Chapter 9
General Regression Models

9.1 Introduction

In this chapter we consider dependent data but move from the linear models of Chap. 8 to general regression models. As in Chap. 6, we consider generalized linear models (GLMs) and, more briefly, nonlinear models. We first give an outline of this chapter. In Sect. 9.2 we describe three motivating datasets to which we return throughout the chapter. The GLMs discussed in Sect. 6.3 can be extended to incorporate dependences in observations on the same unit; as with the linear model, an obvious way to carry out modeling in this case is to introduce unit-specific random effects. Within a GLM a natural approach is for these random effects to be included on the linear predictor scale. The resultant *conditional* models are known as generalized linear mixed models (GLMMs), and these are introduced in Sect. 9.3. In Sects. 9.4 and 9.5 we describe likelihood and conditional likelihood methods of estimation, with Sect. 9.6 devoted to a Bayesian treatment. Section 9.7 illustrates some of the flexibility of GLMMs by describing and applying a particular model for spatial dependence. An alternative random effects specification, based on conjugacy, is described in Sect. 9.8. An important approach to the modeling and analysis of dependent data that is philosophically different from the random effects formulation is via marginal models and generalized estimating equations (GEE), and these are the subject of Sect. 9.9. In Sect. 9.10, a second GEE approach is described in which the estimating equations for the mean are supplemented with a second set for the variances/covariances. For GLMMs, extra care must be taken with parameter interpretation, and Sect. 9.11 discusses this issue, emphasizing how interpretation differs between conditional and marginal models. In Part II of the book, which focused on independent data, Chap. 7 was devoted to models for binary data. For dependent data, models binary data are less well developed, and so we do not devote a complete chapter to their description. However, Sect. 9.12 introduces the modeling of dependent binary data, and, subsequently, Sects. 9.13 and 9.14 describe conditional (mixed) and marginal models for binary data. Section 9.15 considers how nonlinear models, as defined in Sect. 6.10, can be extended

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_9,

to the dependent data case. For such models, many applications concentrate on inference for units, and so the introduction of random effects is again suggested. We refer to the resultant class of models as nonlinear mixed models (NLMMs). Section 9.16 considers issues related to the parameterization of the nonlinear model. Inference for nonlinear mixed models via likelihood and Bayes approaches is covered in Sects. 9.17 and 9.18, while GEE is briefly considered in Sect. 9.19. The assessment of assumptions for general regression models is described in Sect. 9.20, with concluding comments contained in Sect. 9.21. Additional references appear in Sect. 9.22.

9.2 Motivating Examples

In this chapter we will analyze the lung cancer and radon data introduced in Sect. 1.3.3 and three additional datasets.

9.2.1 Contraception Data

Fitzmaurice et al. (2004) reanalyze data originally appearing in Machin et al. (1988) concerning a randomized longitudinal contraception trial. Each of 1,151 women received injections of 100 or 150 mg of depot medroxyprogesterone acetate (DMPA) on the day of randomization and three additional injections at 90-day intervals. There was a final follow-up 3 months after the last injection (a year after the initial injection). The women completed a menstrual diary throughout the study, and the binary response is whether the woman had experienced amenorrhea, the absence of menstrual bleeding for a specified number of days, during each of the four 3-month intervals. There was dropout in this study, but we will not address this issue, important though it is. The sample sizes, across measurement occasions, in the low- and high-dose groups are $[576, 477, 409, 361]$ and $[575, 476, 389, 353]$, respectively. Plotting the individual-level 0/1 data is usually not informative for binary data, and so in Fig. 9.1, we plot the averages, that is, the probabilities of amenorrhea over time for the two treatment groups. We see increasing probabilities of amenorrhea in both groups, with the probabilities in the 150-mg dose group being greater than in the 100-mg dose group.

As we will discuss in Sect. 9.14, for binary data, there is no obvious natural measure of dependence, unlike normal data for which the correlation is routinely used. However, Table 9.1 gives the empirical correlations between responses at different measurement occasions in the low- and high-dose groups, respectively. In both groups there is appreciable correlation between observations on the same woman, with a suggestion that the correlations decrease on measurements taken further apart. To explicitly acknowledge the dependence over time in responses on the same woman, multivariate binary data models are required.

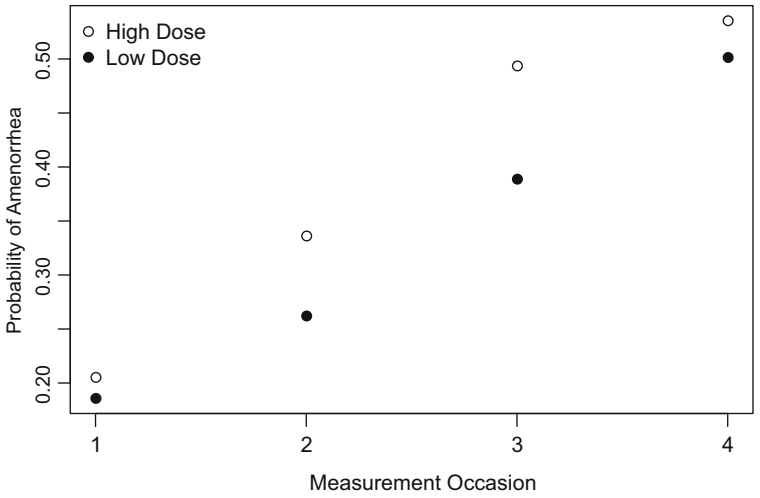

Fig. 9.1 Probability of amenorrhea over time in low- and high-dose groups, in the contraception data

Table 9.1 Empirical variances (on the *diagonal*) and correlations (on the *upper diagonal*), between measurements on the same woman at different observation occasions (1–4), in the low- (*left*) and high- (*right*) dose groups of the contraception data

	1	2	3	4		1	2	3	4
1	0.15	0.40	0.28	0.27	1	0.16	0.31	0.25	0.29
2		0.19	0.45	0.35	2		0.22	0.43	0.43
3			0.24	0.13	3			0.25	0.47
4				0.25	4				0.25

9.2.2 Seizure Data

Thall and Vail (1990) describe data on epileptic seizures in 59 individuals. For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks, after which patients were randomized to one of two groups: treatment with either the antiepileptic drug progabide or with placebo. The numbers of individuals in the placebo and progabide groups were 28 and 31, respectively. The number of seizures was recorded in four consecutive 2-week periods. For these data, let Y_{ij} represent the number of counts for patient i, $i = 1, \ldots, 59$ at occasion j, with $j = 0$ the baseline period and $j = 1, \ldots, 4$ the subsequent set of four 2-week measurement periods. Also, let T_j be the length (in weeks) of the observation period (which is the same for all individuals), with $T_0 = 8$ and $T_j = 2$ for $j = 1, \ldots, 4$. We might consider the model

$$Y_{ij} \mid \mu_{ij} \sim \text{Poisson}(\mu_{ij})$$

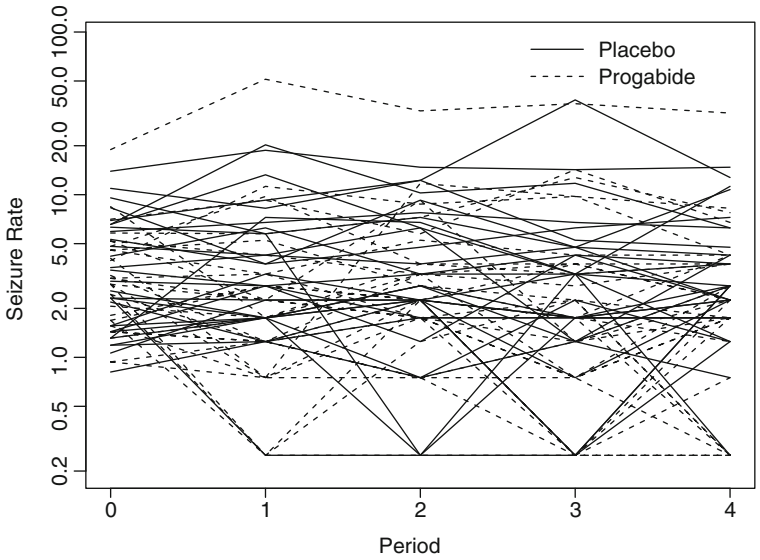

Fig. 9.2 Log of seizure rates by period for individuals on placebo and progabide

where $\mu_{ij} = T_j \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})$ and $\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})$ is a loglinear regression model. There are two immediate issues with this model: data on the same individual are unlikely to be independent and there may be excess-Poisson variation.

As a first look at the data, we plot the log seizure rate $\log[(Y_{ij}+0.5)/T_j]$ for each individual versus period j in Fig. 9.2. The 0.5 is added to avoid taking the log of zero. The line types distinguish the placebo and progabide groups. It is difficult to discern much pattern from this plot. In particular, it is not clear if progabide provides a drop in the rate of seizures, though there is clearly large between-individual variability in the rates. One individual's profile appears to be outlying and high, with the rate of seizures increasing after treatment with progabide.

Figure 9.3 displays the average seizure counts by period and by treatment group. In three out of the four post-baseline periods, the averages are lower in the progabide group. To assess the excess Poisson, we calculate the ratio of the variance of the counts to the mean, that is, $\text{var}(Y_{ij})/\text{E}[Y_{ij}]$, by period and treatment group. Table 9.2 gives these ratios and clearly shows that there is a great deal of excess-Poisson variability for these data.

9.2.3 Pharmacokinetics of Theophylline

Twelve subjects were given an oral dose of the antiasthmatic agent theophylline, with 11 concentration measurements obtained from each individual over 25 h. The doses ranged between 3.10 and 5.86 mg/kg. As is usual with experiments such as

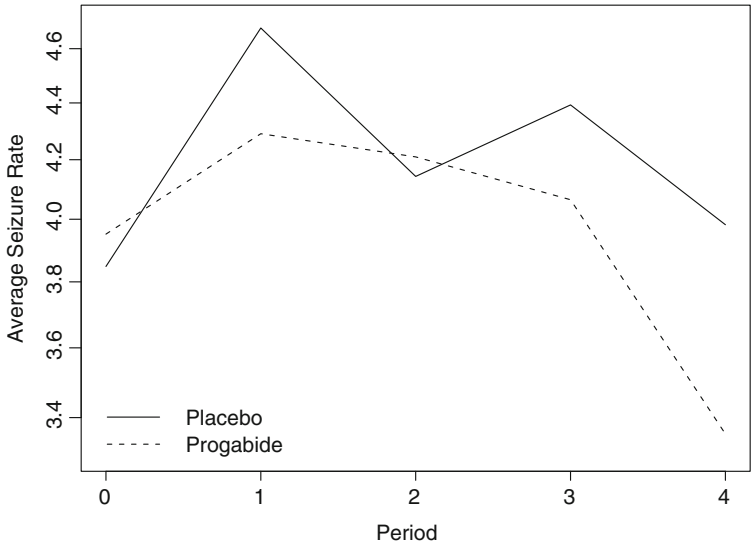

Fig. 9.3 Average seizure rates by period and treatment group

Table 9.2 Ratio of the variance of seizure counts to the mean of the seizure counts, by period and treatment group

	Period				
Group	0	1	2	3	4
Placebo	22.1	11.0	8.0	24.5	7.3
Progabide	24.8	38.8	16.7	23.7	18.9

this, there is abundant sampling at early times in an attempt to capture the absorption phase, which is rapid. Further background on pharmacokinetic modeling is given in Example 1.3.4 where the data for the first individual were presented. Section 6.2 introduced a mean model for these data (for a generic individual) as

$$\frac{Dk_a}{V(k_a - k_e)}\left[\exp(-k_e x) - \exp(-k_a x)\right]$$

where x is the sampling time, $k_a > 0$ is the absorption rate constant, $k_e > 0$ is the elimination rate constant, and $V > 0$ is the (apparent) volume of distribution (that converts total amount of drug into concentration). Figure 9.4 shows the concentration–time data. The curves follow a similar pattern, but there is clearly between-subject variability.

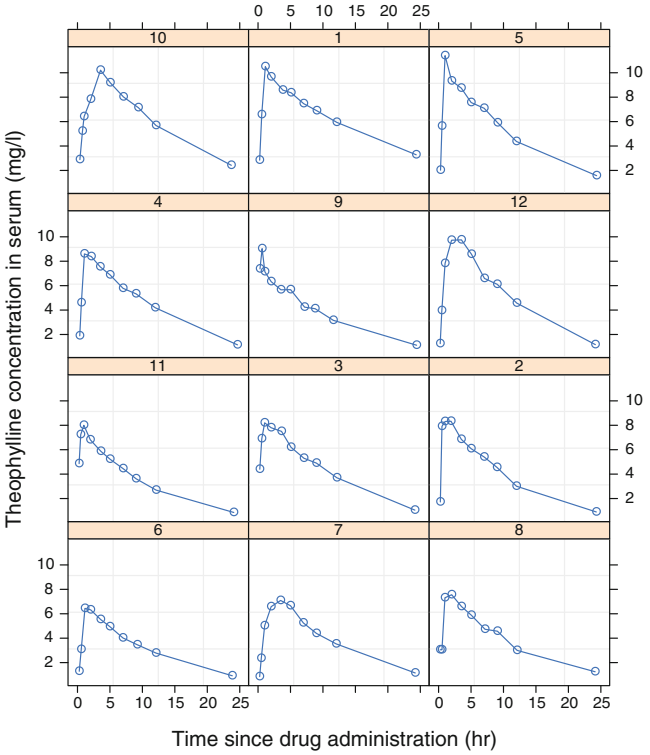

Fig. 9.4 Concentrations versus time for 12 individuals who received the drug theophylline

9.3 Generalized Linear Mixed Models

In this section we describe a modeling framework that allows the introduction of random effects into GLMs; these models induce dependence between responses on the same unit. Adding normal random effects on the linear predictor scale gives a GLMM. The paper of Breslow and Clayton (1993) popularized these models, by discussing implementation and providing a number of cases studies.

We first describe notation. Let Y_{ij} be the jth observation on the ith unit for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. The responses for the ith unit will be denoted $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]^{\mathrm{T}}$, $i = 1, \ldots, m$. Responses on different units will be assumed independent. Let $\boldsymbol{\beta}$ represent a $(k+1) \times 1$ vector of fixed effects and $\boldsymbol{b}_i$ a $(q+1) \times 1$ vector of random effects, with $q \leq k$. Let $\boldsymbol{x}_{ij} = [1, x_{ij1}, \ldots, x_{ijk}]$ be a $(k+1) \times 1$ vector of covariates, so that $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]$ is the design matrix for the fixed effects of unit i, and let $\boldsymbol{z}_{ij} = [1, z_{ij1}, \ldots, z_{ijq}]^{\mathrm{T}}$ be a $(q+1) \times 1$ vector of variables that are a subset of $\boldsymbol{x}_{ij}$, so that $\boldsymbol{z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i}]^{\mathrm{T}}$ is the design matrix for the random effects of unit i.

A GLMM is defined by the following two-stage model:

Stage One: The distribution of the data is $Y_{ij} \mid \theta_{ij}, \alpha \sim p(\cdot)$ where $p(\cdot)$ is a member of the exponential family, that is

$$p(y_{ij} \mid \theta_{ij}, \alpha) = \exp\left\{[y_{ij}\theta_{ij} - b(\theta_{ij})]/a(\alpha) + c(y_{ij}, \alpha)\right\}, \tag{9.1}$$

for $i = 1, \ldots, m$ units and $j = 1, \ldots, n_i$, measurements per unit. The variance is

$$\text{var}(Y_{ij} \mid \theta_{ij}, \alpha) = \alpha v(\mu_{ij}).$$

Let $\mu_{ij} = \text{E}[Y_{ij} \mid \theta_{ij}, \alpha]$ and, for a link function $g(\cdot)$, suppose

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i,$$

so that random effects are introduced on the scale of the linear predictor. This defines the *conditional* part of the model.

Stage Two: The random effects are assigned a normal distribution:

$$\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \text{N}_{q+1}(\,\boldsymbol{0}, \boldsymbol{D}\,).$$

For a number of reasons, including parameter interpretation, it is important to investigate the marginal moments that are induced by the random effects. Since marginal summaries may be calculated for the observed data, comparison with the theoretical forms is useful for model checking. The marginal mean is

$$\begin{aligned}\text{E}[Y_{ij}] &= \text{E}[\,\text{E}_{\boldsymbol{b}_i}(Y_{ij} \mid \boldsymbol{b}_i)\,] \\ &= \text{E}[\,\mu_{ij}\,] = \text{E}_{\boldsymbol{b}_i}[\,g^{-1}(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i)\,].\end{aligned}$$

The variance is

$$\begin{aligned}\text{var}(Y_{ij}) &= \text{E}[\,\text{var}(Y_{ij} \mid \boldsymbol{b}_i)\,] + \text{var}(\,\text{E}[Y_{ij} \mid \boldsymbol{b}_i]\,) \\ &= \alpha\text{E}_{\boldsymbol{b}_i}[\,v\{g^{-1}(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i)\}\,] + \text{var}_{\boldsymbol{b}_i}[\,g^{-1}(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i)\,].\end{aligned}$$

The covariances between outcomes on the same unit are

$$\begin{aligned}\text{cov}(Y_{ij}, Y_{ik}) &= \text{E}[\,\text{cov}(Y_{ij}, Y_{ik} \mid \boldsymbol{b}_i)\,] + \text{cov}[\,\text{E}(Y_{ij} \mid \boldsymbol{b}_i], \text{E}[Y_{ik} \mid \boldsymbol{b}_i)\,] \\ &= \text{cov}_{\boldsymbol{b}_i}[\,g^{-1}(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i), g^{-1}(\boldsymbol{x}_{ik}\boldsymbol{\beta} + \boldsymbol{z}_{ik}\boldsymbol{b}_i)\,] \\ &\neq 0,\end{aligned}$$

for $j \neq k$ due to shared random effects, and

$$\text{cov}(Y_{ij}, Y_{i'k}) = 0,$$

for $i \neq i'$, as there are no random effects in the model that are shared by different units. Explicit forms of the moments are available for some choices of exponential family, as we see later in the chapter, though the marginal distribution of the data is not typically available (outside of the normal case discussed in Chap. 8).

9.4 Likelihood Inference for Generalized Linear Mixed Models

As discussed in Sect. 8.5, there are three distinct sets of parameters for which inference may be required: fixed effects $\boldsymbol{\beta}$, variance components $\boldsymbol{\alpha}$, and random effects $\boldsymbol{b} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]^{\mathrm{T}}$. As with the linear mixed model (LMM), we maximize the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ denote the variance components in $\boldsymbol{D}$ and the scale parameter α (if present). The likelihood is obtained by integrating $[\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$ from the model:

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{m} \int p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{b}_i) \times p(\boldsymbol{b}_i \mid \boldsymbol{\alpha}) \, d\boldsymbol{b}_i.$$

There are m integrals to evaluate, each of dimension equal to the number of random effects, $q + 1$. For non-Gaussian GLMMs, these integrals are not available in closed form, and so some sort of analytical, numerical, or simulation-based approximation is required (Sect. 3.7). Common approaches include analytic approximations such as the Laplace approximation (Sect. 3.7.2) or the use of adaptive Gauss–Hermite numerical integration rules (Sect. 3.7.3). There are two difficulties with inference for GLMMs: carrying out the required integrations and maximizing the resultant (approximated) likelihood function. The likelihood function can be unwieldy, and, in particular, the second derivatives may be difficult to determine, so the Newton–Raphson method cannot be directly used. An alternative is provided by the quasi-Newton approach in which the derivatives are approximated (Dennis and Schnabel 1996).

