9

Multilevel Generalized Linear Models

Germán Rodríguez

Office of Population Research, Princeton University

9.1 Introduction

Two of the most influential papers in applied statistics published in the last few decades are Nelder and Wedderburn [65], introducing generalized linear models (GLMs), and Cox [20], the seminal paper introducing life tables with regression, better known as proportional hazard models. As we will see, these two developments are closely related. Nelder and Wedderburn's unique contribution was to provide a unified conceptual framework for studying a large range of statistical models, including not only classical linear models, but also logit and probit models for binary data, log-linear Poisson models for count data, and others. The unification was not only conceptual, but also led to common estimation procedures in the form of an iteratively re-weighted least squares (IRLS) algorithm. The first implementation of these procedures appeared in the highly successful program GLIM [3], which for many statisticians became synonymous with GLMs.

In this chapter we follow Wong and Mason [94], Longford [54, 56], Goldstein [30], Breslow and Clayton [11], and others in exploring extensions of GLMs to include random effects in a multilevel setting. Chapter 1 in this handbook has described multilevel models for continuous outcomes, while Chapter 6 has focused on multilevel models for categorical outcomes. Here we adopt a unified approach that views the general linear mixed model and many of the random-effects models for categorical data discussed in earlier chapters as special cases of the Multilevel Generalized Linear Model (MGLM). This approach has conceptual merit in emphasizing the similarities among these models, and provides a common framework to study and evaluate estimation methods. Alas, we do not have a single estimation procedure that can be applied to all MGLMs with the same measure of success that IRLS achieved

J. de Leeuw, E. Meijer (eds.), *Handbook of Multilevel Analysis*,

for GLMs. Instead, we must choose between quick but sometimes biased approximations, and more accurate but often compute-intensive maximum likelihood and Bayesian approaches. Part of our task in this chapter is to describe and illustrate the alternatives.

Section 9.2 develops the modeling framework. We introduce generalized linear models (GLMs) as an extension of linear models, and proceed to an analogous derivation of multilevel generalized linear models (MGLMs) as an extension of multilevel linear models. The ideas discussed apply more generally to generalized linear mixed models (GLMMs) and our notation reflects this broader applicability, but we tend to focus the narrative on the multilevel case. We review survival models, note their close connection with GLMs, and describe a natural extension to the multilevel case. We draw an important distinction between conditional and marginal models that is significant in the generalized linear case. Finally, we introduce non-linear mixed models and contrast them with MGLMs.

Section 9.3 is devoted to a discussion of estimation procedures. It turns out that calculation of the likelihood function for MGLMs involves intractable integrals. We discuss several alternatives and assess their performance in realistic situations, referring to some of our earlier work using simulated data and a case study [81, 82] and introducing new results. We review a range of approximate estimation procedures that, unfortunately, can be severely biased when random effects are substantial. We describe maximum likelihood estimation using Gauss-Hermite quadrature, a method that appears to work remarkably well, but is limited to relatively low-dimensional models. We also discuss Bayesian estimation procedures focusing on the Gibbs sampler, a Markov Chain Monte Carlo (MCMC) method that can be applied to more complex models involving high-dimensional integrals, albeit not without difficulty. We close this section with a brief discussion of other approaches to estimation, an active area of current research.

Section 9.4 is devoted to an application of MGLMs to the study of infant and child mortality in Kenya, using data from a national survey conducted in 1998. We use a three-level piece-wise exponential survival model that allows for clustering of infant and child deaths at both the family and community levels, and fit it to data using the equivalent MGLM with Poisson errors and log link. We compare estimates that ignore clustering, and estimates obtained by approximate quasi-likelihood and by full maximum likelihood. The discussion emphasizes interpretation of the results, particularly the family and community random parameters. Finally, we show how the model can be used to estimate measures of intra-family and intra-community correlation in infant and child deaths.

Section 9.5 is a brief discussion and summary of our conclusions.

9.2 Extending Multilevel Models

9.2.1 Generalized Linear Models

Consider briefly the general linear model. We usually view the outcome y_i for the i-th individual as a realization of a random variable (r.v.) $\underline{y}_i$ that depends on a vector $\boldsymbol{x}_i$ of predictors or explanatory variables through the equation

$$\underline{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \underline{\epsilon}_i, \tag{9.1}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $\underline{\epsilon}_i$ is an error term having a normal distribution with mean 0 and variance σ^2.

It will facilitate further generalization if we write this model in a slightly different way, noting that $\underline{y}_i$ has a normal distribution with mean μ_i and variance σ^2, which we write

$$\underline{y}_i \sim \mathcal{N}(\mu_i, \sigma^2), \tag{9.2}$$

and the expected value satisfies the linear model

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{9.3}$$

This approach draws a clear distinction between the stochastic structure of the data, specified in the first equation, and the systematic component, specified in the second.

The Exponential Family

Nelder and Wedderburn [65] generalize this model in two master strokes. First, they assume that the distribution of $\underline{y}_i$ is in an *exponential family* that includes as special cases many of the distributions we encounter in applied work, such as the normal, binomial, Poisson, gamma, and inverse Gaussian. The family may be written as

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{9.4}$$

where θ_i and ϕ are unknown parameters and $a_i(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions. Usually $a_i(\phi) = \phi/p_i$, where p_i is a known prior weight, and this will be assumed in the applications that follow. In this family, the mean is $E(\underline{y}_i) = b'(\theta_i)$ and the variance is $\mathrm{Var}(\underline{y}_i) = b''(\theta_i)a_i(\phi)$. In applied work, we often express the variance as a function of the mean, say $\mathrm{Var}(\underline{y}_i) = \phi\, V(\mu_i)$.

All the distributions mentioned above can be obtained from this general expression by suitable choice of parameters and functions. For example if we set $\theta_i = \mu_i$, $b(\theta_i) = \frac{1}{2}\theta_i^2$, $\phi = \sigma^2$, and $a_i(\phi) = \phi$, we obtain a normal

distribution with mean μ_i and variance σ^2. In this case, the variance function is the identity. The Poisson distribution with mean μ_i has $\theta_i = \log \mu_i$, $b(\theta_i) = e^{\theta_i}$, $a_i(\phi) = \phi$, and $\phi = 1$, and the variance equals the mean. McCullagh and Nelder [59] show how you can obtain other special cases from the general formula.

The Link Function

The second aspect of the generalization is that instead of modeling the expected value of the outcome as a linear function of the covariates, we model a *transformation* of the expected value. Specifically, we introduce a one-to-one continuous differentiable transformation of the mean $\eta_i = g(\mu_i)$ and assume that the transformed mean follows a linear model, so that

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{9.5}$$

The function $g(\cdot)$ is called the *link function*, and connects the mean with the linear predictor $\boldsymbol{x}_i'\boldsymbol{\beta}$ and thus the explanatory variables. The simplest possible link function is the identity, which leads to modeling the mean itself. Other transformations in common use are the logit, probit, log, inverse, and square root.

A key feature of GLMs is that the model for the transformed mean η_i is simple and has a familiar linear structure. Because the link function is one-to-one, we can always invert it to obtain a a model for the mean

$$\mu_i = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}), \tag{9.6}$$

but this model is usually more complicated. In particular, interpretation of the parameters is straightforward in the transformed scale, but may be rather involved in the original scale. Notable exceptions are models with log and logit links, where exponentiated coefficients may be interpreted as multiplicative effects on an expected count or an odds ratio, respectively. An example will follow in Section 9.4.

Link functions can often be motivated as a way to handle range restrictions on the mean μ_i. With count data, for example, a linear model is not attractive because the mean μ_i must be non-negative, but the linear predictor $\boldsymbol{x}_i'\boldsymbol{\beta}$ may yield positive or negative values. Modeling the log of the mean instead solves the problem. The link function can also make the assumption of linearity more plausible. With count data, for example, one often finds that effects are relative rather than absolute; an additive model in the log scale is equivalent to a multiplicative model in the original scale, and can thus represent relative effects. A link function that maps the mean μ_i into the parameter θ_i in the exponential family is said to be a canonical link. The canonical links for the Poisson and Bernoulli distributions are the log and the logit, respectively.

Estimation and Testing

An important practical feature of GLMs is that they can all be fit to data using the same algorithm, a form of iteratively reweighted least squares (IRLS). The algorithm may be motivated by considering a linearized form of the model, a fact that has motivated the adoption of similar strategies for MGLMs. Write the model as

$$\underline{\boldsymbol{y}} = \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}) + \underline{\boldsymbol{\epsilon}}, \tag{9.7}$$

where $\boldsymbol{\mu}(\cdot)$ is the inverse link function applied element-wise to the linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, and $\underline{\boldsymbol{\epsilon}}$ is a vector of independent heteroscedastic error terms with mean $\boldsymbol{\emptyset}$ and (diagonal) variance-covariance matrix $\phi\,\boldsymbol{V}(\boldsymbol{\mu})$. Expanding the link using a first-order Taylor series about a trial parameter value $\boldsymbol{\beta}_0$ and rearranging terms leads to the approximating linear model

$$\underline{\boldsymbol{y}}^* \approx \boldsymbol{X}\boldsymbol{\beta} + \underline{\boldsymbol{\epsilon}}^*, \tag{9.8}$$

where $\underline{\boldsymbol{y}}^* = \boldsymbol{D}^{-1}(\underline{\boldsymbol{y}} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)) + \boldsymbol{X}\boldsymbol{\beta}_0$ is a working response, $\underline{\boldsymbol{\epsilon}}^* = \boldsymbol{D}^{-1}\underline{\boldsymbol{\epsilon}}$ is a new error term with variance $\phi\boldsymbol{W}$, where $\boldsymbol{W} = \boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{\mu})\boldsymbol{D}^{-1}$ is a diagonal matrix of weights, and $\boldsymbol{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\eta}$ is a diagonal matrix of derivatives of the link function with respect to the linear predictor. This approximating linear model may be fit using weighted least squares to obtain an improved estimate of $\boldsymbol{\beta}$, which can then be used to obtain a better approximating model, and so on to convergence. McCullagh and Nelder [59] show that this method is equivalent to Fisher scoring and leads to maximum likelihood estimates.

Under standard regularity conditions, the large sample distribution of the estimator $\underline{\hat{\boldsymbol{\beta}}}$ is approximately normal with mean equal to the true parameter value $\boldsymbol{\beta}$ and variance-covariance matrix $\phi\,(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$. This result provides large-sample standard errors and a basis for Wald tests. Likelihood ratio tests are often preferable, and in the context of GLMs they are usually calculated by reference to a statistic known as the *deviance*. This statistic is constructed by considering a likelihood ratio test that compares the model of interest with a saturated model that has a separate parameter for each observation. The deviance is the product of the scale parameter ϕ and the usual likelihood ratio chi-squared statistic $-2\log\lambda$. A test comparing two nested models can then be computed as the difference of their scaled deviances.

9.2.2 Multilevel Generalized Linear Models

We now consider a similar extension for multilevel linear models. In previous chapters we have written the general linear mixed model in a form analogous to (9.1),

$$\underline{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\underline{\boldsymbol{\delta}} + \underline{\epsilon}_i, \tag{9.9}$$

where $\underline{y}_i$ is the r.v. representing the outcome for the i-th individual, $\boldsymbol{x}_i$ is the i-th row of the model matrix for the fixed effects $\boldsymbol{\beta}$, $\boldsymbol{z}_i$ is the i-th row of the model matrix for the random effects $\underline{\boldsymbol{\delta}}$, and $\underline{\epsilon}_i$ is the individual error term. We assume that the random effects $\underline{\boldsymbol{\delta}}$ have a $\mathcal{N}(\boldsymbol{\emptyset}, \boldsymbol{\Omega})$ distribution, and the error terms are independent and identically distributed (i.i.d.) $\mathcal{N}(0, \sigma^2)$ r.v.'s.

