B, 49, 1-39.

Crouch, E. A. C., and Spiegelman, D. (1990), "The Evaluation of Integrals of the Form $\int f(t)\exp(-t^2)dt$: Application to Logistic Normal Models," *Journal of the American Statistical Association,* 85, 464-469.

Crowder, M. J. (1978), "Beta-Binomial ANOVA for Proportions," *Applied Statistics,* 27, 34-37.

Dempster, A. P. (1972), "Covariance Selection," *Biometrics,* 28, 157-175.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1972), "Maximum Likelihood From Incomplete Observations," *Journal of the Royal Statistical Society,* Ser. B, 39, 1-38.

Firth, D. (1991), "Generalized Linear Models," in *Statistical Theory and Modelling,* eds. D. V. Hinkley, N. Reid, and E. J. Snell, London: Chapman and Hall, pp. 55-82.

Godambe, V. P. (1960), "An Optimum Property of Regular Maximum Likelihood Estimation," *The Annals of Mathematical Statistics,* 31, 1208-1211.

Goldstein, H. (1986), "Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares," *Biometrika,* 73, 43-56.

——— (1988), "Restricted Unbiased Iterative Generalized Least-Squares Estimation," *Biometrika,* 76, 622-623.

——— (1991), "Nonlinear Multilevel Models, With an Application to Discrete Response Data," *Biometrika,* 78, 45-51.

Goldstein, H., Healy, M., and Rasbash, J. (1992), "Time Series Models for Repeated Measures Data," unpublished manuscript, submitted to *Biometrics.*

Good, I. J., and Gaskins, R. A. (1971), "Nonparametric Roughness Penalties for Probability Densities," *Biometrika,* 58, 255-277.

Graybill, F. A. (1983), *Matrices With Applications in Statistics* (2nd ed.), Belmont, CA: Wadsworth.

Green, P. J. (1987), "Penalized Likelihood for General Semi-Parametric Regression Models," *International Statistical Review,* 55, 245-259.

Harrison, P. J., and Stevens, C. F. (1976), "Bayesian Forecasting" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 38, 205-247.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association,* 72, 320-340.

Hinde, J. (1982), "Compound Poisson Regression Models," in *GLIM 82: Proceedings of the International Conference on Generalized Linear Models,* ed. R. Gilchrist, Berlin: Springer, pp. 109-121.

Ii, Y., and Raghunathan, T. E. (1991), "Bayesian Analysis of Series of 2 × 2 Tables," unpublished manuscript submitted to *Statistics in Medicine.*

Jorgensen, B. (1987), "Exponential Dispersion Models" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 49, 127-162.

Kackar, R. N., and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association,* 79, 853-862.

Karim, M. R., and Zeger, S. L. (1992), "Generalized Linear Models With Random Effects: Salamander Mating Revisited," *Biometrics,* 48, 631-644.

Kemp, I., Boyle, P., Smans, M., and Muir, C. (1985), *Atlas of Cancer in Scotland, 1975-1980. Incidence and Epidemiologic Perspective,* IARC Scientific Publication 72, Lyon, France: International Agency for Research on Cancer.

Kneale, G. W. (1971), "Problems Arising in Estimating From Retrospective Survey Data the Latent Period of Juvenile Cancers Initiated by Obstetric Radiography," *Biometrics,* 27, 563-590.

Laird, N. M. (1978), "Empirical Bayes Methods for Two-Way Contingency Tables," *Biometrika,* 65, 581-590.

Laird, N. M., and Louis, T. A. (1987), "Empirical Bayes Confidence Intervals Based on Bootstrap Samples," *Journal of the American Statistical Association,* 82, 739-757.

Lamotte, L. R. (1972), "Notes on the Covariance Matrix of a Random, Nested ANOVA Model," *The Annals of Mathematical Statistics,* 43, 659-662.

Liang, K. Y., and Waclawiw, M. A. (1990), "Extension of the Stein Estimating Procedure Through the Use of Estimating Functions," *Journal of the American Statistical Association,* 85, 435-440.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13-22.

Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate Regression Analyses for Categorical Data" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 54, 3-40.