One approach to the integration/maximization difficulties is the following. In Sect. 6.5.2 the iteratively reweighted least squares (IRLS) algorithm was described as a method for finding MLEs in a GLM. The penalized-IRLS (P-IRLS) algorithm is a variant in which the working likelihood is augmented with a penalization term corresponding to the (log of the) random effects distribution. This algorithm may be used in a GLMM context in order to obtain, conditional on $\boldsymbol{\alpha}$, estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}$, with $\boldsymbol{\alpha}$ being estimated via a profile log-likelihood (Sect. 2.4.2); see Bates (2011). The P-IRLS is also used for nonparametric regression and is described in this context in Sect. 11.5.1.

The method of penalized quasi-likelihood (PQL) was historically popular (Breslow and Clayton 1993) but can be unacceptably inaccurate, in particular, for binary outcomes. See Breslow (2005) for a recent perspective.

Approximate inference for $[\boldsymbol{\beta}, \boldsymbol{\alpha}]$ is carried out via the usual asymptotic normality of the MLE which is, with sloppy notation,

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \boldsymbol{I}_{\beta\alpha} \\ \boldsymbol{I}_{\alpha\beta} & \boldsymbol{I}_{\alpha\alpha} \end{bmatrix}^{-1} \right) \tag{9.2}$$

where $\boldsymbol{I}_{\beta\beta}$, $\boldsymbol{I}_{\beta\alpha}$, $\boldsymbol{I}_{\alpha\beta}$, and $\boldsymbol{I}_{\alpha\alpha}$ are the relevant information matrices. An important observation is that in general $\boldsymbol{I}_{\beta\alpha} \neq \boldsymbol{0}$, and so we cannot separately estimate the regression and variance parameters, so consistency requires correct specification of both mean and variance models. Likelihood ratio tests are available for fixed effects though it requires experience or simulation to determine whether the sample size m is large enough for the null χ^2 distribution to be accurate.

In terms of the random effects, one estimator is

$$\mathrm{E}[\boldsymbol{b}_i \mid \boldsymbol{y}] = \frac{\int_{\boldsymbol{b}_i} \boldsymbol{b}_i p(\boldsymbol{y} \mid \boldsymbol{b}_i) p(\boldsymbol{b}_i \mid \boldsymbol{D})\, d\boldsymbol{b}_i}{\int_{\boldsymbol{b}_i} p(\boldsymbol{y} \mid \boldsymbol{b}_i) p(\boldsymbol{b}_i \mid \boldsymbol{D})\, d\boldsymbol{b}_i}.$$

Unless the first stage is normal, the integrals in numerator and denominator will not be analytically tractable, though Laplace approximations or adaptive Gauss–Hermite may be used. In practice, empirical Bayes estimators, $\mathrm{E}[\boldsymbol{b}_i \mid \boldsymbol{y}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}]$, are used.

Example: Seizure Data

Recall that Y_{ij} is the number of seizures on patient i during period j, $j = 0, 1, 2, 3, 4$, and T_j is the observation period during period j, $j = 0, 1, 2, 3, 4$ with $T_0 = 8$ weeks and $T_j = 2$ weeks for $j = 1, \ldots, 4$. It is clear from Fig. 9.2 that there is considerable between-patient variability in the level of seizures, which suggests that a random effects model should include at least random intercepts. A random intercepts GLMM for the seizure data is:

Stage One: $Y_{ij} \mid \boldsymbol{\beta}, b_i \sim_{ind} \text{Poisson}(\mu_{ij})$, with

$$g(\mu_{ij}) = \log \mu_{ij} = \log T_{ij} + \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i,$$

and where $\boldsymbol{x}_{ij}$ is the design matrix for individual i at period j, with associated fixed effect $\boldsymbol{\beta}$. A particular model will be discussed shortly. The first two-conditional moments are

$$\begin{aligned} \mathrm{E}[Y_{ij} \mid b_i] &= \mu_{ij} = T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i), \\ \mathrm{var}(Y_{ij} \mid b_i) &= \mu_{ij}. \end{aligned}$$

Table 9.3 Parameter interpretation for the model defined by (9.3)

Group	Period 0	Period 1,2,3,4
Placebo	$\exp(\beta_0)$	$\exp(\beta_0 + \beta_2)$
Progabide	$\exp(\beta_0 + \beta_1)$	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)$

Stage Two: $b_i \mid \sigma_0^2 \sim_{iid} \mathrm{N}(0, \sigma_0^2)$.

Writing $\boldsymbol{\alpha} = \sigma_0^2$, the likelihood is

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} \frac{\exp[-\mu_{ij}(b_i)]\mu_{ij}(b_i)^{y_{ij}}}{y_{ij}!} \\ &\qquad \times (2\pi\sigma_0^2)^{-1/2} \exp\left(-\frac{b_i^2}{2\sigma_0^2}\right) db_i \\ &= (2\pi\sigma_0^2)^{-m/2} \prod_{i=1}^{m} \exp\left(\sum_{j=1}^{n_i} y_{ij} x_{ij}\boldsymbol{\beta}\right) \\ &\qquad \int \exp\left(-\mathrm{e}^{b_i} \sum_{j=1}^{n_i} \mathrm{e}^{\boldsymbol{x}_{ij}\boldsymbol{\beta}} + \sum_{j=1}^{n_i} y_{ij} b_i - \frac{b_i^2}{2\sigma_0^2}\right) db_i. \end{aligned}$$

The latter integral is analytically intractable. A Laplace approximation would expand each of the m integrands about the maximizing value of $\boldsymbol{b}_i$, or, alternatively, numerical integration can be used, for example, using adaptive Gauss–Hermite.

Let $x_{1i} = 0/1$ if patient i was assigned placebo/progabide, $x_{2j} = 0/1$ if $j = 0/1, 2, 3, 4$, and $x_{ij3} = x_{1i} \times x_{2j}$ for $j = 0, 1, 2, 3, 4$. Therefore, x_1 is a treatment indicator, x_2 is an indicator of pre-/post-baseline, and x_3 takes the value 1 for progabide individuals who are post-baseline and is zero otherwise. The first model we fit is

$$\boldsymbol{x}_{ij}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2j} + \beta_3 x_{3ij}, \tag{9.3}$$

so that $\boldsymbol{x}_{ij}$ is 1×4 and $\boldsymbol{\beta}$ is 4×1. Table 9.3 summarizes the form of the model across groups and periods.

We first provide an interpretation from a conditional perspective. In the following interpretation, a "typical" patient corresponds to a patient whose random effect is zero, that is, $b = 0$. On the more interpretable rate scale:

- $\exp(\beta_0)$ is the rate of seizures for a typical individual under placebo in time period 0.
- $\exp(\beta_1)$ is the ratio of the seizure rate of a typical individual under progabide to a typical individual under placebo, in time period 0. If the groups are comparable

at the time of treatment assignment and there are no other corrupting factors, we would expect this parameter to be estimated as close to 1.
- $\exp(\beta_2)$ is the ratio of the seizure rate post-baseline (T_j, $j = 1, 2, 3, 4$) as compared to baseline (T_0), for a typical individual in the placebo group.
- $\exp(\beta_3)$ is the ratio of the seizure rate for a typical individual in the progabide group post-baseline, as compared to a typical individual in the placebo group in the same period. Hence, $\exp(\beta_3)$ is the rate ratio parameter of interest.

Alternatively, we may interpret these rates and ratios of rates as being between two individuals with the same baseline rate of seizures (i.e., the same random effect b) prior to treatment assignment.

We now evaluate the implied *marginal model*. We recap the first two moments of a lognormal random variable. If $Z \sim \text{LogNorm}(\mu, \sigma^2)$, then

$$\begin{aligned} \text{E}[Z] &= \exp(\mu + \sigma^2/2) \\ \text{var}(Z) &= \exp(2\mu + \sigma^2) \times [\exp(\sigma^2) - 1] \\ &= \text{E}[Z]^2 \times [\exp(\sigma^2) - 1]. \end{aligned}$$

Therefore, since $\exp(b_i) \sim \text{LN}(0, \sigma_0^2)$, the marginal mean is

$$\begin{aligned} \text{E}[Y_{ij}] &= \text{E}_{b_i}\left[\,\text{E}(Y_{ij} \mid b_i)\,\right] \\ &= T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})\text{E}_{b_i}\left[\,\exp(b_i)\,\right] \\ &= T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \sigma_0^2/2). \end{aligned}$$

Consequently, for this model, relative rates $\exp(\beta_k)$, $k = 1, 2, 3$ (which, recall, are ratios) have a marginal interpretation, since the $\exp(\sigma_0^2/2)$ terms cancel in numerator and denominator (under the model). For example, $\exp(\beta_1)$ is the ratio of the average seizure rate in the progabide group to the average rate in the placebo group, in time period 0. Further discussion of parameter interpretation in marginal and conditional models is provided in Sect. 9.11. The marginal variance is

$$\begin{aligned} \text{var}(Y_{ij}) &= \text{E}_{b_i}[\,\text{var}(Y_{ij} \mid b_i)] + \text{var}_{b_i}[\text{E}(Y_{ij} \mid b_i)\,] \\ &= \text{E}_{b_i}[\,T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)\,] + \text{var}_{b_i}[\,T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)\,] \\ &= \text{E}[Y_{ij}][1 + \text{E}(Y_{ij})][\exp(\sigma_0^2) - 1)\,] \\ &= \text{E}[Y_{ij}][1 + \text{E}(Y_{ij}) \times \kappa] \end{aligned}$$

where

$$\kappa = \exp(\sigma_0^2) - 1 > 0$$

illustrating quadratic excess-Poisson variation which increases as σ_0^2 increases.

The marginal covariance between observations on the same individual is

$$\begin{aligned}\text{cov}(Y_{ij}, Y_{ik}) &= \text{cov}\left[T_{ij}\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i), T_{ij}\exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + b_i)\right] \\ &= T_{ij}T_{ik}\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{x}_{ik}\boldsymbol{\beta}) \times \exp(\sigma_0^2)[\exp(\sigma_0^2) - 1] \\ &= \text{E}[Y_{ij}]\text{E}[Y_{ik}]\kappa.\end{aligned}$$

To summarize, for individual i, the variance–covariance matrix is

$$\begin{bmatrix} \mu_{i1} + \mu_{i1}^2\kappa & \mu_{i1}\mu_{i2}\kappa & \mu_{i1}\mu_{i3}\kappa & \mu_{i1}\mu_{i4}\kappa \\ \mu_{i2}\mu_{i1}\kappa & \mu_{i2} + \mu_{i2}^2\kappa & \mu_{i2}\mu_{i3}\kappa & \mu_{i2}\mu_{i4}\kappa \\ \mu_{i3}\mu_{i1}\kappa & \mu_{i3}\mu_{i2}\kappa & \mu_{i3} + \mu_{i3}^2\kappa & \mu_{i3}\mu_{i4}\kappa \\ \mu_{i4}\mu_{i1}\kappa & \mu_{i4}\mu_{i2}\kappa & \mu_{i4}\mu_{i2}\kappa & \mu_{i4} + \mu_{i4}^2\kappa \end{bmatrix}.$$

For a random intercepts only LMM the marginal correlation is constant within a unit with correlation $\sigma_0^2/(\sigma_0^2 + \sigma_\epsilon^2)$, regardless of μ_{ij}, μ_{ik}. In contrast, for the Poisson random intercepts mixed model, the marginal correlation is

$$\begin{aligned}\text{corr}(Y_{ij}, Y_{ik}) &= \frac{\kappa\sqrt{\mu_{ij}\mu_{ik}}}{\sqrt{(1 + \kappa\mu_{ij})(1 + \kappa\mu_{ik})}} \\ &= \left[1 + \frac{1}{\kappa}\left(\frac{1}{\mu_{ij}} + \frac{1}{\mu_{ik}}\right) + \frac{1}{\kappa^2}\frac{1}{\mu_{ij}\mu_{ik}}\right]^{-1/2} \qquad (9.4)\end{aligned}$$

so that correlations vary in a complicated fashion as a function of the mean responses. However, the correlations increase as κ increases and as the means increase. A deficiency of this model is that we have only a single parameter (σ_0^2) to control both excess-Poisson variability and the strength of dependence over time. We address this in Sect. 9.6 by adding a second random effect to the model. For observations on different individuals, $\text{cov}(Y_{ij}, Y_{i'k}) = 0$ for $i \neq i'$.

Using a Laplace approximation to evaluate the integrals that define the likelihood, we obtain the estimates and standard errors given in Table 9.4. An alternative approach using Gauss–Hermite with 50 points to evaluate the integrals gave the same answers, so we conclude that the Laplace approximation is accurate in this example.

In terms of the parameter of interest β_3, there is an estimated drop in the seizure rates of 10% in the progabide group as compared to placebo, but this drop is not significant when assessed using a likelihood ratio test under conventional significance levels. The estimated value of β_1 indicates that the placebo and progabide groups are comparable at baseline, though the value of β_2 and its standard error suggest there is some evidence that the rate of seizures increased in the placebo group after randomization.

The random intercepts standard deviation is estimated as $\widehat{\sigma}_0^2 = 0.61$ to give $\widehat{\kappa} = 0.84$. For an individual whose rate of seizures is constant over the study period

Table 9.4 MLEs and standard errors from a generalized linear mixed model fit to the seizure data

	Estimate	Std. err.
β_0	1.03	0.15
β_1	−0.024	0.21
β_2	0.11	0.047
β_3	−0.10	0.065
σ_0	0.78	–

Parameter meaning: β_0 is the log baseline seizure rate in the placebo group for a typical individual; β_1 is the log of the ratio of seizure rates between typical individuals in the progabide and placebo groups, at baseline; β_2 is the log of the ratio of seizure rates of typical individuals in the post-baseline and baseline placebo groups; β_3 is the log of the ratio of the seizure rate for a typical individual in the progabide group as compared to a typical individual in the placebo group, post-baseline; σ_0 is the standard deviation of the random intercepts

at levels $\mu_{ij} = \mu_{ik} = 1, 2, 5$, the correlations between responses on this individual, from (9.4), are estimated as $0.46, 0.63, 0.81$. We conclude that the correlations are appreciable.

9.5 Conditional Likelihood Inference for Generalized Linear Mixed Models

An alternative approach to estimation in the GLMM is provided by conditional likelihood (Sect. 2.4.2). The basic idea is to split the data into components $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ in such a way that $\boldsymbol{t}_1$ contains information on parameter of interests, while $\boldsymbol{t}_2$ contains information primarily on nuisance parameters. In a GLMM setting, the aim is to condition on a part of the data that eliminates the random effects, hence avoiding both the need for their estimation and the need to specify their distribution. A consequence of the conditioning is that we also eliminate all regression coefficients in the model that are associated with covariates that are constant within an individual.

We now work through the details and assume a discrete GLMM and a canonical link function so that

$$g(\mu_{ij}) = \theta_{ij} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{ij}^{\mathrm{T}} + \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{z}_{ij}^{\mathrm{T}}.$$

We further assume $\alpha = 1$, as is true for Poisson and binomial models. Viewing both $\boldsymbol{\beta}$ and $\boldsymbol{b}$ as fixed effects gives, from (9.1),

$$\Pr(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{b}) \propto \exp\left[\boldsymbol{\beta}^{\mathrm{T}} \sum_{i=1}^{m}\sum_{j=1}^{n_i} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij} + \sum_{i=1}^{m} \boldsymbol{b}_i^{\mathrm{T}} \sum_{j=1}^{n_i} \boldsymbol{z}_{ij}^{\mathrm{T}} y_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n_i} b(\theta_{ij})\right], \tag{9.5}$$

where $b'(\theta_{ij}) = \mathrm{E}[Y_{ij} \mid \boldsymbol{b}_i]$. Define

$$\boldsymbol{t}_{1i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij}$$

$$\boldsymbol{t}_{2i} = \sum_{j=1}^{n_i} \boldsymbol{z}_{ij}^{\mathrm{T}} y_{ij},$$

and let $\boldsymbol{t}_1 = [\boldsymbol{t}_{11}, \ldots, \boldsymbol{t}_{1m}]^{\mathrm{T}}$ and $\boldsymbol{t}_2 = [\boldsymbol{t}_{21}, \ldots, \boldsymbol{t}_{2m}]^{\mathrm{T}}$ so that $\boldsymbol{t}_1$ and $\boldsymbol{t}_{2i}$ are sufficient statistics for $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, respectively. Conditioning on the sufficient statistics for $\boldsymbol{b}_i$, we obtain

$$\begin{aligned}\Pr\left(\boldsymbol{y}_i \mid \sum_{j=1}^{n_i} \boldsymbol{z}_{ij}^{\mathrm{T}} Y_{ij} = \boldsymbol{t}_{2i}, \boldsymbol{\beta}\right) &= \frac{\Pr\left(\boldsymbol{y}_i, \sum_{j=1}^{n_i} \boldsymbol{z}_{ij}^{\mathrm{T}} Y_{ij} = \boldsymbol{t}_{2i} \mid \boldsymbol{\beta}, \boldsymbol{b}_i\right)}{\Pr\left(\sum_{j=1}^{n_i} \boldsymbol{z}_{ij}^{\mathrm{T}} Y_{ij} = \boldsymbol{t}_{2i} \mid \boldsymbol{\beta}, \boldsymbol{b}_i\right)} \\ &= \frac{\sum_{S_{1i}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{t}_{1i} + \boldsymbol{b}_i \boldsymbol{t}_{2i})}{\sum_{S_{2i}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij} + \boldsymbol{b}_i \boldsymbol{t}_{2i})},\end{aligned}$$

so that the conditional likelihood is

$$L_c(\boldsymbol{\beta}) = \frac{\sum_{S_{1i}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{t}_{1i})}{\sum_{S_{2i}} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij})},$$

where

$$S_{1i} = \left\{ \boldsymbol{y}_i \mid \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij} = \boldsymbol{t}_{1i}, \sum_{j=1}^{n_i} \boldsymbol{z}_{ij}^{\mathrm{T}} y_{ij} = \boldsymbol{t}_{2i} \right\}$$

$$S_{2i} = \left\{ \boldsymbol{y}_i \mid \sum_{j=1}^{n_i} \boldsymbol{z}_{ij}^{\mathrm{T}} y_{ij} = \boldsymbol{t}_{2i} \right\}.$$

The set S_{1i} denotes the possible outcomes Y_{ij}, $j = 1, \ldots, n_i$ that are consistent with $\boldsymbol{t}_{1i}$ and $\boldsymbol{t}_{2i}$, given $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. The conditional MLE has the usual properties of an MLE. In particular, under regularity conditions, it is consistent and asymptotically normally distributed with the variance–covariance matrix determined from the second derivatives of the conditional log-likelihood.

The conditional likelihood approach allows the specification of a model (via the parameters $\boldsymbol{b}_i$) to acknowledge dependence but eliminates these parameters from the model. We emphasize that no distribution has been specified for the $\boldsymbol{b}_i$, as they have been viewed as fixed effects. Depending on the structure of $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$, some of the $\boldsymbol{\beta}$ parameters may be eliminated from the model. For example, if $\boldsymbol{x}_{ij} = \boldsymbol{z}_{ij}$, the collections S_{1i} and S_{2i} coincide and the complete $\boldsymbol{\beta}$ vector would be conditioned away.