We could write this model more compactly in terms of vectors, with $\underline{\boldsymbol{\epsilon}} \sim \mathcal{N}(\boldsymbol{\emptyset}, \sigma^2\boldsymbol{I})$, but that would not be very productive for the generalization that follows. Instead, we will reformulate the model in terms of the *conditional* distribution of the outcomes $\underline{y}_i$ given the random effects $\boldsymbol{\delta}$, which we write as

$$\underline{y}_i \mid \boldsymbol{\delta} \sim \mathcal{N}(\mu_i, \sigma^2). \tag{9.10}$$

In words, we assume that given the random effects the outcomes are independent normally distributed r.v.'s with mean μ_i and variance σ^2. The conditional mean, in turn, follows the linear model

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta}, \tag{9.11}$$

depending on unknown coefficients $\boldsymbol{\beta}$ and given values $\boldsymbol{\delta}$ of the random effects. The essence of this approach is the recognition that *given* the random effects, the outcomes are independent and follow a linear model.

The stage is now set for the generalization. We retain the key assumption of conditional independence. However, instead of assuming that the conditional distribution of the outcomes $\underline{y}_i$ given the random effects $\underline{\boldsymbol{\delta}}$ is normal, we assume that the distribution is in the exponential family (9.4). This extends the general linear mixed model to situations where the conditional distribution of the responses is binomial, Poisson, gamma, or inverse Gaussian.

The second element of the generalization is the introduction of a link function. We assume that a *transformation* of the conditional mean, rather than the mean itself, follows a linear model, so that

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta}. \tag{9.12}$$

The link function can be the identity, log, logit, probit, or any other one-to-one continuous differentiable transformation. This final extension leads to multilevel logit and probit models, multilevel log-linear models for count data, and many other applications.

By focusing on the conditional distribution of the outcomes given the random effects, we can apply without changes the entire conceptual apparatus of generalized linear models. In particular, random and fixed effects can be interpreted in a unified way, the interpretation is simple in the transformed scale because fixed and random effects enter linearly, and can often be translated meaningfully back to the original scale. We will return to these issues in Section 9.4.

9.2.3 Survival Models

Let us now consider models for time-to-event or survival data, which are closely related to GLMs. There is now an extensive literature on survival models; excellent texts include Kalbfleisch and Prentice [43], Cox and Oakes [22], and Therneau and Grambsch [89].

Hazards and Survival

In a standard hazard model, we assume that the survival experiences of different individuals are independent and that the hazard for individual i, or instantaneous risk of occurrence of the event at time t given that it has not occurred earlier, is given by

$$\lambda(t, \boldsymbol{x}_i) = \lambda_0(t) \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}, \tag{9.13}$$

where $\lambda_0(t)$ represents a baseline hazard at time t and $\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}$ is a relative risk associated with covariate values $\boldsymbol{x}$. The special case where $\lambda_0(t) = \lambda_0$ is the exponential survival model of Feigl and Zelen [26]. The model is easily extended to time-varying covariates $\boldsymbol{x}(t)$ and time-varying effects $\boldsymbol{\beta}(t)$. Note that taking logs yields a model that is linear in the relative risk parameters.

The cumulative hazard is defined as $\Lambda(t, \boldsymbol{x}_i) = \int_0^t \lambda(u, \boldsymbol{x}_i)\, \mathrm{d}u$, which for time-fixed covariates is simply the baseline cumulative hazard times the relative risk for individual i. We will also need the survival function or probability of being alive at time t, which can be obtained from the cumulative hazard as $S(t \mid \boldsymbol{x}_i) = \exp\{-\Lambda(t, \boldsymbol{x}_i)\}$, and therefore for fixed covariates satisfies

$$S(t \mid \boldsymbol{x}_i) = S_0(t)^{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}, \tag{9.14}$$

where $S_0(t)$ is the baseline survival function.

Estimation with Censored Data

A distinctive feature of survival models is that observations are often *censored*, in the sense that for some individuals the event of interest has not yet occurred at the time the data are analyzed. Estimation of censored-data hazard models under parametric assumptions for the baseline hazard relies on the standard survival likelihood, to which an individual who dies at t contributes $\lambda(t, \boldsymbol{x})S(t, \boldsymbol{x})$, the density at t, and an individual who is censored at t contributes $S(t, \boldsymbol{x})$, the probability of surviving to t. This likelihood can be derived under the key assumption that censoring is non-informative, so all we know about an individual who is censored at t is that it survived that long; see Kalbfleisch and Prentice [43]. Cox [20, 21] introduced a partial likelihood

that allows estimation of the relative risk coefficients $\boldsymbol{\beta}$ without assumptions about the shape of the baseline hazard $\lambda_0(t)$.

Several authors have noted a close relationship between hazard models and GLMs, and a number of papers show how various survival models can be fit using standard GLM software; see Aitkin and Clayton [4] for the exponential, Weibull, and extreme value distributions, Bennet and Whitehead [8] for the logistic and log-logistic, and Clayton and Cuzick [17] and Whitehead [93] for estimation using Cox's partial likelihood. In this section we focus on connections with Poisson models with log link, which we use in our application in Section 9.4, and binomial models with logit and c-log-log links.

Piece-Wise Exponential Survival

A flexible semi-parametric approach to hazard models is to partition time (or duration of exposure) into J intervals $[\tau_{j-1}, \tau_j)$ for $j = 1, \ldots, J$ with cutpoints $0 = \tau_0 < \tau_1 < \cdots < \tau_J$, and assume that the baseline hazard is constant within each interval, so that $\lambda_0(t) = \lambda_{0j}$ for $t \in [\tau_{j-1}, \tau_j)$. Judicious choice of cutpoints leads to good approximations to a wide range of hazard functions, using more closely spaced boundaries where the hazard varies rapidly and wider intervals where the hazard changes more slowly.

Holford [40] and Laird and Olivier [47] noted that the piece-wise exponential model is equivalent to a Poisson regression model. With censored data, we observe t_i, the total time lived by the i-th individual, and d_i, a death indicator that takes the value 1 if the individual died and 0 otherwise. Imagine defining analogous measures for each duration interval, so t_{ij} is the time lived by the i-th individual in the j-th interval, and d_{ij} is a death indicator that takes the value 1 if individual i died in interval j and 0 otherwise. Then a piece-wise exponential hazard model can be fitted by treating the death indicators d_{ij} as if they were independent Poisson observations with means $\mu_{ij} = \lambda_{ij} t_{ij}$, where λ_{ij} is the hazard for individual i in interval j.

The proof is not hard and can be sketched as follows. The contribution of the i-th individual to the standard survival log-likelihood for censored data has the form $d_i \log \lambda(t_i, \boldsymbol{x}_i) - \Lambda(t_i, \boldsymbol{x}_i)$. Suppose t_i falls in interval $j(i)$ and write $\lambda_{ij(i)}$ as shorthand for $\lambda(t_i, \boldsymbol{x}_i)$. The cumulative or integrated hazard can be computed easily because the hazard is constant in each interval, so $\Lambda(t_i, \boldsymbol{x}_i) = \sum_j \lambda_{ij} t_{ij}$, where the sum is over all intervals up to $j(i)$. There is a slight lack of symmetry in that we have only one term on the death indicator and $j(i)$ terms on the exposure times, but we can easily add the terms for previous intervals, which have $d_{ij} = 0$ and thus are all zero, to obtain

$$\log L_i = \sum_{j=1}^{j(i)} \{ d_{ij} \log \lambda_{ij} - \lambda_{ij} t_{ij} \}. \tag{9.15}$$

This equation coincides with the log-likelihood that we would obtain if we treated d_{ij} as having a Poisson distribution with mean $\mu_{ij} = \lambda_{ij} t_{ij}$ except for a term $d_{ij} \log(t_{ij})$, but this is a constant depending on the data and not the parameters, so it can be ignored.

It is important to note that we have not assumed that the d_{ij} have independent Poisson distributions, because clearly they do not. If individual i died in interval j, then it must have been alive in all prior intervals, so the indicators couldn't possibly be independent. Moreover, each indicator can only take the values 1 and 0, so it couldn't possibly have a Poisson distribution that assigns probability to values greater than 1. The result is more subtle; it is the likelihood functions that coincide. Given a realization of a piece-wise exponential process, we can find a realization of a set of independent Poisson r.v.'s that happens to have the same probability and thus leads to the same estimates. The practical implication is that one can fit a piece-wise exponential model in terms of the equivalent GLM.

Discrete Survival Models

In his original paper, Cox [20] proposed a discrete version of the proportional hazards model by working with the conditional odds of dying at each possible failure time t_j given survival up to that point. Specifically, he proposed the model

$$\frac{\lambda(t_j \mid \boldsymbol{x})}{1 - \lambda(t_j \mid \boldsymbol{x})} = \frac{\lambda_0(t_j)}{1 - \lambda_0(t_j)} \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}, \tag{9.16}$$

where $\lambda_0(t_j)$ is the baseline conditional probability of dying at t_j given survival to that time and $\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}$ is the relative risk. In this model, the conditional log-odds of dying are linear in the relative risk parameters $\boldsymbol{\beta}$.

Cox [20] extended his partial likelihood approach to estimate $\boldsymbol{\beta}$ while treating the baseline hazards $\lambda_0(t_j)$ as nuisance parameters that could be conditioned out of the likelihood. Allison [5] noted that one could estimate the complete model, including a separate parameter for each discrete time of death t_j, by running a logistic regression on a set of pseudo-observations, in a procedure analogous to that described above for piece-wise exponential models.

An alternative extension of hazard models to discrete data assumes that the survival functions satisfy (9.14) and then solves for the conditional hazard at time t_j, to obtain

$$\lambda(t_j \mid \boldsymbol{x}) = 1 - (1 - \lambda_0(t_j))^{\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}}. \tag{9.17}$$

The transformation that makes the right-hand side a linear function of the parameters is the complementary log-log, and the model can be fitted using a GLM with binomial structure and complementary log-log link.

This model can also be obtained by grouping time in a continuous-time proportional hazards model, see Prentice and Gloeckler [72] and Kalbfleisch and Prentice [43] for details. In this approach, time is grouped into intervals $[\tau_{j-1}, \tau_j)$ as before, but all we observe is whether an individual survives or dies in an interval. This construction imposes some constraints on censoring: If an individual is censored inside an interval, we do not know whether he or she would have survived the interval, and therefore must censor the observation back at the beginning of the interval. Unlike the piece-wise exponential setup, we cannot use information about exposure during part of an interval. On the other hand, we do not need to assume that the hazard is constant in each interval.

9.2.4 Multilevel Survival Models

In the last several years, there has been considerable interest in extending survival models by introducing random effects. A classic demographic contribution is Vaupel et al. [91], which introduced a gamma-distributed random effect to represent unobserved heterogeneity of frailty in univariate survival models, see also Aalen [1], Hougaard [41, 42], and Manton et al. [58]. The idea of frailty can be used to represent association of kindred lifetimes in a multivariate setting, see Clayton [14], Clayton and Cuzick [18], and Oakes [68]; to account for association in recurrent events and event history data, see Clayton [15] and Rodríguez [79]; and leads naturally to two- and three-level survival models, see Guo and Rodríguez [37], Sastry [83], and Barber et al. [7].

The multilevel extension follows the same strategy as for MGLMs. We assume that given a vector of random effects $\underline{\boldsymbol{\delta}}$, the survival experiences of different individuals are independent and follow a hazard model with conditional hazard

$$\lambda(t, \boldsymbol{x}_i \mid \boldsymbol{\delta}) = \lambda_0(t) \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta}\}. \tag{9.18}$$

In this generalization, the hazard for individual i depends not only on the fixed effects $\boldsymbol{\beta}$ with model vector $\boldsymbol{x}_i$, but also on the random effects $\boldsymbol{\delta}$ with model vector $\boldsymbol{z}_i$. Once again, the random effects enter a linear predictor in exactly the same form as the fixed effects. Calculation of the conditional cumulative hazard and the conditional survival function follows along the same lines as in ordinary survival models.

We can also calculate unconditional or marginal survival probabilities by integrating out the random effects. Calculation of unconditional hazards requires special care because hazards, by definition, are conditional on survival to time t. The extent of dependence of kindred lifetimes can be expressed in terms of measures of intraclass correlation. Estimation of both discrete- and continuous-time multilevel survival models can proceed by working in terms

of the equivalent MGLM with binomial or Poisson errors. We will revisit these issues in the context of our application in Section 9.4.