Lindstrom, M. J., and Bates, D. M. (1990), "Nonlinear Mixed Effects Models for Repeated Measures Data," *Biometrics,* 46, 673-687.

Longford, N. T. (1987), "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models With Nested Random Effects," *Biometrika,* 74, 817-827.

Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P., and Pellom, A. C. (1989), "Empirical Bayes Procedures for Stabilizing Maps of U.S. Cancer Mortality Rates," *Journal of the American Statistical Association,* 84, 637-650.

Marquardt, D. W. (1963), "An Algorithm for Least Squares Estimation of Nonlinear Parameters," *SIAM Journal,* 11, 431-441.

McCullagh, P. (1984), "On the Elimination of Nuisance Parameters in the Proportional Odds Model," *Journal of the Royal Statistical Society,* Ser. B, 46, 250-256.

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

McGilchrist, C. A., and Aisbett, C. W. (1991), "Restricted BLUP for Mixed Linear Models," *Biometrical Journal,* 33, 131-141.

Moore, D. F. (1986), "Asymptotic Properties of Moment Estimators for Overdispersed Counts and Proportions," *Biometrika,* 73, 583-588.

Morton, R. (1987), "A Generalized Linear Model With Nested Strata of Extra-Poisson Variation," *Biometrika,* 74, 247-257.

——— (1988), "Analysis of Generalized Linear Models With Nested Strata of Variation," *Australian Journal of Statistics,* 30A, 215-224.

Patterson, H. D., and Thompson, R. (1974), "Recovery of Interblock Information When Block Sizes Are Unequal," *Biometrika,* 58, 545-554.

Prentice, R. L. (1988), "Correlated Binary Regression With Covariates Specific to Each Binary Observation," *Biometrics,* 44, 1033-1048.

Prosser, R., Rasbash, J., and Goldstein, H. (1991), *ML3 Software for Three-Level Analysis, Users' Guide for V. 2,* London: Institute of Education.

Ripley, B. D., and Kirkland, M. D. (1990), "Iterative Simulation Methods," *Journal of Computational and Applied Mathematics,* 31, 165-172.

Robinson, G. K. (1991), "That BLUP Is a Good Thing: The Estimation of Random Effects," *Statistical Science,* 6, 15-51.

Rotnitzky, A., and Jewell, N. P. (1990), "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data," *Biometrika,* 77, 484-497.

Schall, R. (1991), "Estimation in Generalized Linear Models With Random Effects," *Biometrika,* 40, 719-727.

SERC (1989), *EGRET Users' Manual,* Seattle: Author.

Sharples, K., and Breslow, N. (1992), "Regression Analysis of Correlated Binary Data: Some Small Sample Results for the Estimating Equation Approach," *Journal of Statistical Computation and Simulation,* 42, 1-20.

Stiratelli, R., Laird, N., and Ware, J. (1984), "Random Effects Models for Serial Observations With Binary Responses," *Biometrics,* 40, 961-971.

Thall, P. F., and Vail, S. C. (1990), "Some Covariance Models for Longitudinal Count Data With Overdispersion," *Biometrics,* 46, 657-671.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association,* 81, 82-86.

Waclawiw, M. A., and Liang, K. Y. (1991), "Prediction of Random Effects in the Generalized Linear Model," Technical Report 704, The Johns Hopkins University, Dept. of Biostatistics.

West, M., Harrison, P. J., and Migon, H. S. (1985), "Dynamic Generalized Linear Models and Bayesian Forecasting" (with discussion), *Journal of the American Statistical Association,* 80, 73–97.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics,* New York: John Wiley.

Williams, D. A. (1982), "Extra-Binomial Variation in Logistic Linear Models," *Applied Statistics,* 31, 144–148.

Zeger, S. L. (1988), "A Regression Model for Time Series of Counts," *Biometrika,* 75, 621–629.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models with Random Effects; A Gibbs Sampling Approach," *Journal of the American Statistical Association,* 86, 79–86.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics,* 44, 1049–1060.

Zelen, M. (1971), "The Analysis of Several Contingency Tables," *Biometrika,* 58, 129–137.