Example: Seizure Data

We derive the conditional likelihood in this example, using the random intercepts only model so that $\boldsymbol{z}_{ij}\boldsymbol{b}_i = b_i$. The loglinear random intercept model is

$$\begin{aligned}\log \mathrm{E}[Y_{ij} \mid \boldsymbol{\beta}^\star, \lambda_i] &= \log T_{ij} + \lambda_i + \beta_2 x_{2j} + \beta_3 x_{3ij} \\ &= \log T_{ij} + \lambda_i + \boldsymbol{x}_{ij}\boldsymbol{\beta}^\star \\ &= \log \mu_{ij}\end{aligned}$$

where $\boldsymbol{\beta}^\star = [\beta_2, \beta_3]^{\mathrm{T}}$ represents the regression coefficients that are not conditioned from the model (since they are associated with covariates that change within an individual), $\boldsymbol{x}_{ij} = [x_{2j}, x_{3ij}]$, and $\lambda_i = \beta_0 + \beta_1 x_{1i} + b_i$. We cannot estimate β_1 because the associated covariate x_{1i} is a treatment indicator and constant within an individual in this study; hence, it is eliminated from the model by the conditioning, along with b_i and β_0. This parameter is not a parameter of primary interest, however.

To derive the conditional likelihood, we first write $c_{1i}^{-1} = \prod_{j=0}^{4} y_{ij}!$, and then the joint distribution of the data for the ith individual is

$$\begin{aligned}p(\boldsymbol{y}_i \mid \boldsymbol{\beta}^\star, \lambda_i) &= c_{1i} \exp\left(\sum_{j=0}^{4} y_{ij} \log \mu_{ij} - \sum_{j=0}^{4} \mu_{ij}\right) \\ &= c_{1i} \exp\left[\lambda_i y_{i+} + \sum_{j=0}^{4} y_{ij}(\log T_{ij} + \boldsymbol{x}_{ij}\boldsymbol{\beta}^\star) - \mu_{i+}\right].\end{aligned}$$

In this case, the conditioning statistic is y_{i+}, and its distribution is straightforward to derive

$$Y_{i+} \mid \boldsymbol{\beta}^\star, \lambda_i \sim \text{Poisson}(\mu_{i+}).$$

Letting $c_{2i}^{-1} = y_{i+}!$, and recognizing that $\mu_{i+} = \exp(\lambda_i) \sum_{j=0}^{4} T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}^\star)$, gives

$$\begin{aligned}p(y_{i+} \mid \boldsymbol{\beta}^\star, \lambda_i) &= c_{2i} \prod_{i=1}^{m} \exp\left(-\mu_{i+} + y_{i+} \log \mu_{i+}\right) \\ &= c_{2i} \prod_{i=1}^{m} \exp\left[\lambda_i y_{i+} + y_{i+} \log\left(\sum_{j=0}^{4} T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}^\star)\right) - \mu_{i+}\right].\end{aligned}$$

Hence,

$$p(\boldsymbol{y}_i \mid \boldsymbol{y}_{i+}, \boldsymbol{\beta}^\star) = \frac{p(\boldsymbol{y}_i \mid \boldsymbol{\beta}^\star, \lambda_i)}{p(y_{i+} \mid \boldsymbol{\beta}^\star, \lambda_i)}$$

simplifies to

$$
\begin{aligned}
p(\boldsymbol{y}_i \mid \boldsymbol{y}_{i+}, \boldsymbol{\beta}^\star) &= \frac{c_{1i}}{c_{2i}} \exp\left[\sum_{j=0}^{4} y_{ij}(\log T_{ij} + \boldsymbol{x}_{ij}\boldsymbol{\beta}^\star) - y_{i+} \log\left(\sum_{j=0}^{4} T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}^\star)\right)\right] \\
&= \frac{c_{1i}}{c_{2i}} \prod_{j=0}^{4} \exp\left\{y_{ij}\left[\log T_{ij} + \boldsymbol{x}_{ij}\boldsymbol{\beta}^\star - \log\left(\sum_{j=0}^{4} T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}^\star)\right)\right]\right\} \\
&= \frac{y_{i+}!}{\prod_{j=0}^{4} y_{ij}!} \prod_{j=0}^{4} \left(\frac{T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}^\star)}{\sum_{l=0}^{4} T_{il} \exp(\boldsymbol{x}_{il}\boldsymbol{\beta}^\star)}\right)^{y_{ij}}
\end{aligned}
$$

which is a multinomial likelihood (we have conditioned a set of Poisson counts on their total so this is no surprise). More transparently,

$$
y_{ij} \mid y_{i+}, \boldsymbol{\beta}^\star \sim \text{Multinomial}_4(y_{i+}, \boldsymbol{\pi}_i)
$$

where $\boldsymbol{\pi}_i = [\pi_{i0}, \ldots, \pi_{i4}]^{\mathrm{T}}$ and

$$
\pi_{ij} = \frac{T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}^\star)}{\sum_{l=0}^{4} T_{il} \exp(\boldsymbol{x}_{il}\boldsymbol{\beta}^\star)}.
$$

Since $\boldsymbol{x}_{i1} = \boldsymbol{x}_{i2} = \boldsymbol{x}_{i3} = \boldsymbol{x}_{i4}$ and $T_{i0} = 8 = \sum_{j=1}^{4} T_{ij}$, we effectively have two observation periods of equal length. Letting $Y_i^\star = \sum_{j=1}^{4} Y_{ij}$,

$$
Y_i^\star \mid y_{i+}, \beta^\star \sim_{ind} \text{Binomial}(y_{i+}, \pi_i^\star)
$$

where the odds are such that

$$
\frac{\pi_i^\star}{1 - \pi_i^\star} = \begin{cases} \exp(\beta_2) & i = 1, \ldots, 28, \quad \text{placebo group} \\ \exp(\beta_2 + \beta_3) & i = 29, \ldots, 59, \quad \text{progabide group.} \end{cases}
$$

Hence, fitting can be simply performed using logistic regression. For the seizure data, the sum of the denominators are 1,825 and 1,967 for placebo and progabide with 963 and 987 total seizures in the post-treatment period. These values result in estimates (standard errors) of $\widehat{\beta}_2 = 0.11$ (0.047) and $\widehat{\beta}_3 = -0.10$ (0.065). The estimate suggests a positive effect of progabide, but the difference from zero is not significant. Performing Fisher's exact test (Sect. 7.7) makes little difference for these data since the counts are large.

The conditional likelihood approach is quite intuitive in this example and results in a two-period design in which each person is acting as their own control. Conditioning on the sum of the two counts results in a single outcome per patient and removes the need to confront the dependency issue.

9.6 Bayesian Inference for Generalized Linear Mixed Models

9.6.1 Model Formulation

A Bayesian approach to inference for a GLMM requires a prior distribution for $\boldsymbol{\beta}, \boldsymbol{\alpha}$. As with the linear mixed model (Sect. 8.6), a proper prior is required for the matrix $\boldsymbol{D}$. A proper prior is not always necessary for $\boldsymbol{\beta}$, but care is required. The exponential family and canonical link lead to a likelihood that is well behaved (in particular, with respect to tail behavior), though it is safer to specify a proper prior since impropriety of the posterior can occur in some cases (e.g., with noncanonical links or when counts are either equal to zero or to the denominator; see Sect. 6.8.1). As with the LMM, closed-form inference is unavailable, but MCMC (Sect. 3.8) is almost as straightforward as in the LMM, and the integrated nested Laplace approximation approach (Sect. 3.7.4) is also available though the approximation is not always accurate for the GLMM (Fong et al. 2010).

Let $\boldsymbol{W} = \boldsymbol{D}^{-1}$, and assume that there are no unknown scale parameters at stage one of the model (i.e., $\alpha = 1$), as is the case for binomial and Poisson models. The joint posterior is

$$p(\boldsymbol{\beta}, \boldsymbol{W}, \boldsymbol{b} \mid \boldsymbol{y}) \propto \prod_{i=1}^{m} [p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{b}_i) p(\boldsymbol{b}_i \mid \boldsymbol{W})] \, \pi(\boldsymbol{\beta}, \boldsymbol{W}).$$

We assume independent hyperpriors:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{N}_{q+1}(\boldsymbol{\beta}_0, \boldsymbol{V}_0) \\ \boldsymbol{W} &\sim \mathrm{Wish}_{q+1}(r, \boldsymbol{R}^{-1}) \end{aligned}$$

where $\mathrm{Wish}_{q+1}(r, \boldsymbol{R}^{-1})$ denotes a Wishart distribution of dimension $q + 1$ with degrees of freedom r and scale matrix $\boldsymbol{R}^{-1}$; see Sect. 8.6.2 for further discussion. The conditional distribution for $\boldsymbol{W}$ is unchanged from the LMM case. There are no closed-form conditional distributions for $\boldsymbol{\beta}$, or for $\boldsymbol{b}_i$, but if an MCMC approach is followed, Metropolis–Hastings steps can be used.

9.6.2 Hyperpriors

In a GLMM we can often specify priors for more meaningful parameters than the original elements of $\boldsymbol{\beta}$. For example, $\exp(\beta)$ is the relative risk/rate in a loglinear model and is the odds ratio in a logistic model. It is convenient to specify lognormal priors for a generic parameter $\theta > 0$, since one may specify two quantiles of the distribution, and directly solve for the two parameters of the prior. Denote

by LogNorm(μ, σ) the lognormal prior distribution for θ with $\mathrm{E}[\log\theta] = \mu$ and $\mathrm{var}(\log\theta) = \sigma^2$, and let θ_1 and θ_2 be the q_1 and q_2 quantiles of this prior. Then, (3.15) and (3.16) give the lognormal parameters. As an example, in a Poisson model, suppose we believe there is a 50% chance that the relative risk is less than 1 and a 95% chance that it is less than 5. With $q_1 = 0.5, \theta_1 = 1.0$ and $q_2 = 0.95, \theta_2 = 5.0$, we obtain lognormal parameters $\mu = 0$ and $\sigma = \log(5/1.96) = 0.98$.

Consider the random intercepts model with $b_i \mid \sigma_0^2 \sim_{iid} \mathrm{N}(0, \sigma_0^2)$. It is not straightforward to specify a prior for σ_0, which represents the standard deviation of the residuals on the linear predictor scale and is consequently not easy to interpret. We specify a gamma prior Ga(a, b) for the precision $\tau_0 = 1/\sigma_0^2$, with parameters a, b specified a priori. The choice of a gamma distribution is convenient since it produces a marginal distribution for the "residuals" in closed form. As discussed in Sect. 8.6.2, the marginal distribution for b_i is $t_d(0, \lambda^2)$, a Student's t distribution with $d = 2a$ degrees of freedom, location zero, and scale $\lambda^2 = b/a$. These summaries allow prior specification based on beliefs concerning the residuals on a natural scale.

As an example, consider a log link, in which case the above prior specification is equivalent to the residual relative risks following a log Student's t distribution. We specify the range $\exp(\pm V)$ within which we expect the residual relative risks to lie with probability q and use the relationship $\pm t^d_{q/2}\lambda = \pm V$, where t^d_q is the qth quantile of a Student's t random variable with d degrees of freedom, to give $a = d/2, b = V^2 d/2(t^d_{q/2})^2$. For example, if we assume a priori that the residual relative risks follow a log Student's t distribution with 2 degrees of freedom and that 95% of these risks fall in the interval [0.5,2.0], then we obtain the prior, Ga$(1, 0.0260)$. In terms of σ_0, this results in [2.5%,97.5%] quantiles of [0.084,1.01] with posterior median 0.19.

It is important to assess whether the prior allows all reasonable levels of variability in the residual relative risks, in particular, small values should not be excluded. The prior Ga(0.001,0.001), which has been widely used under the guise of being relatively non-informative, should be avoided for this reason. This prior corresponds to the relative risks following a log Student's t distribution with 0.002 degrees of freedom, so that the spread is enormous. For example, the 0.01 quantile for σ_0 is 6.4 so that it is unlikely a priori that the standard deviation is small.

Example: Seizure Data

For illustration, we consider three models for the seizure data:

Model 1: The conditional mean model we start with has stages one and two given by:

$$\begin{aligned} Y_{ij} \mid b_i &\sim_{ind} \text{Poisson}[T_{ij}\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)] \\ b_i \mid \sigma_0^2 &\sim_{iid} \mathrm{N}(0, \sigma_0^2). \end{aligned} \tag{9.6}$$

For a Bayesian analysis, we require priors for $\boldsymbol{\beta}$ and σ_0^2. In this and the following two models, we take the improper prior $\pi(\boldsymbol{\beta}) \propto 1$. We assume $\sigma_0^{-2} = \tau_0 \sim \text{Ga}(1, 0.260)$. This prior corresponds to a Student's t_2 distribution for the residual rates with a 95% prior interval of $[0.5, 2.0]$.

Model 2: We assume the same first and second stages as model 1 but address the sensitivity to the prior on τ_0. Specifically, we perturb the prior to $\tau_0 \sim \text{Ga}(2, 1.376)$, which corresponds to a Student's t_4 distribution for the residual rates with a 95% interval $[0.1, 10.0]$.

Model 3: As pointed out in Sect. 9.4, a Poisson mixed model with a single random effect has a single parameter σ_0 only to model excess-Poisson variability *and* within-individual dependence. Therefore, we introduce "measurement error" into the model via the introduction of an additional random effect in the linear predictor. To motivate this model, consider the random intercepts only LMM model:

$$\begin{aligned} \text{E}[Y_{ij} \mid b_i] &= \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i + \epsilon_{ij} \\ b_i \mid \sigma_0^2 &\sim \text{N}(0, \sigma_0^2) \\ \epsilon_{ij} \mid \sigma_\epsilon^2 &\sim \text{N}(0, \sigma_\epsilon^2), \end{aligned}$$

with b_i and ϵ_{ij} independent. By analogy, consider the model:

$$\begin{aligned} Y_{ij} \mid b_i, \epsilon_{ij} &\sim \text{Poisson}[T_{ij} \exp(\boldsymbol{x}_{ij}\beta + b_i + \epsilon_{ij})] \\ b_i \mid \sigma_0^2 &\sim \text{N}(0, \sigma_0^2) \\ \epsilon_{ij} \mid \sigma_\epsilon^2 &\sim \text{N}(0, \sigma_\epsilon^2) \end{aligned}$$

with b_i and ϵ_{ij} independent. There are now two parameters to allow for between-individual variability, σ_0^2, and within-individual variability, σ_ϵ^2 (with both producing excess-Poisson variability). Unfortunately, there is no simple marginal interpretation of σ_0^2 and σ_ϵ^2 since

$$\begin{aligned} \text{E}[Y_{ij}] &= \mu_{ij} = T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \sigma_0^2/2 + \sigma_\epsilon^2/2) \\ \text{var}(Y_{ij}) &= \mu_{ij}\{1 + \mu_{ij}[\exp(\sigma_0^2) - 1][\exp(\sigma_\epsilon^2) - 1]\} \\ \text{cov}(Y_{ij}, Y_{ik}) &= T_{ij}T_{ik} \exp[(\boldsymbol{x}_{ij} + \boldsymbol{x}_{ik})\boldsymbol{\beta}] \exp(\sigma_0^2)[\exp(\sigma_0^2) - 1]. \end{aligned}$$

The expression for the marginal covariance shows that σ_0^2 is controlling the within-individual dependence in the model, with large values giving high dependence. The expression for the marginal variance is quadratic in the mean and is controlled by both σ_0^2 and σ_ϵ^2, with large values corresponding to greater excess-Poisson variability. We assign independent priors $\sigma_0^{-2} \sim \text{Ga}(1, 0.260)$, $\sigma_\epsilon^{-2} \sim \text{Ga}(1, 0.260)$.

All three models were implemented using MCMC. Table 9.5 gives summaries for the three models. Model 1 gives very similar inference to the likelihood approach described in Sect. 9.4 (specifically, the result presented in Table 9.4), which is not surprising given the relatively large sample size and weak priors. Model 2 shows little sensitivity to the prior distribution on σ_0 which is again not surprising given

Table 9.5 Posterior means and standard deviations for Bayesian analyses of the seizure data

	Model 1		Model 2		Model 3	
	Estimate	Std. err.	Estimate	Std. err.	Estimate	Std. err.
β_0	1.03	0.16	1.04	0.16	1.04	0.18
β_1	−0.036	0.21	−0.030	0.22	0.062	0.25
β_2	0.11	0.047	0.11	0.047	0.0064	0.10
β_3	−0.10	0.065	−0.10	0.065	−0.29	0.14
σ_0	0.80	0.078	0.81	0.077	0.82	0.084
σ_ϵ	–	–	–	–	0.39	0.033

See the caption of Table 9.4 for details on parameter interpretation. Models 1 and 2 are standard GLMMs and differ only in the priors placed on σ_0 which is the standard deviation of the random intercepts. Model 3 adds an additional measurement error random effect, with standard deviation σ_ϵ

the number of individuals. Model 3 shows substantive differences, however. The parameter of interest β_3 is now greatly reduced, with a 95% credible interval for the rate being [0.56,0.99]. The reason for the change is that in the progabide group, there is a single individual (as seen in Fig. 9.2) who is very influential; this individual has counts of 151, 102, 65, 72 and 63 in the five time periods. The introduction of measurement error accommodates this individual. The posterior medians of ϵ_{ij} for this individual show a negative error term at baseline, followed by a run of positive terms post-baseline: −0.61, 0.61, 0.17, 0.27, 0.14. The difference in signs explains why the between-individual random effect cannot accommodate this individual's data. Notice also that β_2 (the log ratio of seizure rates in the post-baseline period relative to the baseline period, for typical individuals in the placebo group) is now close to zero, whereas in models 1 and 2, it is 0.11. This shows that the aberrant individual's measurements were responsible for the high value of β_2 in the first two models. The estimate for σ_ϵ is less than half the estimate for σ_0 so that between-individual variability is greater than within-individual variability for these data.

In analyses presented in Diggle et al. (2002), the influential individual was dropped, and in their Table 9.7, the single random effect analysis produced an estimate (standard error) of −0.30 (0.070), which is very similar to that for model 3. We would always prefer to not remove individuals from the analysis, however, unless there are substantive reasons to do so.

Another possibility for modeling excess-Poisson variability, by combining the Poisson likelihood with a gamma random effects distribution, is considered in Sect. 9.8. □

In the last example we saw that the introduction of normal random effects accounted for both measurement error and between-individual variability. This flexibility is a great benefit of the GLMM framework. One way of approaching modeling is to first imagine that the response is continuous and then decide upon a model that would be considered in this case. The same structure can then be assumed for the data at hand but on the linear predictor scale. In the next example, the versatility is further illustrated with a model for spatial dependence.

9.7 Generalized Linear Mixed Models with Spatial Dependence

9.7.1 A Markov Random Field Prior

The topic of modeling residual spatial dependence is vast and here we only scratch the surface and present a model that is popular in the spatial epidemiology literature, and fits within the GLMM framework. We first describe the model and then illustrate its use on the lung cancer and radon data of Sect. 1.3.3.