9.2.5 Conditional and Marginal Models

An alternative approach to the analysis of correlated data that is popular in longitudinal or repeated-measurement studies focuses on the marginal distribution of the responses, see Diggle et al. [25, Chapter 8]. These models assume that the outcomes have a distribution in the exponential family, and that a transformation of the *marginal* mean is a linear function of observed covariates with coefficients $\boldsymbol{\beta}$. The models are usually fit to data using generalized estimating equations (GEE) that take into account the dependence of the observations. The method is very similar to the IRLS algorithm used in GLMs, using the same working dependent variable and the same set of weights, but instead of using weighted least squares (WLS) with a diagonal weight matrix, it uses generalized least squares (GLS) with a more general weight matrix, where the non-diagonal elements reflect the correlation structure of the observations.

In the linear case, the marginal and conditional models coincide, in the sense that in both instances the mean is a linear function of the covariates with the same coefficients $\boldsymbol{\beta}$. This is no longer true in the more general case; except for variance-component probit models, where the conditional and marginal models differ only by a scaling of the coefficients, the two approaches lead to different models. The distinction is particularly important in the case of survival models, where it can give rise to interesting paradoxes, see Vaupel and Yashin [92]. Marginal models are useful when one is interested in making inferences about population averages, whereas conditional models have a subject-specific interpretation, see Neuhaus et al. [66] for a comparison. As will be shown in our application, one can always use a conditional model to compute marginal quantities of interest, so in this sense the MGLM approach is richer, see also Goldstein [31, 32].

9.2.6 Non-Linear Models

Generalized linear models and the extensions considered so far expand the statistician's toolkit beyond the assumption of normally distributed outcomes, while retaining the assumption of linear effects on a transformed scale. Non-linear models are different; they retain the assumption of normally distributed outcomes, but move beyond the assumption of linear effects to consider more general structures where the parameters enter non-linearly. These models often have a natural physical interpretation, may be more parsimonious than linear models, and can provide more reliable predictions outside the observed range of the data. Needless to say, non-linear models have also been extended

to include random effects at various levels of aggregation, see Davidian and Giltinan [24] and Pinheiro and Bates [71, Part II]. In this chapter we focus on MGLMs, but note that the two approaches share common estimation problems and have adopted similar solutions.

9.3 Approaches to Estimation

Estimation of multilevel linear models for normally distributed outcomes using maximum likelihood or restricted maximum likelihood is very well understood. Excellent implementations are available in specialized multilevel packages, namely HLM and MLwiN, as well as in general-purpose statistical packages, including Stata, SAS, and R/S-Plus. When it comes to MGLMs, however, the picture gets more complicated.

Estimation by maximum likelihood requires the marginal distribution of the responses. We assume that the random effects have density $g(\boldsymbol{\delta})$, a multivariate normal density with a patterned covariance structure. We further assume that the conditional density of the outcomes given the random effects, $f(\boldsymbol{y} \mid \boldsymbol{\delta})$, is a product of densities in the exponential family. The product of the marginal and conditional densities gives us the joint density of the outcomes and the random effects. Calculation of the marginal density of the outcome is then a "simple matter" of integrating out the random effects:

$$f(\boldsymbol{y}) = \int f(\boldsymbol{y} \mid \boldsymbol{\delta})\, g(\boldsymbol{\delta})\, \mathrm{d}\boldsymbol{\delta}. \tag{9.19}$$

Unfortunately, this integral is intractable, with no general closed-form solution.

There are some special cases of interest with a single random effect whose distribution is conjugate with the distribution of the outcome, see Lee and Nelder [49] for a general approach. For example, if $\underline{y}$ and $\underline{\delta}$ are scalars, the marginal distribution of the random effect is gamma and the conditional distribution of the outcome given the random effect is Poisson, then the marginal distribution of the outcome is negative binomial, see Lawless [48]. For binary outcomes the beta-binomial combination is popular, see Crowder [23]. A difficulty with these approaches is that they do not extend easily to models involving multiple dependent random effects. The flexibility of the assumption of multivariate normality for the random effects is, in fact, unmatched. To retain this flexibility, we need a way to get around the intractability of (9.19). We now turn to a discussion of the three main approaches to estimation in current use, starting with a simulation study used to evaluate them.

9.3.1 A Simulation Study

To assess the performance of alternative estimation procedures, we will use data from a simulation study described in Rodríguez and Goldman [81]. The study was motivated by work on health care utilization in Guatemala, where exploratory analyses had suggested large family and community effects on the use of modern health care, yet more formal analyses using multilevel models, as implemented in then current software, had failed to confirm the existence of large effects. To resolve this disparity, we ran a number of simulations, using what we then considered small and large variance components. Subsequent work revealed that the actual effects were in fact much larger than the values used in our simulations, see Pebley et al. [69] and the case study in Rodríguez and Goldman [82].

We will focus here on a set of simulations using the actual structure of Guatemalan data on prenatal care, with 2449 births to 1558 mothers who were living in 161 communities. We created three composite explanatory variables summarizing characteristics of the pregnancy, mother, and community, and set their fixed-effect coefficients to 1. We added random effects representing unobserved characteristics of the mother and community, sampled from normal distributions with mean 0 and variance 1. Finally, we simulated a binary response following a 3-level random-intercept logit model. This procedure was used to generate 100 datasets that have been used by several authors and are freely available at `http://data.princeton.edu/multilevel`.

Table 9.1 summarizes the results of trying various estimation procedures on these datasets. The results for MQL-1 and MQL-2 appeared in Rodríguez and Goldman [81]. Goldstein and Rasbash [34] reported results for PQL-2 using the first 25 of our 100 datasets; we have extended the analysis to cover all 100 and added PQL-1. The results using quadrature methods and the Gibbs sampler are new. We will comment on these results as we describe the various procedures. For brevity, we omit presentation and discussion of standard errors.

Browne and Draper [13] have also analyzed the first 25 of our datasets, and went on to generate a further 500 samples with the same multilevel structure, as part of an interesting simulation study contrasting Bayesian and likelihood-based procedures. The comparison includes MQL and PQL as well as a Bayesian approach, but excludes maximum likelihood via quadrature procedures. Their implementation of Bayesian estimation combines the Metropolis algorithm with Gibbs sampling and tries two choices of diffuse priors for the variances of the random effects. The evaluation criteria include the bias of point estimates and also the coverage rates of interval estimates. Their results parallel ours and lead to essentially the same conclusions regarding the relative merits of these methods.

Table 9.1 Estimates for simulated data using the Guatemala structure.

Estimation	Fixed Parameters (β)			Random Parameters (σ)	
Method	Individual	Family	Community	Family	Community
True Value	1.000	1.000	1.000	1.000	1.000
MQL-1	0.738	0.744	0.771	0.100	0.732
MQL-2	0.853	0.859	0.909	0.273	0.763
PQL-1	0.808	0.806	0.831	0.432	0.781
PQL-2	0.933	0.940	0.993	0.732	0.924
ML-5	0.983	0.988	1.037	0.962	0.981
ML-20	0.983	0.990	1.039	0.973	0.979
Gibbs	0.971	0.978	1.022	0.922	0.953

9.3.2 Marginal and Penalized Quasi-Likelihood

Goldstein [30] and collaborators have proposed a general approach to the estimation of MLGMs that relies on a linearization strategy, and has led to four different approximations, known as first- and second-order maximum quasi-likelihood (MQL) and penalized quasi-likelihood (PQL).

MQL-1

To motivate these approximations, we write the MLGM model as

$$\underline{\boldsymbol{y}} = \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\underline{\boldsymbol{\delta}}) + \underline{\boldsymbol{\epsilon}}, \tag{9.20}$$

where $\underline{\boldsymbol{\epsilon}}$ is a heteroscedastic error term with mean $\underline{\boldsymbol{\emptyset}}$ and variance $\boldsymbol{V}(\boldsymbol{\mu})$ depending on the mean. Goldstein [30] approximates the inverse link $\boldsymbol{\mu}(\boldsymbol{\eta})$ using a first-order Taylor series expansion around trial values $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\underline{\boldsymbol{\delta}} = \underline{\boldsymbol{\emptyset}}$, to obtain

$$\underline{\boldsymbol{y}} = \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0) + \boldsymbol{D}\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \boldsymbol{D}\boldsymbol{Z}\underline{\boldsymbol{\delta}} + \underline{\boldsymbol{\epsilon}}, \tag{9.21}$$

where $\boldsymbol{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\eta}_0$ is a diagonal matrix of derivatives of the mean with respect to the linear predictor evaluated at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$. Pre-multiplying both sides of the equation by $\boldsymbol{D}^{-1}$ and rearranging terms gives

$$\underline{\boldsymbol{y}}^* = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\underline{\boldsymbol{\delta}} + \underline{\boldsymbol{\epsilon}}^*, \tag{9.22}$$

where $\underline{\boldsymbol{y}}^* = \boldsymbol{D}^{-1}(\underline{\boldsymbol{y}} - \boldsymbol{\mu}_0) + \boldsymbol{X}\boldsymbol{\beta}_0$ and $\underline{\boldsymbol{\epsilon}}^*$ is an error term with mean $\underline{\boldsymbol{\emptyset}}$ and variance $\boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{\mu})\boldsymbol{D}^{-1}$. (The variance is simpler for logit and other models where the derivative of the link $\boldsymbol{D}$ coincides with the variance function $\boldsymbol{V}(\boldsymbol{\mu})$.)

Equation (9.22) has the structure of a linear mixed model, with mean $E(\underline{\boldsymbol{y}}^*) = \boldsymbol{X}\boldsymbol{\beta}$ and variance

$$\mathrm{Var}(\underline{\boldsymbol{y}}^*) = \boldsymbol{Z}\boldsymbol{\Omega}\boldsymbol{Z}' + \boldsymbol{D}^{-1}\boldsymbol{V}(\boldsymbol{\mu}_0)\boldsymbol{D}^{-1}, \tag{9.23}$$

which has been evaluated at μ_0. Fitting this model by ML or REML leads to an improved estimate of the fixed effects $\boldsymbol{\beta}$, which can then be used to compute a new approximating model. The procedure is iterated to convergence. This method is termed maximum quasi-likelihood (MQL) because the approximating linear mixed model matches the mean and variance of the target model. Interestingly, if there are no random effects, the method coincides exactly with the IRLS algorithm used in GLMs and therefore leads to maximum likelihood estimates.

Longford [54, 56] adopted a different approach that, somewhat surprisingly, leads to an equivalent algorithm. He approximates the conditional likelihood $f(\boldsymbol{y} \mid \boldsymbol{\delta})$ using a second-order Taylor series expansion about $\boldsymbol{\delta} = \boldsymbol{\emptyset}$. The random effects appear in this expansion only in a quadratic form, which can be combined with a similar quadratic form in the marginal density $g(\boldsymbol{\delta})$ of the random effects to carry out the required integration analytically. Longford goes on to derive a Fisher scoring algorithm that provides estimates of both fixed and random effects. This strategy was first implemented in the multilevel package VARCL [55], and turns out to be exactly equivalent to Goldstein's MQL-1 procedure. For further details, see Rodríguez and Goldman [81].

Unfortunately, the results in Table 9.1 show that first-order MQL estimates can be biased, underestimating the fixed effects (β's) by 23–26% and the random parameters (σ's) by 27% at the community and 90% at the family level. For related results, see Breslow and Clayton [11] and Breslow and Lin [12].

MQL-2

Goldstein [30, p. 50] also proposed a quadratic approximation based on a second-order Taylor series expansion. Specifically, he adds the second-order terms corresponding to each of the random effects in the model, but omits second-order terms on the fixed effects as well as mixed derivatives. The resulting squared terms are treated as additional random effects in the approximating linear model. Because these are really not separate terms, their means and variances are not estimated, but rather are calculated from the variances of the original random effects under the assumption of normality. The resulting constrained model is easily fit using MLwiN. We refer to this approximation as MQL-2.