The following three-stage model was introduced by Besag et al. (1991) in the context of disease mapping:

Stage One: The distribution of the response in area i is

$$Y_i \mid \mu_i, \epsilon_i, S_i \sim_{ind} \text{Poisson}[E_i \mu_i \exp(\epsilon_i + S_i)]$$

with loglinear mean model

$$\log \mu_i = \beta_0 + \beta_i x_i, \tag{9.7}$$

where x_i is the radon level in area i. The random effects ϵ_i and S_i represent error terms without and with spatial structure, respectively. We have already encountered the nonspatial version when a Poisson-Gamma model was described for these data in Chap. 6. There are many models one might envision for the spatial terms S_i, $i = 1, \ldots, m$. An obvious isotropic form would be $\boldsymbol{S} = [S_1, \ldots, S_m]^{\mathrm{T}} \sim \mathrm{N}_m(\boldsymbol{0}, \sigma_s^2 \boldsymbol{R})$ with $\boldsymbol{R}$ a correlation matrix with $R_{ii'}$ describing the correlation between areas i and i', $i, i' = 1, \ldots, m$. A common form is $R_{ii'} = \rho^{d_{ii'}}$ where $d_{ii'}$ is the distance between the centroids of areas i and i'. We have already seen this form of correlation in the context of longitudinal data; see in particular (8.14).

Marginally, this model gives

$$\begin{aligned}
\mathrm{E}[Y_i] &= E_i \mu_i \exp(\sigma_\epsilon^2/2 + \sigma_s^2/2) \\
\mathrm{var}(Y_i) &= \mathrm{E}[Y_i] \left\{1 + \mathrm{E}[Y_i][\exp(\sigma_\epsilon^2) - 1][\exp(\sigma_s^2) - 1)]\right\} \\
\mathrm{cov}(Y_i, Y_{i'}) &= E_i \mu_i E_{i'} \mu_{i'} \exp(\sigma_s^2)[\exp(\sigma_s^2) - 1].
\end{aligned}$$

This *isotropic* model is computationally expensive within an MCMC scheme because we need to invert $\boldsymbol{R}$ at each iteration to obtain the conditional distribution. We describe an alternative which is both computationally feasible and statistically appealing.

Stage Two: The random effects distributions are

$$\epsilon_i \mid \sigma_\epsilon^2 \sim_{iid} \mathrm{N}(0, \sigma_\epsilon^2) \tag{9.8}$$

$$S_i \mid S_{i'}, i' \in \mathrm{ne}(i), \sigma_s^2 \sim_{ind} \mathrm{N}\left(\overline{S}_i, \frac{\sigma_s^2}{n_i}\right) \tag{9.9}$$

where $\overline{S}_i = \frac{1}{n_i} \sum_{i' \in \mathrm{ne}(i)} S_{i'}$ is the mean of the "neighbors" of area i, with $\mathrm{ne}(i)$ defining the set of, and n_i the number of, such neighbors. This *intrinsic conditional autoregressive* (ICAR) model is very appealing since it provides local spatial smoothing and may be viewed as providing stochastic interpolation (Besag and Kooperberg 1995). A common definition (which we adopt in the example at the end of this section) is that two areas are neighbors if they share a common boundary. In non-lattice systems, this is clearly ad hoc.

An interesting aspect of this model is that the joint distribution is undefined. The form of the joint "density" is

$$p(\boldsymbol{s} \mid \sigma_s^2) \propto \sigma_s^{-(m-r)} \exp\left[-\frac{1}{2\sigma_s^2} \sum_{i<i'} W_{ii'}(s_i - s_{i'})^2\right], \tag{9.10}$$

where $W_{ii'} = 1$ if areas i and i' are neighbors and $W_{ii'} = 0$ otherwise. In the spatial context, r is the number of connected regions. So if $r = 1$, there are no collection of areas that are not neighbors of the remaining areas, which means that we cannot break the study region into collections of areas that are unconnected. One way of thinking about this model is that it specifies a prior on the differences between levels in different areas but not on the overall level.

There are two equivalent representations of model (9.10) that are commonly used. In one approach, the intercept β_0 is removed from the mean model (9.7), while in the other, we allow an intercept β_0, along with an improper uniform prior for this parameter, and then constrain $\overline{\boldsymbol{S}} = 0$. In the following we assume that the intercept has been excluded from the model. See Besag and Kooperberg (1995) and Rue and Held (2005) for further discussion of this model.

Stage Three: Hyperpriors:

$$\begin{aligned}
\beta_1 &\sim \mathrm{N}(\mu_\beta, \Sigma_\beta) \\
\sigma_\epsilon^{-2} &\sim \mathrm{Gamma}(a_\epsilon, b_\epsilon) \\
\sigma_s^{-2} &\sim \mathrm{Gamma}(a_s, b_s).
\end{aligned}$$

9.7.2 Hyperpriors

Picking a prior for σ_s is not straightforward because of its interpretation as the *conditional* standard deviation. In particular, σ_s and σ_ϵ are not directly comparable since the latter has a marginal interpretation (on the log relative risk scale).

We describe how to simulate realizations from (9.10) to examine candidate prior distributions. As already noted, due to the rank deficiency, (9.10) does not define a probability density, and so we cannot directly simulate from this prior. We need to define some new notation in order to describe the method of simulation. The model can be written in the form

$$p(\boldsymbol{s} \mid \sigma_s^2) = (2\pi)^{-(m-r)/2} |\boldsymbol{Q}^\star|^{1/2} \sigma_s^{-(m-r)} \exp\left(-\frac{1}{2\sigma_s^2} \boldsymbol{s}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{s}\right) \tag{9.11}$$

where $\boldsymbol{s} = [s_1, \ldots, s_m]$ is the collection of random effects, $\boldsymbol{Q}$ is a (scaled) "precision" matrix of rank $m - r$, with

$$Q_{ij} = \sigma_s^{-2} \begin{cases} n_i & \text{if } i = j \\ -1 & \text{if } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases}$$

and $|\boldsymbol{Q}^\star|$ is a generalized determinant which is the product over the $m - r$ nonzero eigenvalues of $\boldsymbol{Q}$.

Rue and Held (2005) give the following algorithm for generating samples from (9.11):

1. Simulate $z_j \sim \mathrm{N}(0, \lambda_j^{-1})$, for $j = m-r+1, \ldots, m$, where λ_j are the eigenvalues of $\boldsymbol{Q}$ (recall there are $m - r$ nonzero eigenvalues as $\boldsymbol{Q}$ has rank $m - r$).
2. Return $\boldsymbol{s} = z_{m-r+1}\boldsymbol{e}_{n-r+1} + z_3\boldsymbol{e}_3 + \ldots + z_n\boldsymbol{e}_m = \boldsymbol{E}\boldsymbol{z}$ where $\boldsymbol{e}_j$ are the corresponding eigenvectors of $\boldsymbol{Q}$, $\boldsymbol{E}$ is the $m \times (m - r)$ matrix with these eigenvectors as columns, and $\boldsymbol{z}$ is the $(m - r) \times 1$ vector containing z_j, $j = m - r + 1, \ldots, m$.

The simulation algorithm is conditioned so that samples are zero in the null-space of $\boldsymbol{Q}$. If $\boldsymbol{s}$ is a sample and the null-space is spanned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, then $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{v}_1 = \boldsymbol{s}^{\mathrm{T}}\boldsymbol{v}_2 = \boldsymbol{0}$. For example, suppose $\boldsymbol{Q}\boldsymbol{1} = \boldsymbol{0}$ so that the null-space is spanned by $\boldsymbol{1}$ and the rank deficiency is 1. Then $\boldsymbol{Q}$ is of rank $m - 1$, since the eigenvalue corresponding to $\boldsymbol{1}$ is zero, and samples $\boldsymbol{s}$ produced by the algorithm are such that $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{1} = \boldsymbol{0}$. It is also useful to note that if we wish to compute the marginal variances, only then simulation is not required, as they are available as the diagonal elements of the matrix $\sum_j \lambda_j^{-1} \boldsymbol{e}_j \boldsymbol{e}_j^T$.

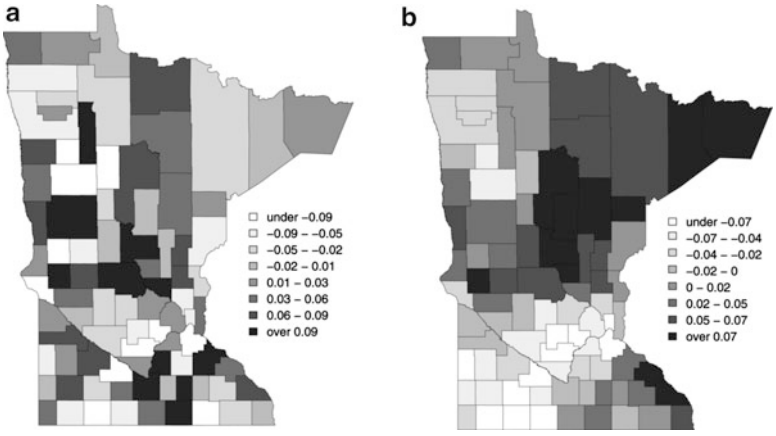

Fig. 9.5 (**a**) Nonspatial and (**b**) spatial random effects for the Minnesota lung cancer data

Example: Lung Cancer and Radon

We apply the Poisson model with nonspatial and spatial normal random effects, that is, the model given by (9.8) and (9.9). We note that model (9.7) does not aggregate correctly from a plausible individual-level model; see Wakefield (2007b) and the discussion leading to model (6.19). The prior on β_1 is $\mathrm{N}(0, 1.17^2)$ which gives a 95% interval for the relative risk of [0.1,10].

The priors on σ_ϵ^2 and σ_s^2 require more care, but we would like to specify priors in such a way that the nonspatial and spatial contributions are approximately equal. This is complicated by σ_s^2 having a conditional interpretation, as just discussed. We specify gamma priors for each of the precisions, σ_ϵ^{-2} and σ_s^{-2}. To make the priors compatible, we first specify a prior for σ_ϵ^{-2} and evaluate the average of the marginal variances over the 87 areas, when $\sigma_s^2 = 1$, as described at the end of Sect. 9.7.2. We then match up the means of the gamma distributions. Following the development of Sect. 9.6.2 for the unstructured variability, we assume that the unstructured residual relative risks lie in the interval [0.2, 5] with probability 0.95 and assume $d = 2$ to give the exponential prior distribution Ga(1,0.140) for σ_ϵ^{-2}. The average of the marginal variances over the study region for the spatial random effects is 0.21; hence, the average of the marginal precisions is approximately 1/0.21. The prior for σ_s^{-2} is therefore Ga(0.21,0.140), to give $\mathrm{E}[\sigma_s^{-2}] = 0.21 \times \mathrm{E}[\sigma_\epsilon^{-2}]$.

The fitting of this model (using INLA) results in the posterior mean estimates $\widehat{\epsilon}_i$ and $\widehat{S}_i$ mapped in Fig. 9.5(a) and (b) respectively. Notice that the scale is narrower in panel (b), since the spatial contribution to the residuals is relatively small here, though the spatial pattern in these residuals is apparent. As we discussed with respect to prior specification, the variances σ_ϵ^2 and σ_s^2 are not directly comparable,

Table 9.6 Parameter estimates for β_1, the area-level log relative risk corresponding to radon, and measures of uncertainty (standard errors and posterior standard deviations) under various models, for the Minnesota lung cancer data

Model	Estimate ($\times 10^2$)	Uncertainty ($\times 10^3$)
Poisson	-3.6	5.4
Quasi-likelihood	-3.6	8.8
Negative binomial	-2.9	8.2
Nonspatial random effects	-2.8	9.1
Nonspatial and ICAR random effects	-2.8	9.7

and so we calculate an approximate proportion of the total residual variance that is spatial by comparing σ_ϵ^2 with an empirical estimate of the marginal variance of the collection of random effects $\{\widehat{S}_i, i = 1, \ldots, m\}$. Specifically, we calculate

$$\frac{\text{var}(\widehat{S}_i)}{\text{var}(\widehat{S}_i) + \widehat{\sigma}_\epsilon^2}$$

where $\text{var}(\widehat{S}_i)$ is the empirical variance of the random effects and $\widehat{\sigma}_\epsilon^2$ is the posterior median. From this calculation, the fraction of the total residual variability that is attributed to the spatial component is 0.13.

Table 9.6 provides estimates and standard error/posterior standard deviations for the log relative risk associated with a unit increase in radon, for a variety of models. We include a model with nonspatial normal random effects only. The Poisson and quasi-likelihood methods assume the same form of (proportional) mean–variance relationship, while the negative binomial and nonspatial normal random effects approaches imply a variance that is quadratic in the mean. The marginal variance does not exist under the improper spatial model, but here the spatial contributions are small. We might therefore expect to see similar conclusions to the negative binomial and nonspatial normal random effects models. This is borne out in the table, with the last three models giving similar estimates that are closer to zero than the first two models. The standard error from the spatial model does increase a little over the nonspatial random effects model.

In general, if strong spatial effects are present and the exposure surface has spatial structure, then when spatial random effects are added to a model, large changes may be seen in the regression coefficient associated with exposure. This phenomenon, which is sometimes known as *confounding by location*, is a big practical headache since it is difficult to decide on whether to attribute spatial variability in risk to the exposure or to the spatial random effects (which may be acting as surrogates for unmeasured confounders). Wakefield (2007b) and Hodges and Reich (2010) provide further discussion.

9.8 Conjugate Random Effects Models

An obvious approach to extending models for independent data is to assume a random effects distribution that is conjugate to the likelihood. We illustrate this approach, and its shortcomings, through two examples.

Example: Lung Cancer and Radon

A Poisson-Gamma conjugate model was fitted to the lung cancer/radon data in Sect. 6.9 with:

Stage One: $Y_i \mid \mu_i, \delta_i \sim_{ind} \text{Poisson}(E_i\delta_i)$, with $\log \mu_i = \beta_0 + \beta_1 x_i$, for $i = 1, \ldots, m$.

Stage Two: $\delta_i \mid b \sim_{iid} \text{Gamma}(b, b)$ for $i = 1, \ldots, m$.

The advantage of this model is that the random effects can be analytically integrated from the model to give $Y_i \mid \mu_i, b \sim_{ind} \text{NegBin}(\mu_i, b)$, $i = 1, \ldots, m$. However, the extension to allow spatial dependence is not obvious, unless one introduces normal random effects, as in the last section.

Example: Seizure Data

Letting $\mu_{ij} = T_{ij} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})$, consider the two-stage model:

$$Y_{ij} \mid \mu_{ij}, \xi_{ij} \sim_{ind} \text{Poisson}(\mu_{ij}\xi_{ij})$$
$$\xi_{ij} \mid b \sim_{iid} \text{Ga}(b, b).$$

This results in $Y_{ij} \mid \mu_i, b \sim_{ind} \text{NegBin}(\mu_i, b)$ with $\text{E}[Y_i] = \mu_{ij}$ and $\text{var}(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}/b)$. This model allows for excess-Poisson variability but not for dependence of observations on the same patient. The introduction of patient-specific random effects allows for the latter but loses the analytical tractability. Specifically, the two-stage model

$$Y_{ij} \mid \mu_{ij}, \delta_i \sim_{ind} \text{Poisson}(\mu_{ij}\delta_i)$$
$$\delta_i \mid b \sim_{iid} \text{Ga}(b, b)$$

leads to a marginal model for the data of the ith individual of

$$\Pr(y_{i0}, \ldots, y_{i4} \mid \mu_{ij}, b) = \left(\prod_{j=0}^{4} \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \right) \frac{b^b}{\Gamma(b)} \frac{\Gamma(b + y_{i+})}{(b + \mu_{i+})^{b+y_{i+}}},$$

which is not of negative binomial form.

9.9 Generalized Estimating Equations for Generalized Linear Models

The GEE approach was described in Sect. 8.7 for linear models. The extension to GLM mean models is conceptually straightforward, since all that is required is specification of a mean model and a working covariance model. The mean is

$$g(\mu_{ij}) = \boldsymbol{x}\boldsymbol{\gamma}$$

where $\mu_{ij} = \mathrm{E}[Y_{ij}]$, $g(\cdot)$ is a link function, $\boldsymbol{x}$ is a $n \times (k+1)$ design matrix, and $\boldsymbol{\gamma}$ is a $(k+1) \times 1$ vector. We use $\boldsymbol{\gamma}$ to denote the parameters in the *marginal* mean model to distinguish them from the parameters $\boldsymbol{\beta}$ which have been used to represent the mixed model *conditional* parameters. The working covariance matrix is

$$\mathrm{var}(\boldsymbol{Y}) = \boldsymbol{W}.$$

and in a GLM setting, $\boldsymbol{W}$ will usually depend on $\boldsymbol{\gamma}$ and on additional parameters $\boldsymbol{\alpha}$ so that $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{\gamma}, \boldsymbol{\alpha})$. Suppose $\widehat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. Then, GEE takes the estimator $\widehat{\boldsymbol{\gamma}}$ that satisfies

$$\boldsymbol{G}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) = \boldsymbol{0},$$

where $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\gamma}$ is $n_i \times (k+1)$ and $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\gamma}, \widehat{\boldsymbol{\alpha}})$ is the $n_i \times n_i$ working covariance model for unit i, $i = 1, \ldots, m$. The estimator $\widehat{\boldsymbol{\gamma}}$ will not be of closed form, unless the link is linear. Under mild regularity conditions,

$$\boldsymbol{V}_{\boldsymbol{\gamma}}^{-1/2} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \rightarrow_d \mathrm{N}_{k+1}(\, \boldsymbol{0}, \mathbf{I}_{k+1}\,),$$

where $\boldsymbol{V}_{\boldsymbol{\gamma}}$ takes the sandwich form

$$\left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1} \left[\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{cov}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right] \left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1}. \tag{9.12}$$

In practice, an empirical estimator of $\mathrm{cov}(\boldsymbol{Y}_i)$ is substituted to obtain $\widehat{\boldsymbol{V}}_{\boldsymbol{\gamma}}$. This produces a consistent estimator of the standard error of $\widehat{\boldsymbol{\gamma}}$, so long as we have independence between units $i \neq i'$, $i, i' = 1, \ldots, m$. For small m, the variance estimator may be unstable, however.

As in the linear case, various assumptions about the form of the working covariance are available. We write

$$\boldsymbol{W}_i = \boldsymbol{\Delta}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{\Delta}_i^{1/2},$$

where $\boldsymbol{\Delta}_i = \text{diag}[\text{var}(Y_{i1}), \ldots, \text{var}(Y_{in_i})]^{\text{T}}$ and $\boldsymbol{R}_i$ is a working correlation model. Common choices include independence, exchangeable, AR(1), and unstructured. For discrete data, there is often no natural choice since, in this setting, the correlation is not an intuitive measure of dependence.

For small m, the sandwich estimator will have high variability, and so model-based variance estimators may be preferable (and we would probably not rely on asymptotic normality if m were small anyway). Model-based estimators are more efficient if the model is correct and efficiency will be improved if we can pick a working correlation matrix that is close to the true structure.

Published comments on whether to assume working independence or a more complex form are a little in conflict: Liang and Zeger (1986) state that there is "little difference when correlation is moderate," in agreement with McDonald (1993) who states "the independence estimator may be recommended for practical purposes." On the other hand, Zhao et al. (1992) assert that assuming independence "can lead to important losses of efficiency," in line with Fitzmaurice et al. (1993) who state that it is "important to obtain a close approximation to $\text{cov}(\boldsymbol{Y}_i)$ in order to achieve high efficiency." The issue is complex since it depends on, among other things, the design and whether the covariates corresponding to the parameters are constant within an individual or not.