Experience suggests that this method is more accurate than MQL-1, although it doesn't always converge. Table 9.1 shows that the bias is reduced to 9–15% for the fixed parameters, and 24% and 73% for the community and family random parameters, respectively; a notable improvement, although substantial bias remains.

PQL-1

Simulations show that MQL-1 and MQL-2 work better when the random effects are small, i.e., their variances are close to zero. This fact should not be surprising considering that the approximation is based on a Taylor series expansion about $\boldsymbol{\delta} = \boldsymbol{\emptyset}$. An alternative procedure would be to expand about $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ with a non-zero pivot, and the obvious candidate is the empirical Bayes estimate of the random effects, defined as $E(\underline{\boldsymbol{\delta}} \mid \boldsymbol{y})$, evaluated at current parameter values. The expansion then becomes

$$\underline{\boldsymbol{y}} = \boldsymbol{\mu}(\boldsymbol{\eta}_0) + \boldsymbol{DX}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \boldsymbol{DZ}(\underline{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) + \underline{\boldsymbol{\epsilon}}, \tag{9.24}$$

and leads to an approximating multilevel linear model with the same form as (9.22), except that the working response is now $\underline{\boldsymbol{y}}^* = \boldsymbol{D}^{-1}(\underline{\boldsymbol{y}} - \boldsymbol{\mu}_0) + \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\delta}_0$. This model can be estimated using ML or REML, and the resulting estimates of both fixed and random effects are used to obtain a new approximating model. The procedure is then iterated to convergence.

The same procedure has been derived by other authors using different approaches. Laird [46] and Stiratelli et al. [87] derive it from a Bayesian perspective as an approximation to a posterior distribution using a diffuse prior. Schall [84] starts from a MGLM and uses a linearized form of the link function applied to the data. Breslow and Clayton [11] derive the procedure using Laplace's method for integral approximation, and term it penalized quasi-likelihood or PQL by relating it to results of Green [36].

Our experience suggests that PQL-1 tends to perform better than MQL-1, is sometimes competitive with MQL-2, and is more likely to converge. For the simulated data, the PQL-1 estimates of the fixed effects are not quite as good as MQL-2, but the estimates of the random parameters are better, although the family standard deviation is still seriously biased.

PQL-2

Goldstein and Rasbash [34] have proposed an improved version of PQL, termed PQL-2, that extends the Taylor series to include second-order terms on the random effects, but no second-order terms on the fixed effects and no mixed derivatives. The resulting squared terms are treated exactly the same way as in MQL-2, as additional random effects whose variance is not estimated but rather calculated from the other parameters.

We have found PQL-2 to be the most accurate method in this series, although sometimes it fails to converge. The results in Table 9.1 show that PQL-2 has only a 1–7% bias for the fixed parameters, and underestimates the community random parameter by 8%, although there is still a 27% bias in the estimation of the family random parameter.

Bootstrapping

One way to reduce the bias in the approximate estimation procedures is by bootstrapping, see Kuk [45] and Goldstein [33], and the detailed discussion in Chapter 11. We used MLwiN to bootstrap MQL-1 and PQL-1 estimates in a case study involving three-level random-intercept logit models [82]. We found that the procedure was successful in correcting the bias of the estimates of both fixed and random parameters. However, the technique is extremely compute-intensive (more so than the MCMC methods discussed below), taking days to converge in one of our datasets and failing after 400 replicates in another. In both cases, however, we noted that the first few iterations achieved large bias corrections, suggesting that one could run a few bootstrap iterations as a diagnostic technique. For more details, see Rodríguez and Goldman [82, Fig. 3].

9.3.3 Gauss-Hermite Quadrature

A second approach to estimation of MGLMs is to calculate the integral (9.19) representing the marginal likelihood using numerical quadrature procedures. Previous work along these lines includes Anderson and Aitkin [6] and Hedeker and Gibbons [38, 39], see also Chapter 6 in this handbook. For an excellent introduction to numerical integration methods with applications to statistics, see Thisted [90, Chapter 5].

Table 9.1 shows the results of computing maximum likelihood estimates for our simulated data using 5-point and 20-point Gauss-Hermite quadrature. We find no evidence of bias in the estimation of the fixed effects, and only about a 2% bias in the estimation of the random parameters, well within the margin of error of our simulations. We now describe the method in some detail.

Quadrature Rules

Quadrature methods approximate an integral as a weighted sum of function values evaluated over a grid of points, so that

$$\int f(x)\ \mathrm{d}x \approx \sum_q w_q f(x_q). \tag{9.25}$$

Simple methods, such as the trapezoidal rule and Simpson's rule, evaluate the integral at equally spaced points and can integrate certain polynomials exactly; in general, k points lead to exact integration of polynomials of degree $k-1$ with appropriate choice of weights.

Gaussian quadrature rules choose not only the weights, but also the evaluation points or abscissæ, and can achieve higher precision with a fixed number

of points. In particular, Gauss-Hermite quadrature (so called because the evaluation points are zeroes of the Hermite polynomials) can be used with integrals of the form $\int f(x)e^{-x^2}\,\mathrm{d}x$, and works best when $f(x)$ can be well approximated by a polynomial. The abscissæ and weights for this rule may be found in Abramowitz and Stegun [2] or may be computed using the function `gauher` in Press et al. [73].

In our applications, we need to evaluate integrals of the form $\int f(z)\,\phi(z)\,\mathrm{d}z$, where $\phi(\cdot)$ is the standard normal density. A simple change of variables leads to the approximation $\sum w_q f(z_q)$, where w_q is the Gauss-Hermite weight divided by $\sqrt{\pi}$ and z_q is the Gauss-Hermite abscissa times $\sqrt{2}$.

Two-Level Likelihood

Consider a two-level random-intercept model with n_j observations in cluster j. Let $\underline{\delta}_j \sim \mathcal{N}(0, \sigma^2)$ denote the cluster effect. We assume that given δ_j the n_j observations are independent and have a distribution in the exponential family $f(y_{ij} \mid \delta_j)$. We further assume that the conditional mean $E(\underline{y}_{ij} \mid \delta_j) = \mu_{ij}$ satisfies a generalized linear model with $g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}+\delta_j$. We write $\underline{\delta}_j = \sigma\underline{z}_j$, so we only need to consider standard normal random effects.

Let $L_j(z_j) = \prod_i f(y_{ij} \mid z_j)$ denote the conditional likelihood for cluster j given the random effect. We can evaluate the marginal likelihood for the cluster using Q-point Gauss-Hermite quadrature as a simple weighted average

$$L_j = \sum_{q=1}^{Q} w_q L_{jq}, \tag{9.26}$$

where we have written L_{jq} as shorthand for $L_j(z_q)$, the likelihood for cluster j evaluated at the q-th quadrature point.

Two-Level Score

First and second derivatives of the likelihood can also be evaluated as weighted averages, but we usually work with the log-likelihood instead. Let $\boldsymbol{\theta}$ denote the model parameters, including $\boldsymbol{\beta}$ and σ (or better still $\log\sigma$, which avoids range restrictions and is usually better behaved).

Let $\boldsymbol{u}_j = \partial \log L_j/\partial\boldsymbol{\theta}$ denote the score vector for cluster j. Simple calculus shows that

$$\boldsymbol{u}_j = \sum_{q=1}^{Q} w^*_{jq}\boldsymbol{u}_{jq}, \tag{9.27}$$

where $\boldsymbol{u}_{jq}$ is the score corresponding to the log-likelihood for cluster j evaluated at the q-th quadrature point, and $w^*_{jq} = w_q L_{jq}/L_j$.

The new weight w^*_{jq} has an interesting interpretation. One can view the approximate likelihood (9.26) as a mixture model where cluster j comes from one of Q discrete classes with random effects z_q and prior probabilities w_q. The new weight w^*_{jq} is the posterior probability that the cluster came from class q given the data $\boldsymbol{y}_j$. Thus, the quadrature score is the posterior average of the scores evaluated at the quadrature points.

Two-Level Hessian

Let $\boldsymbol{H}_j = \partial^2 \log L_j / \partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$ denote the Hessian or matrix of second derivatives of the log-likelihood for cluster j. It can be shown that this matrix satisfies

$$\boldsymbol{H}_j = \sum_{q=1}^{Q} w^*_{jq}\boldsymbol{H}_{jq} + \sum_{q=1}^{Q} w^*_{jq}(\boldsymbol{u}_{jq} - \boldsymbol{u}_j)(\boldsymbol{u}_{jq} - \boldsymbol{u}_j)', \tag{9.28}$$

where $\boldsymbol{H}_{jq}$ is the Hessian for cluster j evaluated at the q-th quadrature point. Thus, the Hessian is the posterior average of the Hessians evaluated at the quadrature points plus the variance of the scores evaluated at the quadrature points.

This equation is formally identical to a well-known result for maximum likelihood estimation using the EM algorithm, which views the random effects $\underline{\delta}_j$ as missing data, and shows that the incomplete data information equals the expected complete data information minus the variance of the scores, which represents the missing information; see Louis [57].

Adaptive Quadrature

Liu and Pierce [53] proposed an extension of Gauss-Hermite quadrature where the variable of integration is transformed so the integrand is sampled in a more appropriate region. The starting point is the observation that the integrand in (9.19) is the product of the prior density of the random effect and the density of the data given the random effect, and is therefore proportional to the posterior distribution of the random effect. This, in turn, can be approximated using a Gaussian density. To fix ideas, consider a two-level variance-components model where the random effect has a $\mathcal{N}(0, \sigma^2)$ prior and write the contribution of a cluster to the likelihood as

$$\int f(\boldsymbol{y} \mid \delta)\,\phi(\delta; 0, \sigma^2)\,\mathrm{d}\delta = \int \left\{ \frac{f(\boldsymbol{y} \mid \delta)\phi(\delta; 0, \sigma^2)}{\phi(\delta; \mu, \gamma^2)} \right\} \phi(\delta; \mu, \gamma^2)\,\mathrm{d}\delta, \tag{9.29}$$

where $\phi(\delta; \mu, \gamma^2)$ denotes the normal density with mean μ and variance γ^2.

Liu and Pierce [53] choose μ and γ^2 to match the mode and the curvature at the mode of the posterior density. The integral on the right-hand side is

then evaluated using Gaussian quadrature, following a change of variables from δ to $(\delta - \mu)/\gamma$. This has the effect of sampling the integrand in a more relevant range, and improves accuracy as long as the ratio in braces is better approximated by a low-order polynomial than the likelihood. The method with a single node is equivalent to the Laplace (or PQL-1) approximation to the integral, so this approach may be viewed as an extension of Laplace approximation.

Pinheiro and Bates [70] derived this algorithm, which they termed *adaptive* Gaussian quadrature, from an interesting perspective. They viewed Gaussian quadrature as a deterministic version of Monte Carlo integration and proposed adaptive quadrature as a deterministic version of importance sampling, which tends to be much more efficient than simple Monte Carlo integration, using a Gaussian density with the same mode and curvature as the posterior density as the importance distribution.

Rabe-Hesketh et al. [74] proposed a slightly different approach that simplifies the calculations required to place the nodes; instead of matching the mode and curvature, they use the posterior mean and variance of the random effects, which are calculated by building on work of Naylor and Smith [64]. Their approach, embodied in the `gllamm` command in Stata, was the first implementation of adaptive quadrature for multilevel models, and has now replaced Gauss-Hermite quadrature in other Stata procedures, including the official commands for random effects logit, probit, and Poisson models. Another implementation of adaptive methods may be found in R's `lme4`.

Although adaptive quadrature requires additional computational effort to place the abscissæ, it usually pays off by requiring many fewer quadrature points. In our original analysis of Guatemalan data reported in Pebley et al. [69], we used Gauss-Hermite quadrature with 20 nodes at each level, so each likelihood evaluation required going over a 400-point grid. Recently, we were able to replicate the results exactly using `gllamm` with the default 6 points per level. The `gllamm` code is slow because it is interpreted, but speed has improved as critical parts of the algorithm have been converted to internal code in Stata. For further details, see Rabe-Hesketh et al. [74].

Extension to More Dimensions

So far we have discussed a two-level model with a single random effect, but the quadrature approach can be extended to higher-dimensional models. Consider first a three-level random-intercept model. Because there is only one random effect at each level, the model can be estimated by recursive application of the method described so far. Specifically, the likelihood for a level-3 unit is computed as a weighted sum of level-3 likelihoods evaluated at the quadrature points. These are products of level-2 likelihoods, each computed using (9.26).

Consider next a two-level random-slope model where we have two random coefficients, say $\underline{\alpha}_j = \alpha + \underline{\delta}_{1j}$ and $\underline{\beta}_j = \beta + \underline{\delta}_{2j}$. Fitting this model requires evaluating a bivariate normal integral, but we can always transform to independence; in the simplest case by using the marginal distribution of $\underline{\delta}_{1j}$ and the conditional distribution of $\underline{\delta}_{2j} \mid \delta_{1j}$, which can then be standardized. Extension to higher-dimensional models follows along similar lines using a Cholesky decomposition.

Optimization Algorithms

The foregoing results can be used in a Newton-Raphson algorithm for maximizing the log-likelihood function. Our experience using the built-in function minimizers in S-Plus and R, as well as code in Press et al. [73], suggests that the extra expense of computing second derivatives is not always worthwhile. Instead, we provide first derivatives only, letting the algorithms compute numerical second derivatives, or use variable-metric methods such as DFP or BFGS that build an approximation to the Hessian in the course of iteration. However, we do use analytic results to evaluate the Hessian after convergence, in order to obtain more accurate standard errors.

The first statistical package to incorporate quadrature methods was Egret [19]. The latest version of Stata can fit two-level random-intercept logit and probit models using adaptive quadrature, and has a nice provision for checking the procedure by comparing results with different numbers of points. A more general implementation of quadrature methods may be found in the package aML [51], which can handle, at least in principle, several levels and multiple random effects.

The computational burden of Gauss-Hermite quadrature increases rapidly with the dimensionality of the problem. For an m-dimensional model using Q quadrature points for each random effect, each evaluation of the likelihood function is equivalent to Q^m evaluations of a GLM likelihood. Using 12 quadrature points, which seems a reasonable standard for general use, one can easily fit three-level random-intercept models and two-level models with two random coefficients, say an intercept and a slope, with each likelihood evaluation the equivalent of 144 GLM likelihoods. But using 12-point quadrature to evaluate the likelihood of a three-level model with two random coefficients at each level is equivalent to evaluating almost 21,000 GLM likelihoods. Obviously, the technique works best for relatively low-dimensional models.

9.3.4 Bayesian Estimation Using the Gibbs Sampler

Recent advances in Bayesian estimation avoid the need for numerical integration by taking repeated samples from the posterior distribution of the parameters of interest. In particular, use of the Gibbs sampler in the context

of MGLMs was first proposed by Zeger and Karim [95], and has been discussed in greater detail by Clayton [16]. See also Chapter 2 in this Handbook and the Browne and Draper [13] evaluation cited earlier.

Gibbs Sampling

To apply this framework, we adopt a Bayesian perspective, treating all parameters as random variables and assigning prior (or hyperprior) distributions to the fixed-effect parameters $\underline{\boldsymbol{\beta}}$ and to the precisions $\underline{\boldsymbol{\tau}}$ (the reciprocals of the variances) of the random effects. To obtain Bayesian estimates that are roughly comparable to maximum likelihood estimates, many analysts use vague or non-informative priors. Fixed effects are typically assumed to come from normal distributions with mean zero and very large variances, and precisions are sampled from diffuse gamma or Pareto distributions, see Spiegelhalter et al. [86].

A popular method for sampling from the posterior distribution of the parameters given the data is the Gibbs sampler, a Markov chain Monte Carlo (MCMC) method that focuses on the so-called full conditional distributions of each parameter given all others, turning a complex multivariate problem into a series of simpler univariate ones. This approach has been combined with a general method for drawing samples from any log-concave distribution, called adaptive rejection sampling [29]. The combination is available in the software package BUGS [86]. Convergence diagnostics can be calculated using a set of R or S-Plus functions, see Best et al. [9].

Results for Simulated Data

We tried the Gibbs sampler on our simulated Guatemalan data. We used non-informative priors, treating all four fixed-effect parameters as i.i.d. normal variates with mean 0 and precision 10^{-6}. For the two random-effect parameters representing the precision of the family and community random effects, we used a $\Gamma(\epsilon, \epsilon)$ distribution with $\epsilon = 0.001$, so the mean is 1 and the variance is 1000. We then ran a naive Gibbs sampler with a burn-in of 200 iterations followed by a further 1000 iterations. We are very grateful to David Clayton for sharing with us a set of C functions for MCMC estimation of generalized linear mixed models and for adapting his driver program to handle our simulated data. These routines have now been incorporated in the R package `GLMMGibbs`, see Myles and Clayton [63].

The results in Table 9.1 are very encouraging, showing practically no bias in the estimation of the fixed effects, about a 5% bias in the estimation of the community effect, and an 8% bias for the family effect. We did some further work exploring the nature of the remaining bias and discovered that we could

essentially eliminate it by either (1) using informative priors for the precisions of the random effects, or (2) using a much larger sample size, simulated by combining our original samples in groups of five. For additional simulation results, see Browne and Draper [13].

Experience with Real Data

Our experience applying MCMC methods to real data has been somewhat mixed. In a case study fitting a three-level random-intercept logit model to data on immunization from Guatemala, we found slow mixing and poor convergence, particularly for parameters representing the variances of the random effects. Deciding whether a run is adequate often requires a battery of diagnostic procedures; we have used tests due to Geweke [28] and Roberts [78], and have found very useful the `gibbsit` software of Raftery and Lewis [75, 76], which provides an estimate of the number of iterations required to estimate credible limits for each parameter with given probability of attaining a desired precision.

Fitting a similar model for prenatal care data characterized by heavier clustering, particularly at the family level, proved substantially more difficult, with estimates of the efficiency of our chains as low as 1%. Rather than running much longer chains, we heeded the advice of Gelman and Rubin [27] and ran multiple chains with different starting values. The S-Plus function `itsim` was very useful in checking the output from multiple chains before pooling them to produce final estimates. In the end, the MCMC approach required extensive computation and judging convergence proved something of an arcane art form. For more details, see Rodríguez and Goldman [82].

9.3.5 Other Approaches to Estimation

High-Order Laplace

Breslow and Lin [12] proposed a fourth-order Laplace approximation for two-level models with a single random effect per cluster, and Lin and Breslow [52] extended the result to multiple independent random effects per cluster. More recently, Raudenbush et al. [77] further extended this approach to high-order approximations for multiple dependent random effects. They report that the method is remarkably accurate and computationally fast, and validate it by comparison to Gauss-Hermite quadrature with up to 40 points, using real and simulated data. This promising strategy was first implemented for two-level models in version 5 of HLM, but has now been extended to three-level models in version 6.

Simulated Maximum Likelihood

Monte Carlo integration is not restricted to Bayesian models, but can also be used for simulating the likelihood; see Lerman and Manski [50] for an early application. Closely related approaches are the method of simulated moments (MSM) introduced by McFadden [61], and the method of simulated scores (MSS), see Keane [44]. These methods are often used by applied economists estimating complex structural models. A useful survey may be found in Gouriéroux and Monfort [35].

In the context of generalized linear mixed models, McCulloch [60] developed Monte Carlo variants of the Expectation-Maximization (EM) and Newton-Raphson algorithms, as well as simulated maximum likelihood (SML). Booth et al. [10] compare several stochastic alternatives to numerical integration, including simulated maximum likelihood using importance sampling. These methods are particularly appropriate for high-dimensional models where quadrature succumbs to the curse of dimensionality.

Recently, Ng et al. [67] evaluated several simulation-based approaches for maximum likelihood estimation in multilevel models with binary outcomes, including bias correction using Kuk's bootstrap (described earlier) and the Robbins-Monro stochastic approximation method, and estimation using simulated maximum likelihood (SML). They conclude that SML performs comparably with the other methods, but has the advantage of yielding variance estimates—which can be used to construct Wald tests and confidence regions—as well as the value of the likelihood at the maximum, which is useful for constructing likelihood ratio tests to compare nested models. They note that SML requires good starting values, confirming results in [60], but is otherwise less prone to computational problems than the other algorithms, and gives results similar to numerical integration.

9.4 Infant and Child Mortality in Kenya

Our illustration of MGLMs uses data from the 1998 Kenya Demographic and Health Survey (KDHS) to study infant and child mortality.

9.4.1 The Kenya Survey

The 1998 Kenya Demographic and Health Survey (KDHS) is a national survey conducted by the National Council for Population and Development (NCPD) in collaboration with the Central Bureau of Statistics (CBS) and Macro International, which provided technical assistance. The survey is national in scope but excluded seven districts accounting for less than 4% of the population. The sample was selected using a two-stage stratified design and relied on a

sampling frame maintained by the CBS. Field work was conducted between February and July 1998, and achieved an overall response rate of 96.8% of households and 95.7% of women aged 15–49 who were eligible for an individual interview. The interview included a retrospective maternity history that collects data on date of birth, survival status, and age at death for all children each woman has given birth to.

We selected for analysis all births in the 10 years preceding the interview, but excluded 170 pairs of twins and one set of triplets, which have much higher mortality risks than singletons. The final sample consists of 10,878 births to 4,939 women who live in 530 communities, defined in terms of the ultimate area units used in the sample design. One objective of our analysis is to determine the extent to which infant and child deaths are clustered within families and within communities.

We must note at the outset a limitation of the data: The community is defined in terms of the respondent's residence at the time of the survey, but our analysis uses retrospective mortality data over a 10-year period. While this is far from ideal, we claim three extenuating circumstances. First, a large fraction of respondents have always lived in the place where they were interviewed, and 80.9% of all births in our sample were born while the mother resided in her current community. Second, migration would tend to attenuate the influence of the community, so our estimates can be considered lower bounds on the true effects. Third, as a sensitivity test we repeated our analysis using only births in the last five years, and discovered that our estimates were remarkably resilient to the choice of reference period.

9.4.2 A Three-Level Hazard Model

Let $\underline{\lambda}_{ijk}(t)$ denote the risk of dying at age t for the i-th child of the j-th mother in the k-th community. We assume that the hazard depends on age t, a set of observed child, family, and community covariates $\boldsymbol{x}_{ijk}$, and unobserved family and community random effects $\underline{\delta}_{jk}$ and $\underline{\delta}_k$ via a conditional proportional hazards model:

$$\underline{\lambda}_{ijk}(t) = \lambda_0(t) \exp\{\boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \underline{\delta}_{jk} + \underline{\delta}_k\}, \tag{9.30}$$

where $\lambda_0(t)$ is a baseline hazard, $\boldsymbol{\beta}$ is a vector of fixed parameters representing the effects of observed covariates, and the unobserved family and community effects are normally distributed, $\underline{\delta}_{jk} \sim \mathcal{N}(0, \sigma_2^2)$ and $\underline{\delta}_k \sim \mathcal{N}(0, \sigma_3^2)$.

Choice of Duration Categories

We assume that the baseline hazard is constant in intervals defined by cut-points $0 = \tau_0 < \tau_1 < \cdots < \tau_D$, so that $\lambda_0(t) = \lambda_{0d}$ if $t \in [\tau_{d-1}, \tau_d)$. The choice

of cutpoints is dictated by the shape of the hazard and constraints in data collection.

The KDHS recorded age at death in days, months, or years. Days are used for neonatal deaths (occurring in the first month of life), months are used mostly for infant deaths (occurring before age 1), and years are used predominantly for deaths at ages 2 or higher. We first tabulated events and exposure by single months up to age 1 and by single years thereafter. In calculating exposure for deaths at ages 2 and higher, we treated deaths as occurring at the midpoint of an interval constrained by the reported age at death in years and the date of interview. No such approximation is required for deaths at earlier ages or for survivors.