9.10 GEE2: Connected Estimating Equations

In an approach coined by Liang et al. (1992) as GEE2, there is a *connected* set of joint estimating equations for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. This approach is particularly appealing if the variance–covariance model is of interest. To motivate a pair of estimating equations, consider the following model for a single individual with n *independent* observations:

$$Y_i \mid \boldsymbol{\gamma}, \alpha \sim_{ind} \text{N}\left[\mu_i(\boldsymbol{\gamma}), \Sigma_i(\boldsymbol{\gamma}, \alpha)\right].$$

For example, we may have $\Sigma_i(\boldsymbol{\gamma}, \alpha) = \alpha\mu_i(\boldsymbol{\gamma})^2, i = 1, \ldots, n$. The log-likelihood is

$$l(\boldsymbol{\gamma}, \alpha) = -\frac{1}{2}\sum_{i=1}^{n}\log(\Sigma_i) - \frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i - \mu_i)^2}{\Sigma_i}.$$

Differentiation gives the score equations as

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\gamma}} &= -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial \Sigma_i}{\partial \boldsymbol{\gamma}}\right)^{\text{T}}\frac{1}{\Sigma_i} + \sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\gamma}}\right)^{\text{T}}\frac{(Y_i - \mu_i)}{\Sigma_i} + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial \Sigma_i}{\partial \boldsymbol{\gamma}}\right)^{\text{T}}\frac{(Y_i - \mu_i)^2}{\Sigma_i^2} \\ &= \sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\gamma}}\right)^{\text{T}}\frac{(Y_i - \mu_i)}{\Sigma_i} + \sum_{i=1}^{n}\left(\frac{\partial \Sigma_i}{\partial \boldsymbol{\gamma}}\right)^{\text{T}}\frac{\left[(Y_i - \mu_i)^2 - \Sigma_i\right]}{2\Sigma_i^2} \qquad (9.13)\end{aligned}$$

and

$$\frac{\partial l}{\partial \alpha} = -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial \Sigma_i}{\partial \alpha}\right)^{\mathrm{T}}\frac{1}{\Sigma_i} + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial \Sigma_i}{\partial \alpha}\right)^{\mathrm{T}}\frac{(Y_i-\mu_i)^2}{\Sigma_i^2}$$
$$= \sum_{i=1}^{n}\left(\frac{\partial \Sigma_i}{\partial \alpha}\right)^{\mathrm{T}}\frac{\left[(Y_i-\mu_i)^2-\Sigma_i\right]}{2\Sigma_i^2}. \tag{9.14}$$

This pair of quadratic estimating functions is unbiased given correct specification of the first two moments; to emphasize, normality of the data is not required. A disadvantage of the use of these functions, compared to the original GEE method (which is sometimes referred to as GEE1), is that if the variance model is wrong, we are no longer guaranteed a consistent estimator of $\boldsymbol{\gamma}$. If the model is correct, however, there will be a gain in efficiency.

Let

$$S_i = (Y_i-\mu_i)^2$$

with $\mathrm{E}[S_i] = \Sigma_i$. Under normality,

$$\mathrm{var}(S_i) = \mathrm{E}[S_i^2] - \mathrm{E}[S_i]^2 = 3\Sigma_i^2 - \Sigma_i^2 = 2\Sigma_i^2$$

Hence, (9.13) and (9.14) can be written

$$\frac{\partial l}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n}\boldsymbol{D}_i^{\mathrm{T}}V_i^{-1}(Y_i-\mu_i) + \sum_{i=1}^{n}\boldsymbol{E}_i W_i^{-1}(S_i-\Sigma_i) \tag{9.15}$$

$$\frac{\partial l}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n}\boldsymbol{F}_i W_i^{-1}(S_i-\Sigma_i) \tag{9.16}$$

where $\boldsymbol{D}_i = \partial\mu_i/\partial\boldsymbol{\beta}$, $\boldsymbol{E}_i = \partial\Sigma_i/\partial\boldsymbol{\beta}$, $\boldsymbol{F}_i = \partial\Sigma_i/\partial\alpha$, $V_i = \Sigma_i$, and $W_i = 2\Sigma_i^2$. This pair of estimating equations can be compared with the usual estimating equation specification

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\boldsymbol{D}_i^{\mathrm{T}}V_i^{-1}(Y_i-\mu_i).$$

The additional term is the information about $\boldsymbol{\gamma}$ in the variance.

We turn to the dependent data situation and let $\boldsymbol{\mu}_i$ denote the $n_i \times 1$ mean vector and $\boldsymbol{\Sigma}_i$ the $n_i \times n_i$ covariance matrix. The general form of estimating equations is

$$\sum_{i=1}^{m}\begin{bmatrix}\boldsymbol{D}_i & \mathbf{0}\\ \boldsymbol{E}_i & \boldsymbol{F}_i\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{V}_i & \boldsymbol{C}_i\\ \boldsymbol{C}_i^{\mathrm{T}} & \boldsymbol{W}_i\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{Y}_i-\boldsymbol{\mu}_i\\ \boldsymbol{S}_i-\boldsymbol{\Sigma}_i\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$, $\boldsymbol{E}_i = \partial\boldsymbol{\Sigma}_i/\partial\boldsymbol{\beta}$, and $\boldsymbol{F}_i = \partial\boldsymbol{\Sigma}_i/\partial\alpha$ and we have "working" variance–covariance structure

$$\begin{aligned}\boldsymbol{V}_i &= \text{var}(\boldsymbol{Y}_i)\\ \boldsymbol{C}_i &= \text{cov}(\boldsymbol{Y}_i, \boldsymbol{S}_i)\\ \boldsymbol{W}_i &= \text{var}(\boldsymbol{S}_i).\end{aligned}$$

When $\boldsymbol{C}_i = \boldsymbol{0}$, we obtain

$$\boldsymbol{G}_1(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) + \sum_{i=1}^{m} \boldsymbol{E}_i \boldsymbol{W}_i^{-1}(\boldsymbol{S}_i - \boldsymbol{\Sigma}_i) \tag{9.17}$$

$$\boldsymbol{G}_2(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \boldsymbol{F}_i \boldsymbol{W}_i^{-1}(\boldsymbol{S}_i - \boldsymbol{\Sigma}_i) \tag{9.18}$$

which are the dependent data version of the normal score equations we obtained earlier, that is, (9.15) and (9.16). In the dependent data pair of equations, we have freedom in choosing $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$. In particular, the latter need not be chosen to coincide with that under a multivariate normal model, and, since this choice is difficult, we could instead choose working independence.

It can be shown (Prentice and Zhao 1991, Appendix 2) that (9.17) and (9.18) arise from the *quadratic exponential model*

$$p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i) = \Delta_i^{-1} \exp[\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{\theta}_i + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\lambda}_i + c_i(\boldsymbol{y}_i)], \tag{9.19}$$

where $\boldsymbol{\theta}_i = [\theta_{i1}, \ldots, \theta_{in_i}]^{\mathrm{T}}$ is the canonical parameter,

$$\boldsymbol{w}_i = [y_{i1}^2, y_{i1}y_{i2}, \ldots, y_{i2}^2, y_{i2}y_{i3}, \ldots]^{\mathrm{T}}$$

is the vector of squared responses, $c_i(\cdot)$ is a function that defines the "shape," $\Delta_i = \Delta_i(\boldsymbol{\theta}_i, \boldsymbol{\lambda}_i, c_i)$ is a normalization constant, and $\boldsymbol{\lambda}_i = [\lambda_{i11}, \lambda_{i12}, \ldots, \lambda_{i22}, \lambda_{i23}, \ldots]^{\mathrm{T}}$. As an example of this form, if all the responses are continuous on the whole real line and $c_i = 0$, the multivariate normal is recovered (Exercise 9.2). Gourieroux et al. (1984) showed that the quadratic exponential family is unique in giving consistent estimates of the mean and covariance parameters, even in the situation in which the data actually arise from outside this family. So, as the exponential family produces desirable consistency properties for mean parameters, the quadratic exponential family has the same properties when mean and variance parameters are of interest.

To emphasize: For consistency of $\widehat{\boldsymbol{\gamma}}$, we require the models for both $\boldsymbol{Y}_i$ and $\boldsymbol{S}_i$ to be correct, and there is increased efficiency over the single estimating equation version (GEE1) if this is the case. This approach is useful if the variance–covariance parameters are of primary interest as, for example, in some breeding and genetic applications. Otherwise, if can, be prudent to stick with GEE1.

9.11 Interpretation of Marginal and Conditional Regression Coefficients

To illustrate the differences in interpretation of marginal and conditional coefficients, we examine the meaning of parameters for a loglinear model. In a marginal model, such as is considered under GEE, we have

$$\mathrm{E}[Y \mid x] = \exp(\gamma_0 + \gamma_1 x),$$

in which case $\exp(\gamma_1)$ is the multiplicative change in the average response over two populations of individuals whose x values differ by one unit . Under the conditional mixed model, the interpretation of regression coefficients is conditional on the value of the random effect. For the model

$$\mathrm{E}[Y \mid x, b] = \exp(\beta_0 + \beta_1 x + b),$$

with $b \mid \sigma_0^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$, $\exp(\beta_1)$ is therefore the change in the expected response for two individuals with identical random effects. Sometimes, the comparison is described as between two *typical* (i.e., $b = 0$) individuals who differ in x by one unit. The marginal mean corresponding to this model follows from the variance of a lognormal distribution:

$$\mathrm{E}[Y \mid x] = \mathrm{E}_b[\,\mathrm{E}(Y \mid x, b)\,] = \exp(\beta_0 + \sigma^2/2 + \beta_1 x).$$

Therefore, for the random intercepts, loglinear model $\exp(\beta_1)$ has the same marginal interpretation to $\exp(\gamma_1)$ and the marginal intercept is $\gamma_0 = \beta_0 + \sigma^2/2$.

We now consider the random intercepts and slopes model

$$\mathrm{E}[Y \mid x, \boldsymbol{b}] = \exp\left[(\beta_0 + b_0) + (\beta_1 + b_1)x\right]$$

where $\boldsymbol{b} = [b_0, b_1]$ and

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \sim \mathrm{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{bmatrix}\right).$$

In this model $\exp(\beta_1)$ is the relative risk between two individuals with the same $\boldsymbol{b}$ but with x values that differ by one unit. That is,

$$\exp(\beta_1) = \frac{\mathrm{E}[Y \mid x, \boldsymbol{b}]}{\mathrm{E}[Y \mid x - 1, \boldsymbol{b}]}.$$

An alternative interpretation is to say that it is the expected change between two "typical individuals," that is, individuals with specific values of the random effects, $\boldsymbol{b} = \mathbf{0}$. Under this model, the marginal mean is

$$\mathrm{E}[Y \mid x] = \exp\left[\beta_0 + D_{00}/2 + x(\beta_1 + D_{01}) + x^2 D_{11}/2\right]$$

so that a quadratic loglinear marginal model has been induced by the conditional formulation. The marginal *median* is $\exp(\beta_0 + \beta_1 x)$ so that $\exp(\beta_1)$ is the ratio of median responses between two populations whose x values differ by one unit. There is no such simple interpretation in terms of marginal means.

Hence, marginal inference is possible under a mixed model formulation, though care must be taken to derive the exact form of the marginal model. Estimation of marginal parameters via GEE produces a consistent estimator in more general circumstances than mixed model estimation, though there is an efficiency loss if the random effects model is correct.

Example: Seizure Data

The marginal mean version of the conditional model fitted previously in this chapter is

$$\mathrm{E}[Y_{ij}] = T_{ij} \exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{ij2} + \gamma_3 x_{i1} x_{ij2}).$$

The parameters are interpreted as follows:

- $\exp(\gamma_0)$ is the expected rate of seizures in the placebo group during the baseline period, $j = 0$ (this expectation is over the hypothetical population of individuals who were assigned to the placebo group).
- $\exp(\gamma_1)$ is the ratio of the expected seizure rate in the progabide group, compared to the placebo group, during the baseline period.
- $\exp(\gamma_2)$ is the ratio of the expected seizure rate post-baseline as compared to baseline, in the placebo group.
- $\exp(\gamma_3)$ is the ratio of the expected seizure rates in the progabide group in the post-baseline period, as compared to the placebo group, in the same period. Hence, $\exp(\gamma_3)$ is a period by treatment effect and is the parameter of interest.

The loglinear mean model suggests the variance model $\mathrm{var}(Y_{ij}) = \alpha_1 \mu_{ij}$. We consider various forms for the working correlation. Table 9.7 gives estimates and standard errors under various models. The Poisson, quasi-likelihood, and working independence GEE models have estimating equation

$$\boldsymbol{G}(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} (\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_i) = \boldsymbol{0}.$$

Consequently, the point estimates coincide but the models differ in the manner by which the standard errors are calculated. The Poisson standard errors are clearly much too small. The coincidence of the estimates and standard errors for independence and exchangeable working correlations is a consequence of the balanced design. The quasi-likelihood standard errors are increased by $\sqrt{19.7} = 4.4$ (in line with the empirical estimates in Table 9.2) but do not acknowledge dependence of observations on the same individual (so estimation is carried out

Table 9.7 Parameter estimates and standard errors under various models for the seizure data

	Estimates and standard errors									
	Poisson		Quasi-Lhd		GEE independence		GEE exchangeable		GEE AR(1)	
γ_0	1.35	0.034	1.35	0.15	1.35	0.16	1.35	0.16	1.31	0.16
γ_1	0.027	0.047	0.027	0.21	0.027	0.22	0.027	0.22	0.015	0.21
γ_2	0.11	0.047	0.11	0.21	0.11	0.12	0.11	0.12	0.16	0.11
γ_3	−0.10	0.065	−0.10	0.29	−0.10	0.22	−0.10	0.22	−0.13	0.27
α_1, α_2	1.0	0	19.7	0	19.4	0	19.4	0.78	20.0	0.89

Parameter meaning: γ_0 is the log baseline seizure rate in the placebo group; γ_1 is the log of the ratio of seizure rates between the progabide and placebo groups, at baseline; γ_2 is the log of the ratio of seizure rates in the post-baseline and baseline placebo groups; γ_3 is the log of the ratio of the seizure rate in the progabide group as compared to the placebo group, post-baseline; α_1 and α_2 are variance and correlation parameters, respectively

as if we have 59×5 independent observations). The standard errors of estimated parameters that are associated with time-varying covariates (γ_2 and γ_3) are reduced under GEE, since within-person comparisons are being made and a longitudinal design can be very efficient in such a study, if there is strong within-individual dependence (as discussed in Sect. 8.3). In none of the analyses would the treatment effect of interest be judged significantly different from zero, under conventional levels.

9.12 Introduction to Modeling Dependent Binary Data

Binary outcomes are the simplest form of data but are, ironically, one of the most challenging to model. For a single binary variable Y all moments are determined by $p = \mathrm{E}[Y]$. Specifically, $\mathrm{E}[Y^r] = p$ for $r \geq 1$, so that Bernoulli random variables cannot be overdispersed. Before turning to observations on multiple units, we initially adopt a simplified notation and consider n binary observations $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^{\mathrm{T}}$. Under conditional independence and with probabilities $p_j = \mathrm{E}[Y_j]$,

$$\Pr(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{p}) = \prod_{j=1}^{n} p_j^{y_j} (1 - p_j)^{1-y_j},$$

with $\boldsymbol{p} = [p_1, \ldots, p_n]^{\mathrm{T}}$. In Chap. 7, we saw that a common mean form is the logistic regression model with $\log[p_j/(1-p_j)] = \boldsymbol{x}_j \boldsymbol{\beta}$. In this chapter we wish to formulate models that allow for dependence between binary outcomes, with a starting point being the specification of a multivariate binary distribution. Such a joint distribution can be used with a likelihood-based approach, or one can use the first one or two moments only within a GEE approach. The difficulty with multivariate binary data is that there is no natural way to characterize dependence between pairs, triples, etc., of binary responses. In the dependent binary data situation, we will show that

correlation parameters are tied to the means, making estimation from a model based on means and correlations unattractive.

To specify the joint distribution of n binary responses requires 2^n probabilities so that the saturated model has $2^n - 1$ parameters. This may be contrasted with a saturated multivariate normal model which has n means, n variances, and $n(n-1)/2$ correlations. As n becomes large, the number of parameters in the binary saturated model is very large. With $n = 10$, for example, there are $2^{10} - 1 = 1,023$ parameters in the binary model as compared to 65 in the normal model. Our aim is to reduce the $2^n - 1$ distinct probabilities to give formulations that allow both a parsimonious description and the interpretable specification of a regression model.

We begin our description of models for multivariate binary data in Sect. 9.13 with a discussion of mixed models, with likelihood, Bayesian and conditional likelihood approaches to inference. Next, in Sect. 9.14, marginal models are described.

9.13 Mixed Models for Binary Data

9.13.1 Generalized Linear Mixed Models for Binary Data

In Sect. 7.5, we discussed a beta-binomial model for overdispersed data. This form is not very flexible, for the reasons described in Sect. 9.8, and so we describe an alternative mixed model with normal random effects. Let Y_{ij} be the binary "success" indicator with $j = 1, \ldots, n_i$ trials on each of $i = 1, \ldots, m$ units.

Consider the GLMM with logistic link:

Stage One: Likelihood: $Y_{ij} \mid p_{ij} \sim_{ind} \text{Bernoulli}(n_{ij}, p_{ij})$ with the linear logistic model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i.$$

In this model, $\boldsymbol{\beta}$ represents a $(k+1) \times 1$ vector of fixed effects and $\boldsymbol{b}_i$ a $(q+1) \times 1$ vector of random effects, with $q \leq k$. Let $\boldsymbol{x}_{ij} = [1, x_{ij1}, \ldots, x_{ijk}]$ be a $(k+1) \times 1$ vector of covariates, so that $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]$ is the design matrix for the fixed effects, and let $\boldsymbol{z}_{ij} = [1, z_{ij1}, \ldots, z_{ijq}]^{\mathrm{T}}$ be a $(k+1) \times 1$ vector of variables that are a subset of $\boldsymbol{x}_{ij}$, so that $\boldsymbol{z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i}]^{\mathrm{T}}$ is the design matrix for the random effects.

Stage Two: Random effects distribution: $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \mathrm{N}_{q+1}(\boldsymbol{0}, \boldsymbol{D})$ for $i = 1, \ldots, m$.

As we have repeatedly stressed, the conditional parameters $\boldsymbol{\beta}$ and the marginal parameters $\boldsymbol{\gamma}$ have different interpretations in nonlinear situations, and for a logistic model, there is no exact analytical relationship between the two. However, we

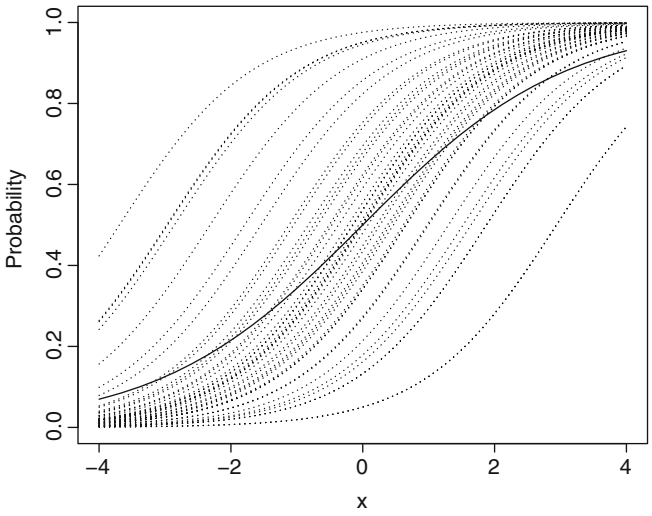

Fig. 9.6 Individual-level curves (*dotted lines*) from a random intercept logistic GLMM, along with marginal curve (*solid line*). The specific model is logit $(\mathrm{E}[Y \mid b]) = \beta_0 + \beta_1 x$, with $\beta_0 = 0, \beta_1 = 1$ and $b \sim_{iid} \mathrm{N}(0, 2^2)$. The approximate attenuation factor of the marginal curve, which is given by the denominator of (9.21), is 1.54

may approximate the relationship. For the random intercepts model $b_i \mid \sigma_0^2 \sim_{iid} \mathrm{N}(0, \sigma_0^2)$, we have, for a generic Bernoulli response Y with associated random effect b,

$$\begin{aligned} \mathrm{E}[Y] &= \frac{\exp(\boldsymbol{x\gamma})}{1+\exp(\boldsymbol{x\gamma})} = \mathrm{E}_b[\,\mathrm{E}(Y \mid b)\,] \\ &= \mathrm{E}_b\left[\frac{\exp(\boldsymbol{x\beta}+b)}{1+\exp(\boldsymbol{x\beta}+b)}\right] \approx \frac{\exp(\boldsymbol{x\beta}/[c^2\sigma_0^2+1]^{1/2})}{1+\exp(\boldsymbol{x\beta}/[c^2\sigma_0^2+1]^{1/2})} \end{aligned} \tag{9.20}$$

where $c = 16\sqrt{3}/(15\pi)$ (Exercise 9.1), so that

$$\boldsymbol{\gamma} \approx \frac{\boldsymbol{\beta}}{[c^2\sigma_0^2+1]^{1/2}}. \tag{9.21}$$

Consequently, the marginal coefficients are attenuated toward zero. Figure 9.6 illustrates this phenomena for particular values of $\beta_0, \beta_1, \sigma_0^2$. We observe that the averaging of the conditional curves results in a flattened marginal curve. This attenuation was first encountered in Sect. 7.9 when the lack of collapsibility of the odds ratio was discussed. We emphasize that one should not view the difference in marginal and conditional parameter estimates as bias. If $\sigma_0 > 0$ and $\beta_1 \neq 0$, the parameters will differ, but they are estimating different quantities. In practice, if we fit marginal and conditional models and we do not see attenuation, then the

approximation could be poor (e.g., if σ_0^2 is large) or some of the assumptions of the conditional model could be inaccurate.

For the general logistic mixed model

$$\log\left(\frac{\mathrm{E}[Y \mid \boldsymbol{b}]}{1-\mathrm{E}[Y \mid \boldsymbol{b}]}\right) = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{z}\boldsymbol{b}$$

with $\boldsymbol{b} \mid \boldsymbol{D} \sim_{iid} \mathrm{N}_{q+1}(\boldsymbol{0}, \boldsymbol{D})$, we obtain

$$\mathrm{E}[Y] = \frac{\exp(\boldsymbol{x}\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}\boldsymbol{\gamma})} \approx \frac{\exp\left(\boldsymbol{x}\boldsymbol{\beta}/ \mid c^2\boldsymbol{D}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}} + \boldsymbol{I}_{q+1} \mid^{(q+1)/2}\right)}{1+\exp\left(\boldsymbol{x}\boldsymbol{\beta}/ \mid c^2\boldsymbol{D}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}} + \boldsymbol{I}_{q+1} \mid^{(q+1)/2}\right)}$$

so that

$$\boldsymbol{\gamma} \approx \frac{\boldsymbol{\beta}}{\mid c^2\boldsymbol{D}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}} + \boldsymbol{I}_{q+1} \mid^{(q+1)/2}}.$$

With random slopes or more complicated random effects structures, it is therefore far more difficult to understand the relationship between conditional and marginal parameters.

Marginal inference is possible with mixed models, but one needs to do a little work. Specifically, if one requires marginal inference, then the above approximations may be invoked, or one may directly calculate the required integrals using a Monte Carlo estimate. For example, the marginal probability at $\boldsymbol{x}$ is

$$\widehat{\mathrm{E}}[Y \mid \boldsymbol{x}] = \frac{1}{S}\sum_{s=1}^{S} \frac{\exp(\boldsymbol{x}\widehat{\boldsymbol{\beta}} + \boldsymbol{b}^{(s)})}{1+\exp(\boldsymbol{x}\widehat{\boldsymbol{\beta}} + \boldsymbol{b}^{(s)})} \tag{9.22}$$

where the random effects are simulated as $\boldsymbol{b}^{(s)} \mid \widehat{\boldsymbol{D}} \sim \mathrm{N}_{q+1}(\boldsymbol{0}, \widehat{\boldsymbol{D}})$, $s = 1, \ldots, S$. A more refined Bayesian approach would replace $\widehat{\boldsymbol{D}}$ by samples from the posterior $p(\boldsymbol{D} \mid \boldsymbol{y})$.

An important distinction between conditional and marginal modeling through GEE is that the latter is likely to be more robust to model misspecification, since it directly models marginal associations.

Recall that the logistic regression model for binary data can be derived by consideration of an unobserved (latent) continuous logistic random variable (Sect. 7.6.1). This latent formulation can be extended to the mixed model. In particular, assume $U_{ij} = \mu_{ij} + b_i$, where $b_i \mid \sigma_0^2 \sim \mathrm{N}(0, \sigma_0^2)$ and U_{ij} follows the standard logistic distribution, that is, $U_{ij} \mid b_i \sim_{ind} \mathrm{Logistic}(\mu_{ij} + b_i, 1)$. Without loss of generality set, $Y_{ij} = 1$ if $U_{ij} > c$ and 0 otherwise. Then

$$\Pr(Y_{ij} = 1 \mid b_i) = \Pr(U_{ij} > c \mid b_i) = \frac{\exp(\mu_{ij} + b_i - c)}{1+\exp(\mu_{ij} + b_i - c)}$$

and taking $\mu_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + c$ produces the random effects logistic model.

An interpretation of σ_0^2 is obtained by comparing its magnitude to $\pi^2/3$ (the variance of the logistic distribution, which can be viewed as the within-person variability) via the intra-class correlation:

$$\rho = \text{corr}(U_{ij}, U_{ik}) = \frac{\sigma_0^2}{\sigma_0^2 + \pi^3/3}.$$

Note that ρ is the marginal correlation (averaged over the random effects) among the unobserved latent variables U_{ij} and not the marginal correlation among the Y_{ij}'s. See Fitzmaurice et al. (2004, Sect. 12.5) for further discussion.

We examine the marginal moments further. The marginal mean is $\text{E}[Y_{ij}] = \Pr(Y_{ij} = 1) = \text{E}_{b_i}[p_{ij}]$ where we continue to consider the random intercepts only model

$$p_{ij} = \frac{\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}{1 + \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}.$$

The expectation is over the distribution of the random effect. We have already derived the approximate marginal mean (9.20), which we write as

$$p_{ij}^{\star} = \frac{\exp[\boldsymbol{x}_{ij}\boldsymbol{\beta}/(c^2\sigma_0^2 + 1)^{1/2}]}{1 + \exp[\boldsymbol{x}_{ij}\boldsymbol{\beta}/(c^2\sigma_0^2 + 1)^{1/2}]}.$$

The variance is

$$\begin{aligned}
\text{var}(Y_{ij}) &= \text{E}[\text{var}(Y_{ij} \mid b_i)] + \text{var}[\text{E}(Y_{ij} \mid b_i)] \\
&= \text{E}[p_{ij} - p_{ij}^2] + \text{E}[p_{ij}^2] - \text{E}[p_{ij}]^2 \\
&= p_{ij}^{\star}(1 - p_{ij}^{\star}),
\end{aligned}$$

illustrating again that there is no overdispersion for a Bernoulli random variable. This gives the diagonal elements of the marginal variance–covariance matrix. The covariances between responses on the same unit i are

$$\begin{aligned}
\text{cov}(Y_{ij}, Y_{ik}) &= \text{cov}\left(\frac{\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}{1 + \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}, \frac{\exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + b_i)}{1 + \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + b_i)}\right) \\
&= \text{E}\left[\left(\frac{\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}{1 + \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i)}\right)\left(\frac{\exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + b_i)}{1 + \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta} + b_i)}\right)\right] - p_{ij}^{\star}p_{ik}^{\star},
\end{aligned}$$

so note that the marginal covariance is not constant and not of easily interpretable form. With a single random effect, the correlations are all determined by the single parameter σ_0.

9.13.2 Likelihood Inference for the Binary Mixed Model

As with the GLMMs described in Sect. 9.4, the integrals required to evaluate the likelihood for the fixed effects $\boldsymbol{\beta}$ and variance components $\boldsymbol{\alpha} = \boldsymbol{D}$ are analytically intractable. Unfortunately the Laplace approximation method may not be reliable for binary GLMMs, particularly if the random effects variances are large. For this reason adaptive Gauss–Hermite quadrature methods are often resorted to, though care in implementation is required to ensure that sufficient points are used to obtain an accurate approximation. When maximization routines encounter convergence problems, it may be an indication that either the model being fitted is not supported by the data or that the data do not contain sufficient data to estimate all of the parameters.

9.13.3 Bayesian Inference for the Binary Mixed Model

A Bayesian approach to binary GLMMs requires priors to be specified for $\boldsymbol{\beta}$ and $\boldsymbol{D}$. As in Sect. 9.6.2, the priors may be specified in terms of interpretable quantities, for example, the residual odds of success. The information in binary data is limited, and so sensitivity to the priors may be encountered, particularly the prior on **D**. As with likelihood-based approaches, greater care is required in computation with binary data. Fong et al. (2010) report that the INLA method is relatively inaccurate for binary GLMMs so that MCMC is the more reliable method if the binomial denominators are small.

Example: Contraception Data

We illustrate likelihood inference for a binary GLMM using the contraception data introduced in Sect. 9.2.1. Let $Y_{ij} = 0/1$ denote the absence/presence of amenorrhea in the ith woman at time t_{ij}, where the latter takes the values 1, 2, 3, or 4. Also, let $d_i = 0/1$ represent the randomization indicators to doses of 100 mg/150 mg, for $i = 1, \ldots, 1151$ women (576 and 575 women received the low and high doses, respectively). There are n_i observations per woman, up to a maximum of 4. We consider the following two-stage model:

Stage One: The response model is $Y_{ij} \mid p_{ij} \sim_{ind}$ Bernoulli (p_{ij}) with

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 d_i t_{ij} + \beta_4 d_i t_{ij}^2 + b_i, \tag{9.23}$$

so that we have separate quadratic models in time for each of the two-dose levels.

Table 9.8 Mixed effects model parameter estimates for the contraception data

	Likelihood Laplace		Likelihood G–H[a]		Bayesian MCMC	
Parameter	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
Intercept	−3.8	0.27	−3.8	0.30	−3.6	0.27
Low-dose time	1.1	0.25	1.1	0.27	0.99	0.25
Low-dose time^2	−0.044	0.052	−0.042	0.055	−0.015	0.052
High-dose time	0.55	0.18	0.56	0.21	0.55	0.18
High-dose time^2	−0.11	0.051	−0.11	0.050	−0.11	0.058
σ_0	2.1	–	2.3	0.11	2.2	0.13

[a] Adaptive Gauss–Hermite with 50 points

Stage Two: The random effects model is $b_i \mid \sigma_0^2 \sim_{iid} \mathrm{N}(0, \sigma_0^2)$.

We do not include a term for the main effect of dose, since we assume that randomization has ensured that the two-dose groups are balanced at baseline ($t = 0$). The conditional odds ratios $\exp(\beta_1)$ and $\exp(\beta_2)$ represent linear and quadratic terms in time for a typical individual ($b_i = 0$) in the low-dose group. Similarly, $\exp(\beta_1 + \beta_3)$ and $\exp(\beta_2 + \beta_4)$ represent linear and quadratic terms in time for a typical individual ($b_i = 0$) in the high-dose group.

Table 9.8 gives parameter estimates and standard errors for a number of analyses, including Laplace and adaptive Gauss–Hermite rules for likelihood calculation. We initially concentrate on the Gauss–Hermite results which are more reliable than those based on the Laplace implementation. Informally, comparing the estimates with the standard errors, the linear terms in time are clearly needed, while it is not so obvious that the quadratic terms are required.

In terms of substantive conclusions, a woman assigned the high dose, when compared to a woman assigned the low dose, both with the same baseline risk of amenorrhea (i.e., with the same random effect) will have increased odds at time t of

$$\exp(\widehat{\beta}_3 t + \widehat{\beta}_4 t^2)$$

giving increases of 1.6, 2.0, 2.0, 1.6 at times 1, 2, 3, 4, respectively. Hence, the difference between the groups increases and then decreases as a function of time, though it is always greater than zero.

The standard deviation of the random effects $\widehat{\sigma} = 2.3$ is substantial here. An estimate of a 95% interval for the risk of amenorrhea in the low-dose group at occasion 1 is

$$\frac{\exp(-3.8 + 1.1 - 0.042 \pm 1.96 \times 2.3)}{1 + \exp(-3.8 + 1.1 - 0.042 \pm 1.96 \times 2.3)} = [0.0007, 0.85],$$

so that we have very large between-woman variability in risk. The marginal intra-class correlation coefficient is estimated as $\rho = 0.61$ (recall this is the correlation for the latent variable and not for the marginal responses).

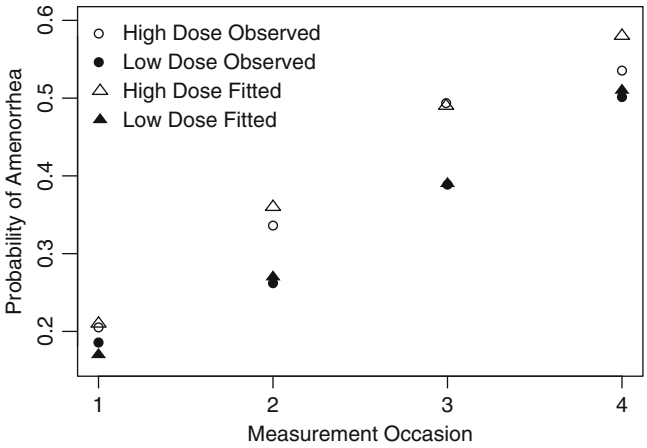

Fig. 9.7 Probability of amenorrhea over time in low- and high-dose groups in the contraception data, along with fitted probabilities. The latter are calculated via Monte Carlo simulation, with likelihood estimation in the mixed model, implemented with Gauss–Hermite quadrature

Table 9.9 Monte Carlo estimated variances (on the *diagonal*) and correlations (*upper diagonal*), between measurements on the same woman, at different observations occasions (1–4), in the low- (*left*) and high- (*right*) dose groups

	1	2	3	4		1	2	3	4
1	0.14	0.38	0.36	0.33	1	0.17	0.39	0.36	0.33
2		0.20	0.41	0.39	2		0.23	0.42	0.40
3			0.24	0.43	3			0.25	0.43
4				0.25	4				0.24

These estimates are based on likelihood estimation in the mixed model, implemented with Gauss–Hermite quadrature

Allowing the random effects variance to vary by covariate groups is important to investigate since missing such dependence can lead to serious inaccuracies (Heagerty and Kurland 2001). The assumption of a common σ_0 in the two groups is important for accurate inference in this example. We fit separate logistic GLMMs to the two-dose groups and obtain estimates of 2.3 and 2.2, illustrating that a common σ_0 is supported by the data.

We evaluate the marginal means calculation using Monte Carlo integration. These means are shown, along with the observed proportions, in Fig. 9.7. We see that the overall fit is good, apart from the last time point (for which there is reduced data due to dropout).

In Table 9.9, we estimate the marginal variance–covariance and correlation matrices for the two-dose groups using Monte Carlo integration. As we have already discussed in Sect. 9.13.1 a random intercepts only model does not lead to correlations that are constant across time (unlike the linear model). In general, the estimates are in reasonable agreement with the empirical variances and correlations reported in Table 9.1.

For the Bayesian analysis, the prior for the intercept was relatively flat, $\beta_0 \sim \mathrm{N}(0, 2.38^2)$. If there was no effect of time (i.e., if $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$) then a 95% interval for the probabilities for a typical individual would be $\exp(\pm 1.96 \times 2.38) = [0.009, 0.99]$. For the regression coefficients, we specify $\beta_k \sim \mathrm{N}(0, 0.98^2)$ which gives a 95% interval for the odds ratios of $\exp(\pm 1.96 \times 0.98) = [0.15, 6.8]$. Finally, for σ_0^{-2}, we assume a Gamma(0.5,0.1) prior which gives a 95% interval for σ_0 of [0.06,4.5]. More informatively, a 95% interval for the residual odds is [0.17,6.0]. These priors are not uninformative but correspond to ranges for probabilities and odds ratios that are consistent with the application.

The posterior means and standard deviations are given in Table 9.8, and we see broad agreement with the MLEs and standard errors found using Gauss–Hermite. The intra-class correlation coefficient is estimated as 0.60 with 95% credible interval [0.55, 0.67].