Following some exploratory work, we decided to use separate exposure categories for the first month of life, and then for ages 1–5, 6–11, 12–23, and 24–59 completed months, with more detail at ages where the hazard is changing rapidly. These five categories capture more than 90% of the variation in the hazard by duration (as measured by the deviance in a marginal Poisson model), and yield 48,094 pseudo-observations. For some preliminary analyses, we used only three categories: the first month of life, the rest of the first year, and older ages, which reduced the number of pseudo-observations to 30,456 and yielded very similar results.

Selection of Explanatory Variables

Our selection of variables has been guided by previous work in the field; see Mosley and Chen [62] for a conceptual framework. We included only one community-level variable, type of place of residence, classified as urban or rural. Residence is coded at the time of the survey, so the same caveat we discussed for community effects applies here.

Our only family-level variable is mother's education, which can be coded in terms of completed years or using dummy variables to mark achievements such as completing primary or secondary school. Our exploratory analysis indicated that the most efficient way to capture the educational effect was to use linear and quadratic terms. We found that mortality increased as one moved up from no education to complete primary, and decreased only when one went past secondary education, but this tendency became less noticeable after controlling for mother's age, which plays the role of a confounding factor: The children of very young mothers have higher mortality risks, but young women also tend to be more educated than older women, a fact that actually lowers their children's risk.

All remaining variables are defined at the individual level. Males are known to have higher mortality than females, so we included a dummy variable for sex. First- and high-order births are also at increased risk. We considered using dummy variables for first births and for births of order six and higher,

but noticed that linear and quadratic terms did a better job of capturing what appeared to be a gradual increase in risk with birth order.

An important determinant of mortality is length of the preceding birth interval, which of course is defined only for births of order two or higher. Children born shortly after a previous birth are known to have much higher risks, either because of maternal depletion or because they have to compete with older siblings for scarce resources. To capture this effect, we used a linear spline defined as $30-i$ (where i is interval length) for intervals shorter than 30 months and 0 for first births and for longer intervals. The linear spline proved significantly better than a simple dummy for short intervals.

The final individual variable in our model is age of the mother at the time of birth of the child, which is known to have a U-shaped relationship with mortality, with higher hazards for the youngest and oldest mothers. We tried dummy variables for mothers aged < 20 and 40+ at the time of birth of the child, but discovered that linear and quadratic terms on age at birth did a better job.

As part of our exploratory work, we allowed all of these variables to interact with child's age. We found no evidence of non-proportional effects except possibly for mother's education, which appeared to have a larger effect beyond the first month of life. However, the reduction in deviance did not justify the additional number of parameters required, as judged by Akaike's information criterion, so we retained the simpler proportional hazards model.

Estimation Results

Table 9.2 shows the results of fitting our final model by first-order MQL, first-order PQL, and maximum likelihood via 12-point Gauss-Hermite quadrature (ML). We also include for comparison results from a marginal Poisson model that ignores clustering at the family and community levels. Unlike some of the results we have obtained for heavily clustered binary data, in this application all three methods yield similar estimates of the fixed effects. In fact, the results are very similar to the marginal model as well, except possibly for cohort and birth order. However, the marginal model underestimates standard errors by an average of 8%, and does a poor job estimating the precision of the urban effect. The estimates of the random parameters, reported here in terms of the standard deviation of the family and community effects, are unusual in that MQL and PQL lead to slightly larger values than Gauss-Hermite quadrature.

First-order MQL converged quickly and uneventfully. First-order PQL alternated between two sets of estimates of the random parameters, one of which had the family variance component set to zero. The other, reported in Table 9.2, yielded results similar to MQL. We tried second-order MQL and PQL, but both failed repeatedly from a variety of starting points. We also tried these procedures with the smaller sample of 30,456 pseudo-observations

Table 9.2 Parameter estimates for the multilevel model of infant and child survival in Kenya.

Variable	Term	GLM	MQL-1	PQL-1	ML
		Fixed Effects			
Constant	1	−4.189	−4.163	−4.164	−4.588
		(0.095)	(0.105)	(0.106)	(0.118)
Age	1–5	−1.669	−1.646	−1.647	−1.642
(months)		(0.089)	(0.090)	(0.090)	(0.089)
	6–11	−2.062	−2.005	−2.007	−1.998
		(0.096)	(0.097)	(0.097)	(0.097)
	12–23	−2.912	−2.830	−2.834	−2.822
		(0.105)	(0.104)	(0.105)	(0.106)
	24–59	−3.748	−3.641	−3.646	−3.632
		(0.108)	(0.106)	(0.108)	(0.109)
Sex	male	0.080	0.087	0.087	0.087
		(0.065)	(0.067)	(0.067)	(0.068)
Cohort	1993+	0.195	0.173	0.173	0.173
		(0.066)	(0.068)	(0.069)	(0.069)
Mother's	$a - 25$	−0.060	−0.048	−0.048	−0.047
Age		(0.010)	(0.011)	(0.011)	(0.011)
	$(a - 25)^2$	0.003	0.003	0.003	0.003
		(0.001)	(0.001)	(0.001)	(0.001)
Birth	$o - 3$	0.079	0.046	0.047	0.043
Order		(0.035)	(0.038)	(0.038)	(0.039)
	$(o - 3)^2$	0.005	0.004	0.004	0.004
		(0.004)	(0.005)	(0.005)	(0.005)
Birth	$(30 - i)_+$	0.039	0.036	0.036	0.036
Interval		(0.006)	(0.006)	(0.006)	(0.006)
Mother's	$e - 7$	−0.074	−0.066	−0.066	−0.068
Education		(0.014)	(0.015)	(0.015)	(0.015)
	$(e - 7)^2$	−0.008	−0.007	−0.007	−0.007
		(0.002)	(0.003)	(0.003)	(0.003)
Residence	urban	0.022	−0.001	0.001	0.040
		(0.102)	(0.144)	(0.144)	(0.142)
		Random Effects			
Family	σ_2	–	0.732	0.696	0.613
	$\log \sigma_2$	–	−0.312	−0.363	−0.489
		–	(0.102)	(0.096)	(0.140)
Community	σ_3	–	0.747	0.745	0.680
	$\log \sigma_3$	–	−0.291	−0.294	−0.386
		–	(0.068)	(0.058)	(0.081)
Log-likelihood		−5688.86	–	–	−5602.12

Standard errors shown in parentheses.

using only three duration categories and obtained similar results. We believe that further exploration of the properties of MQL and PQL for Poisson data with moderate and large amounts of clustering would be useful. The ML estimates converged quickly. We verified our calculations for two-level models that included only the family or community effect by running Stata's `xtpois` procedure, which uses adaptive Gaussian quadrature for normal random effects, obtaining practically identical results.

Testing Random Parameters

A final technical point before we turn to the interpretation of the results concerns testing for family and community effects. In Table 9.2, we report standard errors for $\log \sigma$ rather than σ because normal approximations tend to work better in the unconstrained scale. One must be careful not to divide the estimate by its standard error, as this would test the hypothesis $H_0 : \sigma = 1$ rather than $H_0 : \sigma = 0$. Instead, we build 95% confidence intervals in the log scale and exponentiate to obtain intervals for σ. In our example, the confidence intervals are $(0.467, 0.807)$ for the family and $(0.580, 0.797)$ for the community σ, indicating large effects. Note that by construction these intervals cannot include zero, so they should not be used as formal tests.

Likelihood ratio tests are preferable, but are not without difficulties. Because the null hypothesis $H_0 : \sigma = 0$ is on the boundary of the parameter space, the likelihood ratio test does not have the usual large sample chi-squared distribution with degrees of freedom equal to the number of parameters set to zero, see Self and Liang [85] and Stram and Lee [88]. These authors suggest treating the test for $H_0 : \sigma_2 = \sigma_3 = 0$ as a 50-50 mixture of χ^2_1 and χ^2_2 rather than the nominal χ^2_2. Similarly, a test of $H_0 : \sigma_2 = 0$ or $H_0 : \sigma_3 = 0$ would be treated as an equal mixture of zero and a χ^2_1. Pinheiro and Bates [71] simulate likelihood ratio statistics in the context of linear mixed models and note that these adjustments are not always successful. A simpler approach is to use the nominal degrees of freedom, understanding that the test would then be conservative. In our application, twice the difference in log-likelihoods between the marginal and conditional models is 173.5, and the effect is highly significant no matter how we treat the test criterion.

9.4.3 Fixed Effects Estimates

The first thing to note in Table 9.2 is the remarkable decline in risk with age. Exponentiating the coefficients for durations 1–5 and 12–23 we see that by ages one to five completed months, the risk is 81% lower—and by age one completed year, it is 95% lower—than in the first month of life. Males in this sample have a 9% higher risk than females with the same characteristics, but this difference is not significant.

Children born in the period since January 1993, however, have 19% *higher* risk at any given age than children born in 1992 or earlier. We examined this result closely for possible artifacts, including sensitivity to the choice of duration categories, and found it to be robust. We also looked at survival to age 1 using logit models to compare births in the periods 1–4 and 5–9 years before the survey, with similar results. It seems clear that infant and child mortality increased in Kenya in the late 1990s, an unfortunate development that is probably related to the AIDS pandemic.

Mother's age at the time of birth of the child has a significant effect on survival. The left panel in Fig. 9.1 shows the expected U-shaped relationship. The risk reaches its minimum around age 32, at which point a 10-year difference in either direction increases the risk by as much as 40%, everything else being equal. Birth order, on the other hand, has no significant effect on survival, with sample estimates suggesting, if anything, a linear increase in risk with parity. The excess risk often observed for first-order births appears to have been captured by mother's age.

Short birth intervals have a strong negative effect on infant and child survival, as expected. The hazard increases 4% for each month that the interval falls short of 30, the arbitrary cutoff point in our linear spline. This translates into a 24% excess risk for children born two years after a sibling, compared to children born after an interval of two and a half years.

Mother's education, which ranges from 0 to 19 years with quartiles at 4, 7, and 8, has a large effect on infant and child mortality. The right panel in Fig. 9.1 shows the overall relationship: We see little if any effect of just a few years of primary education, but a large (and increasing) effect after that. Around the median, each year of education is associated with a 7% decline in risk.

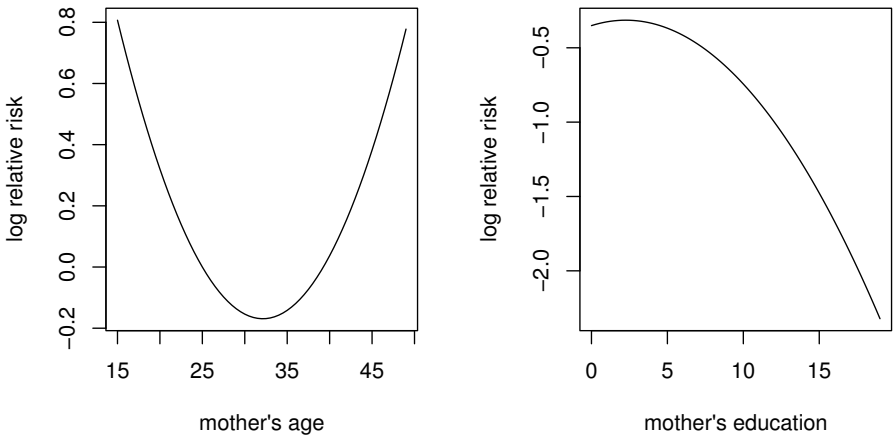

Fig. 9.1 The effects of mother's age and education on the log relative risk.

Finally, we find no significant effect of residence on child survival. Interestingly, urban residents in our sample have a 4% higher risk than their rural counterparts. We speculate that the erosion of the traditional differential that favors urban residence may be associated with higher prevalence of AIDS in the cities.

9.4.4 The Random Parameters

The most remarkable feature of our results concerns the large amount of clustering observed at both the family and community levels. In an analysis of family effects on infant and child mortality in Guatemala, Guo and Rodríguez [37] find much smaller family effects, and note that their results are in line with previous work in the area; see also Sastry [83].

Consider first the family random effect, which is estimated to have a standard deviation of 0.61. Because this effect is in the scale of the linear predictor, it can be interpreted exactly the same way as a fixed coefficient pertaining to an observed covariate. In particular, the children of a mother who is one standard deviation above the mean in a latent distribution of family frailty have 85% higher risk than the children of an average mother. In contrast, the children of a mother who is one standard deviation below the mean enjoy 45% lower risk than the children of the average mother. In both cases, the comparison is with children with identical observed characteristics who live in the same community.

The community random effect is, surprisingly, even larger, with a standard deviation of 0.68. Children who live in a community whose frailty is one standard deviation above the mean have almost double the risk—while those who live in communities one standard deviation below the mean have about half the risk—compared to children with the same observed characteristics who live in an average community. From a public health point of view, it would be interesting to identify communities with large estimated random effects, in search for an explanation of these findings.

One way to put these results in perspective is to look at the effect of observed characteristics other than age of the child. We computed the observed log relative risk, defined as the linear predictor omitting the constant, the dummy variables representing duration, and both random effects. The way we coded our covariates, this risk is zero for a third child, female, born before 1993, born at least two and a half years after the second birth, whose mother was 25 at the time of birth, had completed seven years of education, and lived in a rural area. For a similar male born after 1993 in a city, the log relative risk is 0.30. In our sample, log relative risks range from -2.04 to 2.16; selected percentiles are shown in Table 9.3.

Exponentiating these numbers, we find that children in the third quartile of relative risk have 61% higher risk than those in the first quartile. In contrast,

Table 9.3 Selected percentiles of log relative risk.

P	1	5	10	25	50	75	90	95	99
lrr	−0.78	−0.30	−0.13	0.15	0.38	0.63	0.87	1.02	1.31

the inter-quartile ranges in unobserved family and community characteristics translate into 2.3-fold and 2.5-fold increases in risk, respectively. Similarly, the range from the first to the 99th percentile in observed risk factors translates into an 8-fold increase in risk, whereas the equivalent ranges in the normal distributions representing family and community effects translate into 17-fold and 24-fold increases in risk, respectively. By this account, substantial relative risks associated with family and community frailty remain unobserved.

9.4.5 Survival Probabilities

We now translate our results into conditional and marginal probabilities of surviving to (or dying by) selected ages. This calculation can be done for selected values of the covariates, and helps present the results of hazard models in a less technical language.

Table 9.4 shows the conditional probabilities of infant and child death for our reference category and for children at the first and third quartile of observed risk factors and unobserved family and community effects. The underlying survival probabilities are all estimated as

$$S(t \mid \boldsymbol{x}_{ijk}, \delta_{jk}, \delta_k) = \exp\{-\Lambda_0(t) \exp\{\boldsymbol{x}'_{ijk}\hat{\beta} + \delta_{jk} + \delta_k\}\}, \tag{9.31}$$

with the log relative risk $\boldsymbol{x}'_{ijk}\hat{\boldsymbol{\beta}}$ set to the observed quartiles 0.15 and 0.63, and the unobserved frailties set to the normal quartiles $\pm 0.67\hat{\sigma}_2$ and $\pm 0.67\hat{\sigma}_3$.

Table 9.4 Estimated infant and child mortality at first and third quartiles of observed and unobserved risk.

Risk Factor			Mortality	
Observed	Family	Community	Infant	Child
Q1	Q1	Q1	0.014	0.022
		Q3	0.034	0.053
	Q3	Q1	0.031	0.049
		Q3	0.075	0.118
Q3	Q1	Q1	0.022	0.035
		Q3	0.054	0.085
	Q3	Q1	0.049	0.078
		Q3	0.119	0.183
Baseline			0.028	0.044

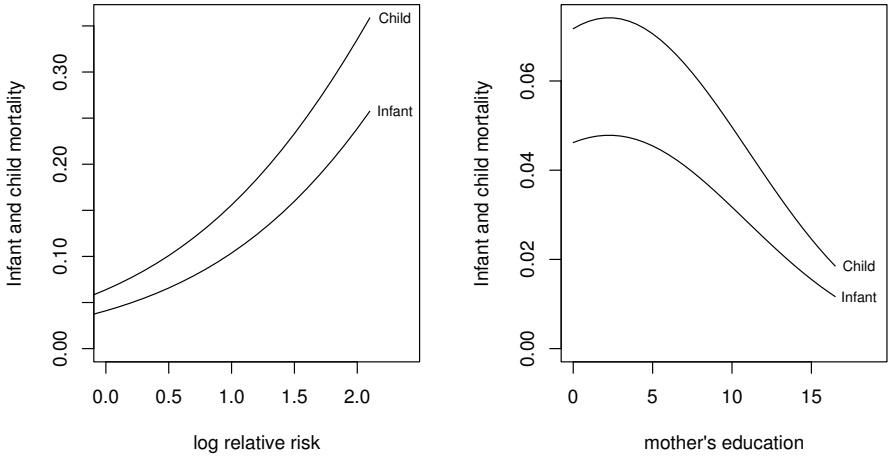

Fig. 9.2 Marginal probabilities of infant and child death by log relative risk and by mother's education.

As we move up the quartiles of observed and unobserved risk factors, the probability of an infant death increases from 14 to 119 per thousand, and the probability of a child death increases from 22 to 183 per thousand.

Figure 9.2 shows the marginal probabilities of infant and child death as a function of the log relative risk that combines all observed predictors (left panel), and as a function of mother's education with all other variables set to their reference values (right panel). The corresponding survival probabilities are estimated by evaluating the double integral

$$S(t \mid \boldsymbol{x}_{ijk}) = \int\int S(t \mid \boldsymbol{x}_{ijk}, \delta_{jk}, \delta_k)\ \mathrm{d}\delta_{jk}\ \mathrm{d}\delta_k \tag{9.32}$$

using 12-point Gauss-Hermite quadrature with conditional probabilities estimated using (9.31).

The marginal probability of infant death varies from 6 to 258 per thousand as a function of observed risk factors. The equivalent range for mortality up to age five is 9 to 359 per thousand. The effect of mother's education is fairly substantial. The probability that a child in our reference category will die before age one ranges from 47 per thousand if the mother has only a few years of education to 25 per thousand for high school graduates (and even less for the few women with higher education), after averaging over unobserved family and community attributes. Similarly, the probability of dying before age five declines from 75 to 39 per thousand, on average, as mother's education increases through upper primary and high school.

9.4.6 Intraclass Correlations

The variance parameters in random intercept models are closely related to measures of intraclass correlation. In a two-level linear model, the Pearson correlation between any two observations in the same cluster is $\rho = \sigma_2^2/(\sigma_2^2+\sigma_1^2)$. In a two-level logit model, the correlation is usually calculated by reference to the *latent* variable formulation of the model, setting $\sigma_1^2 = \pi^2/3$, the variance of the underlying standard logistic error, see Chapter 6. Rodríguez and Elo [80] show that the correlation of observed or *manifest* binary outcomes in two-level models can be quite different, and provide a Stata command `xtrho` that can be used to compute marginal and joint probabilities, and hence measures of correlation such as Person's r or Yule's Q, by numerical integration. Their ideas are easily extended to three-level survival, as shown below.

In the context of survival models, Oakes [68] has shown that the variance in a two-level model where frailty has a gamma distribution is closely related to Kendall's τ, a coefficient of ordinal association. No similar results have been obtained in general, but having fitted a multilevel survival model we can estimate any measure of association as a function of the estimated joint and marginal distributions. Because we followed children up to age 5 only, we are not in a position to estimate the correlation of lifetimes, but we can estimate correlation in survival up to ages one and five.

We calculate three marginal probabilities that are useful in constructing measures of intraclass correlation. First, we need the probability that a child with covariates $\boldsymbol{x}$ will live to age t, which is given by (9.32). Second, we need the probability that two children of the same mother both survive to age t. Because the survival experiences of these two children are independent given the family and community random effects, we can calculate the bivariate survival probability as

$$S_2(t,t \mid \boldsymbol{x}_{ijk}, \boldsymbol{x}_{i'jk}) = \int\int S(t \mid \boldsymbol{x}_{ijk}, \delta_{jk}, \delta_k)\, S(t \mid \boldsymbol{x}_{i'jk}, \delta_{jk}, \delta_k)\, \mathrm{d}\delta_{jk}\, \mathrm{d}\delta_k, \quad (9.33)$$

where the double integral is evaluated by Gauss-Hermite quadrature. We usually set $\boldsymbol{x}_{ijk} = \boldsymbol{x}_{i'jk}$, although only variables at levels 2 and 3 would need to be the same. Third, we need the probability that two children of different mothers who live in the same community will both survive to age t. Given the community random effect δ_k the survival experiences of these two children are independent, and the probability of surviving to age t can be calculated for each one by integrating out the corresponding family effect. The probability in question is then

$$S_3(t, t \mid \boldsymbol{x}_{ijk}, \boldsymbol{x}_{i'j'k}) = \int \left(\int S(t \mid \boldsymbol{x}_{ijk}, \delta_{jk}, \delta_k) \, \mathrm{d}\delta_{jk} \right. \\ \left. \times \int S(t \mid \boldsymbol{x}_{i'j'k}, \delta_{j'k}, \delta_k) \, \mathrm{d}\delta_{j'k} \right) \mathrm{d}\delta_k, \quad (9.34)$$

and can also be evaluated by Gauss-Hermite quadrature. We usually set $\boldsymbol{x}_{ijk} = \boldsymbol{x}_{i'j'k}$, although only variables at level 3 need be the same.

With these three probabilities in hand, we can now calculate any measure of correlation for binary outcomes. For example, the Pearson correlation between the indicators of survival to age t for two children of the same mother with observed covariates $\boldsymbol{x}$ is given by

$$\rho_2(t, \boldsymbol{x}) = \frac{S_2(t, t \mid \boldsymbol{x}, \boldsymbol{x}) - S(t \mid \boldsymbol{x})^2}{S(t \mid \boldsymbol{x})[1 - S(t \mid \boldsymbol{x})]}, \quad (9.35)$$

where $S_2(t, t \mid \boldsymbol{x}, \boldsymbol{x})$ is the joint survival probability from (9.33) and $S(t \mid \boldsymbol{x})$ is the marginal probability from (9.32). A similar expression applies to the correlation for children of different mothers living in the same community, but using (9.34) for the joint probability. These measures of intraclass correlation are a function of the marginal and joint probabilities of survival to age one or five, which in turn depend on the linear predictor as well as the variances of the random effects.

Figure 9.3 shows these correlations calculated over the entire range of observed relative risks in Kenya using the estimated values of σ_2 and σ_3 in

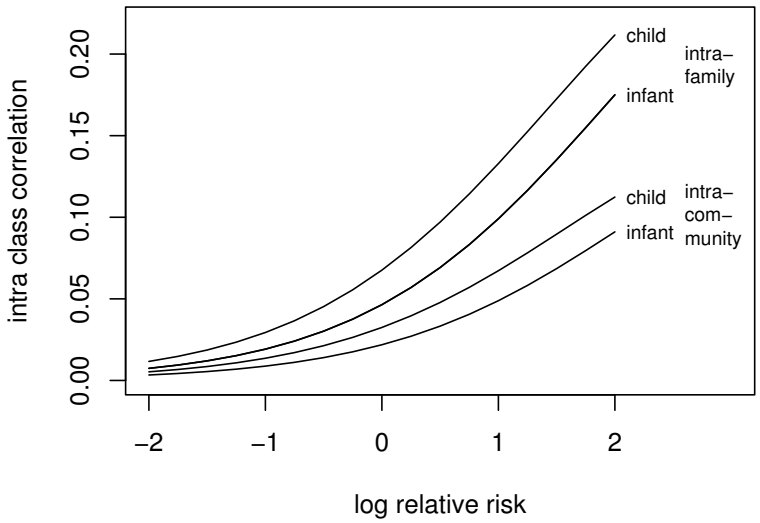

Fig. 9.3 Intra-family and intra-community correlations in infant and child mortality, by log relative risk.