9.13.4 Conditional Likelihood Inference for Binary Mixed Models

Recall that conditional likelihood is a technique for eliminating nuisance parameters, in this case the random effects in the mixed model. Following from Sect. 9.5, we outline the approach as applied to the binary mixed model with random intercepts. Consider individual i with binary observations $y_{i1}, \ldots, y_{in_i}$ and assume the random intercepts model $Y_{ij} \mid \lambda_i, \boldsymbol{\beta}^\star \sim \text{Bernoulli}(p_{ij})$, where

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{x}_{ij}\boldsymbol{\beta}^\star + \lambda_i$$

and $\lambda_i = \boldsymbol{x}_i\boldsymbol{\beta}^\dagger + b_i$ so that $\boldsymbol{\beta}^\dagger$ represents those parameters associated with covariates that are constant within an individual and $\boldsymbol{\beta}^\star$ those that vary. Mimicking the development in Sect. 9.5, the joint distribution for the responses of the ith unit is

$$\begin{aligned}
\Pr(y_{i1}, \ldots, y_{in_i} \mid \lambda_i, \boldsymbol{\beta}^\star) &= \prod_{j=1}^{n_i} \frac{\exp\left(\lambda_i y_{ij} + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij}\right)}{1 + \exp\left(\lambda_i + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{ij}^{\mathrm{T}}\right)} \\
&= \frac{\exp\left(\lambda_i \sum_{j=1}^{n_i} y_{ij} + \boldsymbol{\beta}^{\star\mathrm{T}} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij}\right)}{\prod_{j=1}^{n_i}\left[1 + \exp\left(\lambda_i + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{ij}^{\mathrm{T}}\right)\right]} \\
&= \frac{\exp\left(\lambda_i t_{2i} + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{t}_{1i}\right)}{\prod_{j=1}^{n_i}\left[1 + \exp\left(\lambda_i + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{ij}^{\mathrm{T}}\right)\right]} \\
&= \frac{\exp\left(\lambda_i t_{2i} + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{t}_{1i}\right)}{k(\lambda_i, \boldsymbol{\beta})} \\
&= p(\boldsymbol{t}_{1i}, t_{2i} \mid \lambda_i, \boldsymbol{\beta}^\star)
\end{aligned}$$

where

$$\boldsymbol{t}_{1i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij}, \quad t_{2i} = \sum_{j=1}^{n_i} y_{ij} = y_{i+}$$

and

$$k(\lambda_i, \boldsymbol{\beta}^\star) = \prod_{j=1}^{n_i} \left[1 + \exp\left(\lambda_i + \boldsymbol{\beta}^{\star\mathrm{T}} \boldsymbol{x}_{ij}^{\mathrm{T}}\right)\right].$$

Therefore, the conditioning statistic is the number of successes on the ith unit. We have conditional likelihood

$$L_c(\boldsymbol{\beta}) = \prod_{i=1}^{m} p(\boldsymbol{t}_{1i} \mid t_{2i}, \boldsymbol{\beta}^\star) = \prod_{i=1}^{m} \frac{p(\boldsymbol{t}_{1i}, t_{2i} \mid \lambda_i, \boldsymbol{\beta}^\star)}{p(t_{2i} \mid \lambda_i, \boldsymbol{\beta}^\star)}$$

where

$$p(\boldsymbol{t}_{2i} \mid \lambda_i, \boldsymbol{\beta}^\star) = \frac{\sum_{l=1}^{\binom{n_i}{y_{i+}}} \exp\left(\lambda_i y_{i+} + \boldsymbol{\beta}^{\star\mathrm{T}} \sum_{k=1}^{n_i} \boldsymbol{x}_{ik}^{\mathrm{T}} y_{ik}^{(l)}\right)}{k(\lambda_i, \boldsymbol{\beta}^\star)},$$

and the summation is over the $\binom{n_i}{y_{i+}}$ ways of choosing y_{i+} ones out of n_i and $\boldsymbol{y}_i^{(l)} = [y_{i1}^{(l)}, \ldots, y_{in_i}^{(l)}]$, $l = 1, \ldots, \binom{n_i}{y_{i+}}$ is the collection of these ways. Inference may be based on the conditional likelihood

$$\begin{aligned} L_c(\boldsymbol{\beta}^\star) &= \prod_{i=1}^{m} \frac{\exp\left(\lambda_i y_{i+} + \boldsymbol{\beta}^{\star\mathrm{T}} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij}\right)}{\sum_{l=1}^{\binom{n_i}{y_{i+}}} \exp\left(\lambda_i y_{i+} + \boldsymbol{\beta}^{\star\mathrm{T}} \sum_{k=1}^{n_i} \boldsymbol{x}_{ik}^{\mathrm{T}} y_{ik}^{(l)}\right)} \\ &= \prod_{i=1}^{m} \frac{\exp\left(\boldsymbol{\beta}^{\star\mathrm{T}} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}^{\mathrm{T}} y_{ij}\right)}{\sum_{l=1}^{\binom{n_i}{y_{i+}}} \exp\left(\boldsymbol{\beta}^{\star\mathrm{T}} \sum_{k=1}^{n_i} \boldsymbol{x}_{ik}^{\mathrm{T}} y_{ik}^{(l)}\right)}. \end{aligned}$$

Hence, there is no need to specify a distribution for the unit-specific parameters that allow for within-unit dependence, as they are eliminated by the conditioning argument.

As an example, if $n_i = 3$ and $\boldsymbol{y}_i = [0, 0, 1]$ so that $y_{i+} = 1$, then

$$\boldsymbol{y}_i^{(1)} = [1, 0, 0], \quad \boldsymbol{y}_i^{(2)} = [0, 1, 0], \quad \boldsymbol{y}_i^{(3)} = [0, 0, 1]$$

and the contribution to the conditional likelihood is

$$\frac{\exp(\boldsymbol{\beta}^{\star\mathrm{T}} \boldsymbol{x}_{i3}^{\mathrm{T}})}{\exp(\boldsymbol{\beta}^{\star\mathrm{T}} \boldsymbol{x}_{i1}^{\mathrm{T}}) + \exp(\boldsymbol{\beta}^{\star\mathrm{T}} \boldsymbol{x}_{i2}^{\mathrm{T}}) + \exp(\boldsymbol{\beta}^{\star\mathrm{T}} \boldsymbol{x}_{i3}^{\mathrm{T}})}.$$

As a second example, if $n_i = 3$ and $\boldsymbol{y}_i = [1, 0, 1]$ so that $y_{i+} = 2$, then

$$\boldsymbol{y}_i^{(1)} = [1, 1, 0], \quad \boldsymbol{y}_i^{(2)} = [1, 0, 1], \quad \boldsymbol{y}_i^{(3)} = [0, 1, 1],$$

and the contribution to the conditional likelihood is

$$\frac{\exp(\boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i1}^{\mathrm{T}} + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i3}^{\mathrm{T}})}{\exp(\boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i1}^{\mathrm{T}} + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i2}^{\mathrm{T}}) + \exp(\boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i1}^{\mathrm{T}} + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i3}^{\mathrm{T}}) + \exp(\boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i2}^{\mathrm{T}} + \boldsymbol{\beta}^{\star\mathrm{T}}\boldsymbol{x}_{i3}^{\mathrm{T}})}.$$

There is no contribution to the conditional likelihood from individuals with $n_i = 1$ or $y_{i+} = 0$ or $y_{i+} = n_i$. The conditional likelihood can be computationally expensive to evaluate if n_i is large, for example, if $n_i = 20$ and $y_{i+} = 10$ there are $\binom{n_i}{y_{i+}} = 184,756$ variations. The similarity to Cox's partial likelihood (e.g., Kalbfleisch and Prentice 2002, Chap. 4) may be exploited to carry out computation, however.

We reiterate that the conditional likelihood estimates those elements of $\boldsymbol{\beta}^{\star}$ that are associated with covariates that vary within individuals. If a covariate only varies between individuals, then its effect cannot be estimated using conditional likelihood. For covariates that vary both between and within individuals, only the within-individual contrasts are used.

9.14 Marginal Models for Dependent Binary Data

We now consider the marginal modeling of dependent binary data. We begin by describing how the GEE approach of Sect. 9.9 can be used for binary data and then describe alternative approaches.

9.14.1 Generalized Estimating Equations

For the marginal Bernoulli outcome $Y_{ij} \mid \mu_{ij} \sim \text{Bernoulli}(\mu_{ij})$ and with a logistic regression model, we have the exponential family representation

$$\begin{aligned}\Pr(Y_{ij} = y_{ij} \mid \boldsymbol{x}_{ij}) &= \mu_{ij}^{y_{ij}}(1 - \mu_{ij})^{1-y_{ij}} \\ &= \exp\left\{y_{ij}\theta_{ij} - \log[1 + \exp(\theta_{ij})]\right\},\end{aligned}$$

where

$$\theta_{ij} = \log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = \boldsymbol{x}_{ij}\boldsymbol{\gamma}.$$

For independent responses, the likelihood is

$$\begin{aligned}\Pr(\boldsymbol{Y} \mid \boldsymbol{x}) &= \exp\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}\theta_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n_i}\log[1+\exp(\theta_{ij})]\right\} \\ &= \exp\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} l_{ij}\right).\end{aligned}$$

To find the MLEs, we consider the score equation

$$\begin{aligned}\boldsymbol{G}(\gamma) = \frac{\partial l}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{\partial l_{ij}}{\partial \theta_{ij}}\frac{\partial \theta_{ij}}{\partial \boldsymbol{\gamma}} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}(y_{ij}-\mu_{ij}) = \sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\end{aligned}$$

with $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{in_i}]^{\mathrm{T}}$. This form is identical to the use of GEE with working independence and so can be implemented with standard software, though we need to "fix up" the standard errors via sandwich estimation. Hence, the above estimating equation construction offers a very simple approach to inference which may be adequate if the dependence between observations on the same unit is small. If the correlations are not small, then efficiency considerations suggest that nonindependence working covariance models should be entertained.

As with other types of data (Sect. 9.9), we can model the correlation structure (Liang and Zeger 1986) and assume $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i$ with $\boldsymbol{W}_i = \boldsymbol{\Delta}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{\Delta}_i^{1/2}$ with $\boldsymbol{\Delta}_i$ a diagonal matrix with jth diagonal entry $\mathrm{var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$ and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ a working correlation model depending on parameters $\boldsymbol{\alpha}$. In this case, the estimating function is

$$\boldsymbol{G}(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = \sum_{i=1}^{m}\boldsymbol{D}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i), \tag{9.24}$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\gamma}$. As usual, an estimate of $\boldsymbol{\alpha}$ is required, with an obvious choice being a method of moments estimator. The variance of the estimator takes the usual sandwich form (9.12).

9.14.2 Loglinear Models

We now consider another approach to constructing models for dependent binary data that may form the basis for likelihood or GEE procedures. Loglinear models are a

Table 9.10 Probabilities of the four possible outcomes for two binary variables via a loglinear representation

y_1	y_2	$\Pr(Y_1 = y_1, Y_2 = y_2)$
0	0	$c(\boldsymbol{\theta})$
1	0	$c(\boldsymbol{\theta}) \exp(\theta_1^{(1)})$
0	1	$c(\boldsymbol{\theta}) \exp(\theta_2^{(1)})$
1	1	$c(\boldsymbol{\theta}) \exp(\theta_1^{(1)} + \theta_2^{(1)} + \theta_{12}^{(2)})$

popular choice for cross-classified discrete data (Cox 1972; Bishop et al. 1975). We begin by returning to the situation in which we have n responses on a single unit, $y_j, j = 1, \ldots, n$. A saturated loglinear model is

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{n} \theta_j^{(1)} y_j + \sum_{j_1 < j_2} \theta_{j_1 j_2}^{(2)} y_{j_1} y_{j_2} + \ldots + \theta_{12\ldots n}^{(n)} y_1 \ldots y_n \right),$$

with $2^n - 1$ parameters $\boldsymbol{\theta} = [\theta_1^{(1)}, \ldots, \theta_n^{(1)}, \theta_{12}^{(2)}, \ldots, \theta_{n-1,n}^{(2)}, \ldots, \theta_{12\ldots n}^{(n)}]^{\mathrm{T}}$, and normalizing constant $c(\boldsymbol{\theta})$. To provide an interpretation of the parameters, consider the case of $n = 2$ trials for which

$$\Pr(Y_1 = y_1, Y_2 = y_2) = c(\boldsymbol{\theta}) \exp\left(\theta_1^{(1)} y_1 + \theta_2^{(1)} y_2 + \theta_{12}^{(2)} y_1 y_2 \right),$$

where $\boldsymbol{\theta} = [\theta_1^{(1)}, \theta_2^{(1)}, \theta_{12}^{(2)}]^{\mathrm{T}}$ and

$$c(\boldsymbol{\theta})^{-1} = \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} \exp\left(\theta_1^{(1)} y_1 + \theta_2^{(1)} y_2 + \theta_{12}^{(2)} y_1 y_2 \right).$$

Table 9.10 gives the forms of the probabilities for the loglinear representation, from which we can determine the interpretation of the three parameters:

$$\exp(\theta_1^{(1)}) = \frac{\Pr(Y_1 = 1 \mid y_2 = 0)}{\Pr(Y_1 = 0 \mid y_2 = 0)}$$

is the odds of an event at trial 1, given no event at trial 2,

$$\exp(\theta_2^{(1)}) = \frac{\Pr(Y_2 = 1 \mid y_1 = 0)}{\Pr(Y_2 = 0 \mid y_1 = 0)}$$

is the odds of an event at trial 2, given no event at trial 1, and

$$\exp(\theta_{12}^{(12)}) = \frac{\Pr(Y_2 = 1 \mid y_1 = 1) / \Pr(Y_2 = 0 \mid y_1 = 1)}{\Pr(Y_2 = 1 \mid y_1 = 0) / \Pr(Y_2 = 0 \mid y_1 = 0)}$$

is the ratio of the odds of an event at trial 2 given an event at trial 1, divided by the odds of an event at trial 2 given no event at trial 1. Consequently, if this parameter is larger than 1, there is positive dependence between Y_1 and Y_2.

For general n, a simplified version of the loglinear model is provided when third- and higher-order terms are set to zero, so that

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{n} \theta_j^{(1)} y_j + \sum_{j<k} \theta_{jk}^{(2)} y_j y_k \right). \tag{9.25}$$

For this model,

$$\frac{\Pr(Y_j = 1 \mid Y_k = y_k, Y_l = 0, l \neq j, k)}{\Pr(Y_j = 0 \mid Y_k = y_k, Y_l = 0, l \neq j, k)} = \exp(\theta_j^{(1)} + \theta_{jk}^{(2)} y_k).$$

so that $\exp(\theta_j^{(1)})$ is the (conditional) odds of an event at trial j, given all other responses are zero. Further, $\exp(\theta_{jk}^{(2)})$ is the odds ratio describing the association between Y_j and Y_k, given all other responses are set equal to zero, that is,

$$\frac{\Pr(Y_j = 1, Y_k = 1 \mid Y_l = 0, l \neq j, k) \Pr(Y_j = 0, Y_k = 0 \mid Y_l = 0, l \neq j, k)}{\Pr(Y_j = 1, Y_k = 0 \mid Y_l = 0, l \neq j, k) \Pr(Y_j = 0, Y_k = 1 \mid Y_l = 0, l \neq j, k)}$$
$$= \exp(\theta_{jk}^{(2)}).$$

The quadratic model (9.25) was described in Sect. 9.10 and was suggested for the analysis of binary data by Zhao and Prentice (1990). Recall that this model has the appealing property of consistency so long as the first two moments are correctly specified. The quadratic exponential model is unique in this respect.

Unfortunately, parameterizing in terms of the $\boldsymbol{\theta}$ parameters is unappealing for regression modeling where the primary aim is to model the response as a function of $\boldsymbol{x}$. To illustrate, consider binary longitudinal data with a binary covariate x and suppose we let the parameters $\boldsymbol{\theta}$ depend on x. The difference between the log odds $\theta_j^{(1)}(x = 1)$ and $\theta_j^{(1)}(x = 0)$ represents the effect of x on the *conditional* log odds of an event at period j, *given* that there were no events at any other trials, which is difficult to interpret. We would rather model the *marginal* means $\boldsymbol{\mu}$, and these are a function of both $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$. For example, for the $n = 2$ case presented in Table 9.10, the marginal means are

$$\mathrm{E}[Y_1] = c(\boldsymbol{\theta}) \exp(\theta_1^{(1)})[1 + \exp(\theta_1^{(1)} + \theta_{12}^{(2)})]$$
$$\mathrm{E}[Y_2] = c(\boldsymbol{\theta}) \exp(\theta_2^{(1)})[1 + \exp(\theta_2^{(1)} + \theta_{12}^{(2)})],$$

and these forms do not lend themselves to straightforward incorporation of covariates. Hence, alternative approaches have been proposed as we now discuss.

9.14.3 Further Multivariate Binary Models

A number of approaches are based on assuming a marginal mean model, to overcome the problems described in the previous section, along with a second set of parameters to model the dependence.

First, we may reparameterize the model via the mean vector $\boldsymbol{\mu}$ and second- and higher-order loglinear parameters. For example, we may consider second-order parameters only and work with $\boldsymbol{\mu}$ and the loglinear parameters $\boldsymbol{\theta}^{(2)}$, as suggested by Fitzmaurice and Laird (1993). The latter used maximum likelihood for estimation. There are two disadvantages to this approach. First, the interpretation of the $\boldsymbol{\theta}^{(2)}$ parameters depends on the number of responses n. This is particularly a problem in a longitudinal setting with differing n_i. Hence, this approach is most useful for data that have $n_i = n$ for all i. Second, if interest lies in understanding the structure of the dependence, the conditional odds ratio parameters do not have the attractive simple interpretation of *marginal* odds ratios.

A second approach is based on modeling the correlations in addition to the means. Let

$$\begin{aligned} e^{\star}_{ijk} &= \frac{Y_{ij} - \mu_{ij}}{[\mu_{ij}(1-\mu_{ij})]^{1/2}} \\ \rho_{ijk} &= \text{corr}(Y_{ij}, Y_{ik}) = \text{E}[e^{\star}_{ij} e^{\star}_{ik}] \\ \rho_{ijkl} &= \text{E}[e^{\star}_{ij} e^{\star}_{ik} e^{\star}_{il}] \\ \cdots &\quad \cdots \\ \rho_{i1\ldots n_i} &= \text{E}[e^{\star}_{i1} e^{\star}_{i2} \ldots e^{\star}_{in_i}]. \end{aligned}$$

The correlations have marginal interpretations. For example, ρ_{ijkl} is a three-way association parameter. Bahadur (1961) defined a multivariate binary model based on the marginal means and these correlations. The probability for the set of outcomes on unit i is

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{1-y_{ij}} \times \left(1 + \sum_{j<k} \rho_{ijk} e^{\star}_{ij} e^{\star}_{ik} + \sum_{j<k<l} \rho_{ijkl} e^{\star}_{ij} e^{\star}_{ik} e^{\star}_{il} + \ldots + \rho_{i1\ldots n} e^{\star}_{i1} e^{\star}_{i2} \ldots e^{\star}_{in_i}\right).$$

Unfortunately, the correlations are constrained in complicated ways by the marginal means. As an example, consider two measurements on a single individual, Y_{i1} and Y_{i2}, with means μ_{i1} and μ_{i2}. The correlation is

$$\text{corr}(Y_{i1}, Y_{i2}) = \frac{\Pr(Y_{i1} = 1, Y_{i2} = 1) - \mu_{i1}\mu_{i2}}{[\mu_{i1}(1-\mu_{i1})\mu_{i2}(1-\mu_{i2})]^{1/2}}$$

Table 9.11 Notation in the case of $n_i = 2$ binary responses on individual i

		Y_{i2}		
		0	1	
Y_{i1}	0	$1 - \mu_{i1} - \mu_{i2} + \mu_{i12}$	$\mu_{i2} - \mu_{i12}$	$1 - \mu_{i1}$
	1	$\mu_{i1} - \mu_{i12}$	μ_{i12}	μ_{i1}
		$1 - \mu_{i2}$	μ_{i2}	

and

$$\max(0, \mu_{i1} + \mu_{i2} - 1) \leq \Pr(Y_{i1} = 1, Y_{i2} = 1) \leq \min(\mu_{i1}, \mu_{i2}),$$

which implies complicated constraints on the correlation. For example, if $\mu_{i1} = 0.8$ and $\mu_{i2} = 0.2$, then $0 \leq \text{corr}(Y_{i1}, Y_{i2}) \leq 0.25$. The message here is that correlations are not a natural measure of dependence for binary data so that the Bahadur representation is not appealing.

A third approach (Lipsitz et al. 1991; Liang et al. 1992) is to parameterize in terms of the marginal means and the marginal odds ratios defined by. Let

$$\begin{aligned}\delta_{ijk} &= \frac{\Pr(Y_{ij} = 1, Y_{ik} = 1) \Pr(Y_{ij} = 0, Y_{ik} = 0)}{\Pr(Y_{ij} = 1, Y_{ik} = 0) \Pr(Y_{ij} = 0, Y_{ik} = 1)} \\ &= \frac{\Pr(Y_{ij} = 1 \mid Y_{ik} = 1) / \Pr(Y_{ij} = 0 \mid Y_{ik} = 1)}{\Pr(Y_{ij} = 1 \mid Y_{ik} = 0) / \Pr(Y_{ij} = 0 \mid Y_{ik} = 0)},\end{aligned}$$

which is the odds (for individual i) that the jth observation is a 1, given the kth observation is a 1, divided by the odds that the jth observation is a 1, given the kth observation is a 0. Therefore, we have a set of marginal odds ratios, and if $\delta_{ijk} > 1$, we have positive dependence between outcomes j and k. It is then possible to obtain the joint distribution in terms of the means $\boldsymbol{\mu}$, where $\mu_{ij} = \Pr(Y_{ij} = 1)$, the odds ratios $\boldsymbol{\delta}_i = [\delta_{i12}, \ldots, \delta_{i,n_i-1,n_i}]$ and contrasts of odds ratios. To determine the probability distribution of the data, we need to find

$$\mu_{ijk} = \mathrm{E}[Y_{ij} Y_{ik}] = \Pr(Y_{ij} = 1, Y_{ik} = 1),$$

so that we can write down either the likelihood function or an estimating function.