Table 9.2. The intra-family correlations, which result from children sharing unobserved family and community characteristics, are always higher than the intra-community correlations, which result from sharing unobserved community characteristics only. The correlations are also higher for child than for infant mortality (or their complements, survival to ages five and one, respectively), and increase with the relative risk as measured from observed covariates. For our reference cell, the intra-family correlation is 0.05 for infant and 0.07 for child deaths, but these numbers increase to 0.18 and 0.21 at the highest levels of risk. The fact that the correlation between observed outcomes in the same family or community increases with the level of risk parallels the results obtained for two-level logit models in [80].

9.5 Summary and Conclusions

In this chapter we described generalizations of the multilevel model that go beyond normally distributed outcomes to cover a wide range of continuous and discrete responses, including binary, count, and survival data. The distinguishing feature of the generalization is the assumption that *conditional* on a set of random effects, the outcomes are independent and follow a standard generalized linear model. In this extension, a transformation of the conditional mean given a set of observed covariates and unobserved random effects follows a linear model. In a survival context, the conditional hazard has a similar structure. We contrasted this approach with models that focus on the *marginal* distribution of the outcomes, and with models that assume Gaussian outcomes but a non-linear structure of effects.

We reviewed the three main approaches to estimation, including marginal and penalized quasi-likelihood, maximum likelihood using Gauss-Hermite quadrature, and Bayesian estimation using the Gibbs sampler. We reported results of a simulation study showing that for heavily clustered binary responses quasi-likelihood estimates can be severely biased, while maximum likelihood estimates are approximately unbiased. Bayesian estimates showed a small bias that could be eliminated by using informative priors or larger samples. We also commented on a case study using binary data from Guatemala that leads to similar conclusions, but reveals some of the convergence problems that arise with bootstrapping and Bayesian estimates. Finally, we presented an application to survival data from Kenya where the approximate procedures fared better. On balance, there is a clear need for fast and accurate estimation procedures that can be applied to a wide variety of models and datasets.

Our analysis of infant and child mortality in Kenya illustrates the close connection between piece-wise exponential survival models and generalized linear models with Poisson errors and log link. We showed how the risk of death varies between birth and age five as a function of observed character-

istics of the child, mother, and community, as well as unobserved random effects representing heterogeneity of frailty across families and communities. We found large effects on the hazard, and translated these into marginal and conditional probabilities of dying by age one and by age five. Finally, we developed measures of intra-family and intra-community correlation in infant and child deaths. The study illustrates how much more can be learned from a dataset by taking into account the group structure in the framework of multilevel generalized linear models.

Acknowledgements I am grateful to David Clayton for sharing his Gibbs sampling code and to Noreen Goldman and Erik Meijer for helpful comments on the manuscript. This work was supported by National Institute of Child Health and Human Development grant R01 HD35277.

References

1. O. O. Aalen. Heterogeneity in survival analysis. *Statistics in Medicine*, 7: 1121–1137, 1988.
2. M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions.* Number 55 in National Bureau of Standards Applied Mathematics Series. U.S. Government Printing Office, Washington, DC, 1964.
3. M. Aitkin, D. Anderson, B. Francis, and J. Hinde. *Statistical Modelling in GLIM.* Clarendon Press, Oxford, 1989.
4. M. Aitkin and D. G. Clayton. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*, 29:156–163, 1980.
5. P. D. Allison. Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13:61–98, 1982.
6. D. A. Anderson and M. Aitkin. Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B*, 47:203–210, 1985.
7. J. S. Barber, S. A. Murphy, W. G. Axinn, and J. Maples. Discrete-time multilevel hazard analysis. *Sociological Methodology*, 30:201–235, 2000.
8. S. Bennet and J. Whitehead. Fitting logistic and log-logistic regression models to censored data using GLIM. *GLIM Newsletter*, 4:12–19, 1981.
9. N. G. Best, M. K. Cowles, and S. K. Vines. *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, version 0.40.* Medical Research Council Biostatistics Unit, Cambridge, UK, 1997.
10. J. G. Booth, J. P. Hobert, and W. Jank. A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling*, 1:333–349, 2001.
11. N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

12. N. E. Breslow and X. Lin. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91, 1995.
13. W. J. Browne and D. Draper. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1:473–549, 2006. (with discussion)
14. D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151, 1978.
15. D. G. Clayton. The analysis of event history data: Review of progress and outstanding problems. *Statistics in Medicine*, 7:819–841, 1988.
16. D. G. Clayton. Generalized linear mixed models. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 275–301. Chapman & Hall, London, 1996.
17. D. G. Clayton and J. Cuzick. The EM algorithm for Cox's regression model using GLIM. *Applied Statistics*, 34:148–156, 1985.
18. D. G. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series B*, 148:82–117, 1985. (with discussion)
19. C. Corcoran, B. Coull, and A. Patel. *Egret for Windows User Manual*. Cytel Software Corporation, Cambridge, MA, 1999.
20. D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972. (with discussion)
21. D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.
22. D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman & Hall, London, 1984.
23. M. J. Crowder. Beta-binomial Anova for proportions. *Applied Statistics*, 27: 34–37, 1978.
24. M. Davidian and D. M. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London, 1995.
25. P. J. Diggle, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK, 1994.
26. P. Feigl and M. Zelen. Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 21:826–838, 1967.
27. A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992. (with discussion)
28. J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 169–194. Oxford University Press, Oxford, UK, 1992.
29. W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348, 1992.
30. H. Goldstein. Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78:45–51, 1991.
31. H. Goldstein. Multilevel models and generalised estimating equations. *Multilevel Modelling Newsletter*, 5(2):2, 1993.

32. H. Goldstein. Multilevel unit specific and population average generalised linear models. *Multilevel Modelling Newsletter*, 7(3):4–5, 1995.
33. H. Goldstein. Consistent estimators for multilevel generalised linear models using an iterated bootstrap. *Multilevel Modelling Newsletter*, 8(1):3–6, 1996.
34. H. Goldstein and J. Rasbash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, 159: 505–513, 1996.
35. C. Gouriéroux and A. Monfort. *Simulation-Based Econometric Methods*. Oxford University Press, Oxford, UK, 1996.
36. P. J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55:245–259, 1987.
37. G. Guo and G. Rodríguez. Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *Journal of the American Statistical Association*, 87: 969–976, 1992.
38. D. Hedeker and R. D. Gibbons. A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50:933–944, 1994.
39. D. Hedeker and R. D. Gibbons. MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49:157–176, 1996.
40. T. R. Holford. The analysis of rates and survivorship using log-linear models. *Biometrics*, 36:299–306, 1980.
41. P. Hougaard. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71:75–83, 1984.
42. P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396, 1986.
43. J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*, 2nd edition. Wiley, New York, 2002.
44. M. P. Keane. Simulation estimation for panel data models with limited dependent variables. In G. S. Maddala, C. R. Rao, and H. D. Vinod, editors, *Handbook of Statistics*, volume 11, pages 545–571. North-Holland, Amsterdam, 1993.
45. A. Y. C. Kuk. Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society, Series B*, 57: 395–407, 1995.
46. N. M. Laird. Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65:581–590, 1978.
47. N. M. Laird and D. Olivier. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76:231–240, 1981.
48. J. F. Lawless. Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82:808–815, 1987.
49. Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, 58:619–678, 1996.
50. S. R. Lerman and C. F. Manski. On the use of simulated frequencies to approximate choice probabilities. In C. F. Manski and D. McFadden, editors, *Structural*

Analysis of Discrete Data with Econometric Applications, pages 305–319. MIT Press, Cambridge, MA, 1981.

51. L. A. Lillard and C. W. A. Panis. *aML: Multilevel Multiprocess Statistical Software, Version 2.0*. EconWare, Los Angeles, CA, 2003.
52. X. Lin and N. E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016, 1996.
53. Q. Liu and D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81: 624–629, 1994.
54. N. T. Longford. A quasi-likelihood adaptation for variance component analysis. In *American Statistical Association Proceedings of the Statistical Computing Section*, pages 137–142. 1988.
55. N. T. Longford. *VARCL: Software for Variance Component Analysis of Data with Nested Random Effects (Maximum Likelihood)*. Educational Testing Service, Princeton, NJ, 1988.
56. N. T. Longford. Logistic regression with random coefficients. *Computational Statistics & Data Analysis*, 17:1–15, 1994.
57. T. A. Louis. Finding the observed information matrix when using the *EM* algorithm. *Journal of the Royal Statistical Society, Series B*, 44:226–233, 1982.
58. K. G. Manton, E. Stallard, and J. W. Vaupel. Alternative models for the heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association*, 81:635–644, 1986.
59. P. McCullagh and J. A. Nelder. *Generalized Linear Models*, 2nd edition. Chapman & Hall, London, 1989.
60. C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170, 1997.
61. D. McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.
62. W. H. Mosley and L. C. Chen. An analytical framework for the study of child survival in developing countries. *Population and Development Review*, 10:25–45, 1984.
63. J. Myles and D. G. Clayton. *`GLMMGibbs`: An `R` Package for Estimating Bayesian Generalised Linear Mixed Models by Gibbs Sampling*. Comprehensive R Archive Network (`devel` section), 2001. URL `http://cran.r-project.org`
64. J. C. Naylor and A. F. M. Smith. Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31:214–225, 1980.
65. J. A. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series B*, 135:370–384, 1972.
66. J. M. Neuhaus, J. D. Kalbfleisch, and W. W. Hauck. A comparison of cluster-specific and population-averaged approaches to analyzing correlated binary data. *International Statistical Review*, 59:25–35, 1991.
67. E. S. W. Ng, J. R. Carpenter, H. Goldstein, and J.Rasbash. Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling*, 6:23–42, 2006.

68. D. Oakes. A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*, 44:414–422, 1982.
69. A. R. Pebley, N. Goldman, and G. Rodríguez. Prenatal and delivery care and childhood immunization in Guatemala: Do family and community matter? *Demography*, 33:231–247, 1996.
70. J. C. Pinheiro and D. M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35, 1995.
71. J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000.
72. R. L. Prentice and L. A. Gloeckler. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34:57–67, 1978.
73. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*, 2nd edition. Cambridge University Press, Cambridge, MA, 1992.
74. S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1–21, 2002.
75. A. E. Raftery and S. M. Lewis. How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, Oxford, UK, 1992.
76. A. E. Raftery and S. M. Lewis. Implementing MCMC. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 115–130. Chapman & Hall, London, 1996.
77. S. W. Raudenbush, M.-L. Yang, and M. Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9: 141–157, 2000.
78. G. O. Roberts. Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 45–57. Chapman & Hall, London, 1996.
79. G. Rodríguez. Event history analysis. In S. Kotz, C. B. Read, and D. L. Banks, editors, *Encyclopedia of Statistical Sciences, Update Volume*, pages 222–230. Wiley, New York, 1997.
80. G. Rodríguez and I. Elo. Intra-class correlation in random-effects models for binary data. *The Stata Journal*, 3:32–46, 2003.
81. G. Rodríguez and N. Goldman. An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, 158:73–89, 1995.
82. G. Rodríguez and N. Goldman. Improved estimation procedures for multilevel models with binary response: A case-study. *Journal of the Royal Statistical Society, Series A*, 164:339–355, 2001.
83. N. Sastry. A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. *Journal of the American Statistical Association*, 92:426–435, 1997.
84. R. Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727, 1991.

85. S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association*, 82:605–610, 1987.
86. D. J. Spiegelhalter, A. Thomas, N. G. Best, and W. R. Gilks. *BUGS: Bayesian Inference Using Gibbs Sampling*. Medical Research Council Biostatistics Unit, Cambridge, UK, 1996.
87. R. Stiratelli, N. M. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971, 1984.
88. D. O. Stram and J. W. Lee. Variance components testing in the longitudinal mixed-effects model. *Biometrics*, 50:1171–1177, 1994.
89. T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.
90. R. A. Thisted. *Elements of Statistical Computing: Numerical Computation*. Chapman & Hall, London, 1988.
91. J. Vaupel, K. G. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.
92. J. Vaupel and A. Yashin. Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *American Statistician*, 39:176–185, 1985.
93. J. Whitehead. Fitting Cox's regression model to survival data using GLIM. *Applied Statistics*, 29:268–275, 1980.
94. G. Y. Wong and W. M. Mason. The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80:513–524, 1985.
95. S. L. Zeger and M. R. Karim. Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86, 1991.