For the case of $n_i = 2$ (see Table 9.11), we have

$$\delta_{i12} = \frac{\Pr(Y_{i1} = 1, Y_{i2} = 1) \Pr(Y_{i1} = 0, Y_{i2} = 0)}{\Pr(Y_{i1} = 1, Y_{i2} = 0) \Pr(Y_{i1} = 0, Y_{i2} = 1)} = \frac{\mu_{i12}(1 - \mu_{i1} - \mu_{i2} + \mu_{i12})}{(\mu_{i1} - \mu_{i12})(\mu_{i2} - \mu_{i12})},$$

and so

$$\mu_{i12}^2(\delta_{i12} - 1) + \mu_{i12} b_i + \delta_{i12} \mu_{i1} \mu_{i2} = 0,$$

where $b_i = (\mu_{i1} + \mu_{i2})(1 - \delta_{i12}) - 1$, to give

$$\mu_{i12} = \frac{-b_i \pm \sqrt{b_i^2 - 4(\delta_{i12} - 1)\mu_{i1}\mu_{i2}\delta_{i12}}}{2(\delta_{i12} - 1)}$$

if $\delta_{i12} \neq 1$ and $\mu_{i12} = \mu_{ij}\mu_{ik}$ if $\delta_{i12} = 1$. The likelihood is

$$\mu_{i1}^{y_{i1}}(1-\mu_{i1})^{1-y_{i1}}\mu_{i2}^{y_{i2}}(1-\mu_{i2})^{1-y_{i2}} + (-1)^{(y_{i1}-y_{i2})}(\mu_{i12} - \mu_{i1}\mu_{i2}) \quad (9.26)$$

(Exercise 9.3).

As the number of binary responses increases so does the complexity of solving for the μ_{ijk}'s; see Liang et al. (1992) for further details. In the case of large n_i, there are a large numbers of nuisance odds ratios, and assumptions such as $\delta_{ijk} = \delta$ for $i = 1, \ldots, m, j, k = 1, \ldots, n_i$ may be made.

In a longitudinal setting, another possibility is to take

$$\log \delta_{ijk} = \alpha_0 + \alpha_1 |t_{ij} - t_{ik}|^{-1},$$

so that the degree of association is inversely proportional to the time between observations. Computation may be carried out by setting up an estimating equation for $\boldsymbol{y}_i$ and a method of moments estimator for estimation of the covariance parameters. As an alternative, GEE2 may be used with a pair of linked estimating equations (Sect. 9.10).

Letting $\alpha_{ijk} = \log \delta_{ijk}$, Carey et al. (1993) suggest the following approach for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. It is easy to show that

$$\begin{aligned}\frac{\Pr(Y_{ij}=1 \mid Y_{ik}=y_{ik})}{\Pr(Y_{ij}=0 \mid Y_{ik}=y_{ik})} &= \exp(y_{ik}\alpha_{ijk})\frac{\Pr(Y_{ij}=1, Y_{ik}=0)}{\Pr(Y_{ij}=0, Y_{ik}=0)} \\ &= \exp(y_{ik}\alpha_{ijk})\left(\frac{\mu_{ij}-\mu_{ijk}}{1-\mu_{ij}-\mu_{ik}+\mu_{ijk}}\right),\end{aligned}$$

which can be written as a logistic regression model for the conditional probabilities $\mathrm{E}[Y_{ij} \mid Y_{ik}]$:

$$\begin{aligned}\text{logit}\,(\mathrm{E}[Y_{ij} \mid Y_{ik}]) &= \log\left(\frac{\Pr(Y_{ij}=1 \mid Y_{ik}=y_{ik})}{\Pr(Y_{ij}=0 \mid Y_{ik}=y_{ik})}\right) \\ &= y_{ik}\alpha_{ijk} + \log\left(\frac{\mu_{ij}-\mu_{ijk}}{1-\mu_{ij}-\mu_{ik}+\mu_{ijk}}\right)\end{aligned}$$

where the term on the right is an offset (given estimates of the means). Suppose, for simplicity, that $\alpha_{ijk} = \alpha$. Then, given current estimates of $\boldsymbol{\beta}, \alpha$, we can fit a logistic regression model by regressing Y_{ij} on Y_{ik} for $1 \leq j < k \leq n_i$, to reestimate α. The offset is a function of α and $\boldsymbol{\beta}$ so iteration is required. Consequently, Carey et al. (1993) named this approach *alternating logistic regressions*. Once the α parameters are estimated, one may solve for $\text{var}(\boldsymbol{Y}_i)$ in order to use the estimating function (9.24).

In some situations, interest may focus on estimating/modeling the within-unit dependence. Basing a model on correlation parameters is not appealing, but using marginal log odds ratios suggests the model $\alpha_{ijk} = \boldsymbol{x}_{ijk}^{\star}\boldsymbol{\Psi}$ for a set of covariates of interest $\boldsymbol{x}_{ijk}^{\star}$ with associated regression coefficients $\boldsymbol{\Psi}$.

Table 9.12 GEE parameter estimates for the contraception data

	GEE independence		GEE exchangeable		GEE ALR[a]	
Parameter	Est.	Std. err.	Est.	Std. err.	Est.	Std. err.
Intercept	−2.2	0.18	−2.2	0.18	−2.3	0.16
Low-dose time	0.67	0.16	0.70	0.16	0.70	0.15
Low-dose time^2	−0.030	0.033	−0.033	0.032	−0.033	0.031
High-dose time	0.30	0.11	0.33	0.11	0.34	0.11
High-dose time^2	−0.062	0.030	−0.064	0.029	−0.067	0.028

[a] Alternating logistic regression

Example: Contraception Data

Table 9.12 gives parameter estimates and standard errors for various implementations of GEE, for the marginal model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_0 + \gamma_1 t_{ij} + \gamma_2 t_{ij}^2 + \gamma_3 d_i t_{ij} + \gamma_4 d_i t_{ij}^2, \tag{9.27}$$

where the γ notation emphasizes that we are estimating marginal parameters. We initially implement GEE with working independence; in general, this is not to be recommended unless it is thought that the outcomes within a cluster are close to independent. We also allow a working exchangeable structure, with the latter parameterized in terms of correlations. Finally, we assume a working exchangeable model parameterized in terms of a common (marginal) log odds ratio. For these data, there are few substantive differences between the approaches. Under the exchangeable models, the common correlation is estimated as 0.36 (0.024) (which is in line with the correlations in Table 9.1), while the common log odds ratio is estimated as 2.0 (0.11). The latter is log of the ratio of the the odds of amenorrhea at time t, given amenorrhea at time s, to the odds of amenorrhea at time t, given no amenorrhea at time s, $s \neq t$.

We may compare these results with a random intercept GLMM. The Bayesian marginal estimates obtained by dividing the posterior means and the posterior standard deviations by $(c^2\widehat{\sigma}_0^2 + 1)^{1/2}$ result in the estimates (standard errors): −2.3 (0.17), 0.68 (0.15), −0.019 (0.032), 0.34 (0.11), and −0.066 (0.035), which are in close agreement with the point and interval estimates in Table 9.12. The marginal probabilities from the GEE exchangeable model were identical to those obtained via Monte Carlo integration in the mixed model (and displayed on Fig. 9.7).

As we have already mentioned, model checking is very difficult with binary data. For data with replication across common $\boldsymbol{x}$ variables, one may obtain empirical probabilities and/or logits (as in Fig. 9.1), which may suggest model forms in an exploratory model building exercise or may be compared with fitted summaries. Similarly, the dependence structure may be examined across covariate groups, via empirical correlations or odds.

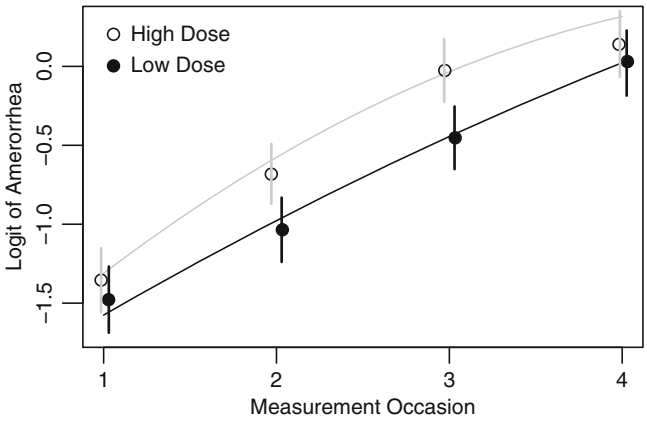

Fig. 9.8 Logit of probability of amenorrhea over time in high- and low- dose groups with marginal fits from exchangeable GEE model

Figure 9.8 shows the fitted logistic curves in each dose group versus time along with the logits of the probabilities of amenorrhea. The vertical lines represent 95% confidence intervals for the logits. These intervals increase slightly in width over time as dropout occurs. Here, we would conclude that the model fit is reasonable.

9.15 Nonlinear Mixed Models

We now turn attention to the nonlinear mixed model (NLMM). Our development will be much shorter for this class of models. One reason for this is that the nonlinearity results in very little analytical theory being available. Also, traditionally, dependent nonlinear data have been analyzed with mixed models and not GEE because the emphasis is often on unit-level inference. The fitting, inferential summarization and assessment of assumptions will be illustrated using the theophylline data described in Sect. 9.2.3.

In a nonlinear mixed model (NLMM), the first stage of a linear mixed model is replaced by a nonlinear form. We describe a specific two-stage form that is useful in many longitudinal situations. The response at time t_{ij} is y_{ij}, and $\boldsymbol{x}_{ij}$ are covariates measured at these times, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Let $N = \sum_{i=1}^{m} n_i$:

Stage One: Conditional on random effects, $\boldsymbol{b}_i$, the response model is

$$y_{ij} = f(\eta_{ij}, t_{ij}) + \epsilon_{ij}, \tag{9.28}$$

where $f(\cdot,\cdot)$ is a nonlinear function and

$$\eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i,$$

with a $(k+1)\times 1$ vector of fixed effects $\boldsymbol{\beta}$, a $(q+1)\times 1$ vector of random effects, $\boldsymbol{b}_i$, with $q \le k$, $\boldsymbol{x}_i = [\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{in_i}]^{\mathrm{T}}$ the design matrix for the fixed effect with $\boldsymbol{x}_{ij} = [1, x_{ij1},\ldots,x_{ijk}]^{\mathrm{T}}$ and $\boldsymbol{z}_i = [\boldsymbol{z}_{i1},\ldots,\boldsymbol{z}_{in_i}]^{\mathrm{T}}$ the design matrix for the random effects with $\boldsymbol{z}_{ij} = [1, z_{ij1},\ldots,z_{ijq}]^{\mathrm{T}}$.

Stage Two: Random terms:

$$\begin{aligned} \mathrm{E}[\boldsymbol{\epsilon}_i] &= \mathbf{0}, \;\; \mathrm{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{E}_i(\boldsymbol{\alpha}), \\ \mathrm{E}[\boldsymbol{b}_i] &= \mathbf{0}, \;\; \mathrm{var}(\boldsymbol{b}_i) = \boldsymbol{D}(\boldsymbol{\alpha}), \\ \mathrm{cov}(\boldsymbol{b}_i, \boldsymbol{\epsilon}_i) &= \mathbf{0} \end{aligned}$$

where $\boldsymbol{\alpha}$ is the vector of variance–covariance parameters. A common model assumes

$$\begin{aligned} \boldsymbol{\epsilon}_i &\sim_{ind} \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}), \\ \boldsymbol{b}_i &\sim_{iid} \mathrm{N}(\mathbf{0}, \boldsymbol{D}). \end{aligned}$$

For this model, $\boldsymbol{\alpha} = [\sigma_\epsilon^2, \boldsymbol{D}]$.

For nonlinear models even the first two moments are not available in closed form. In general:

$$\mathrm{E}[Y_{ij}] = \mathrm{E}_{\boldsymbol{b}_i}[f(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i, t_{ij})] \neq f(\boldsymbol{x}_{ij}\boldsymbol{\beta}, t_{ij})$$

where $f(\boldsymbol{x}_{ij}\boldsymbol{\beta}, t_{ij})$ is the nonlinear curve evaluated at $\boldsymbol{b}_i = \mathbf{0}$. Hence, unlike the LMM, the nonlinear curve at a time point averaged across individuals is not equal to the nonlinear curve at that time for an average individual (i.e., one with $\boldsymbol{b}_i = \mathbf{0}$). The variance is

$$\mathrm{var}(Y_{ij}) = \sigma_\epsilon^2 + \mathrm{var}_{\boldsymbol{b}_i}[f(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i, t_{ij})]$$

so that the marginal variance of the response is not constant across time, even when we have a random intercepts only model (unlike the LMM). For responses on the same individual, dependence is induced through the common random effects:

$$\mathrm{cov}(Y_{ij}, Y_{ij'}) = \mathrm{cov}_{\boldsymbol{b}_i}[f(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i, t_{ij}), f(\boldsymbol{x}_{ij'}\boldsymbol{\beta} + \boldsymbol{z}_{ij'}\boldsymbol{b}_i, t_{ij'})]$$

but, as with the GLMM, there is no closed form for the covariance. Finally, for observations on different individuals:

$$\mathrm{cov}(Y_{ij}, Y_{i'j'}) = 0$$

for $i \neq i'$. The data do not have a closed-form marginal distribution. These forms illustrate that picking particular random effect structures cannot be based on specific requirements in terms of the marginal variance and covariance. Rather, this choice should be based on the context and on data availability.

In a NLMM, the interpretation of parameters is usually tied to the particular model. In a GLMM, one can make use of linearity on the linear predictor scale to have an interpretation in terms of unit changes in covariates (as we have illustrated for loglinear and logistic linear models). In a NLMM, this will not be possible, however (since the model is *nonlinear*!).

We next briefly consider parameterization of the model, before considering likelihood and Bayesian inference in Sects. 9.17 and 9.18, respectively. A GEE approach is briefly considered in Sect. 9.19, but as previously mentioned, this is not as popular as likelihood and Bayes approaches, and so this section is short. The nonlinearity of the model means there is no sufficient statistic for $\boldsymbol{\beta}$, and so conditional likelihood cannot be used.

9.16 Parameterization of the Nonlinear Model

In contrast to LMMs and GLMMs, there is no obvious way to parameterize a NLMM, and the way one proceeds is an art form. Given the normal random effects distribution, one usually parameterizes to quantities on the whole real line. This issue relates to the discussion of the solution locus and the parameterization of nonlinear models given in Sect. 6.15.

Example: A Simple Pharmacokinetic Model

The simplest pharmacokinetic model is

$$\mathrm{E}[Y \mid V, k_e] = \frac{D}{V} \exp(-k_e t)$$

where D is the known dose, $V > 0$ is the volume of distribution, and $k_e > 0$ is the elimination rate constant. The obvious parameterization is $\beta_0 = \log V, \beta_1 = \log k_e$. A key parameter of interest is the clearance, defined as $Cl = V \times k_e$, and so one may alternatively take $\beta_1^\star = \log Cl$ with $\beta_0^\star = \beta_0$ as before. This parameterization has a number of advantages. A first advantage is that the clearance for individual i is often modeled as a function of covariates, for example, via a loglinear model of the form

$$\log Cl = \alpha_0 + \alpha_1 x_i \tag{9.29}$$

where x_i is a covariate of interest such as weight. A second advantage is that the clearance is a very stable parameter to estimate. The clearance is the dose D divided by the area under the concentration–time curve, and this area tends to be very well estimated (unless there are few sample points at large times) and hence so does the clearance, Cl.

If a Bayesian approach is adopted, then the prior must clearly be specific to the parameterization. For example, for $\boldsymbol{\beta} = [\beta_0, \beta_1]^{\mathrm{T}}$ and $\boldsymbol{\beta}^* = [\beta_0^*, \beta_1^*]^{\mathrm{T}}$ the prior $\boldsymbol{\beta} \sim \mathrm{N}_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with fixed $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$, will clearly give different inference to assuming $\boldsymbol{\beta}^\star \sim \mathrm{N}_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. □

There is some theoretical work on choosing parameterizations (Bates and Watts 1980), but good parameterizations are often found through experience with particular models. The accuracy of asymptotic approximations is also crucially dependent on the choice of parameterization, with stable parameters likely to display good asymptotic properties. The examination of likelihood contours (as was done in Sect. 6.12) can indicate whether asymptotic distributions are likely to be accurate or not.

With many nonlinear models, care must be taken to ensure the model is identifiable in the sense that if $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, $f(\boldsymbol{\theta}) \neq f(\boldsymbol{\theta}')$. If there is non-identifiability, then one may either reparameterize the model or enforce identifiability through the prior. The latter can be messy, however.

Unfortunately, preserving identifiability and retaining an interpretable parameter cannot usually be simultaneously achieved. We illustrate the problems with an example.

Example: Pharmacokinetics of Theophylline

As discussed in Sect. 6.2, the one-compartment open model is non-identifiable. We illustrate by parameterizing as $[k_e, k_a, Cl]$ to give the mean model, for a generic individual, as

$$\mathrm{E}[Y] = \frac{D k_e k_a}{Cl(k_a - k_e)} \left[\exp(-k_e t) - \exp(-k_a t)\right]. \tag{9.30}$$

This form is known as the "flip-flop" model because the parameters $[k_e, k_a, Cl]$ give the same curve as the parameters $[k_a, k_e, Cl]$. To enforce identifiability, it is typical to assume that $k_a > k_e > 0$, since for many drugs, absorption is faster than elimination. This suggests the parameterization $[\log k_e, \log(k_a - k_e), \log Cl]$.

9.17 Likelihood Inference for the Nonlinear Mixed Model

As with the linear mixed and generalized linear mixed models already considered, the likelihood is defined with respect to fixed effects $\boldsymbol{\beta}$ and variance components $\boldsymbol{\alpha}$:

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{m} \int_{\boldsymbol{b}_i} p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma_\epsilon^2) \times p(\boldsymbol{b}_i \mid \boldsymbol{D})\, d\boldsymbol{b}_i, \tag{9.31}$$

with $\boldsymbol{\alpha} = [\boldsymbol{D}, \sigma_\varepsilon^2]$.

The first difficulty to overcome is how to calculate the required integrals, which for nonlinear models are analytically intractable (recall for the LMM they were available in closed form). As with the GLMM, two obvious approaches are to resort to Laplace approximations or adaptive Gauss–Hermite. Pinheiro and Bates (2000, Chap. 7) contains extensive details on these approaches (see also Bates 2011). We wish to evaluate

$$p(\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = (2\pi\sigma_\epsilon^2)^{-n_i/2}(2\pi)^{-(q+1)/2}|\boldsymbol{D}|^{-1/2} \int \exp[\, n_i g(\boldsymbol{b}_i)\,]\, d\boldsymbol{b}_i,$$

where

$$-2n_i g(\boldsymbol{b}_i) = [\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i)]^{\mathrm{T}}[\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_i)]/\sigma_\epsilon^2 + \boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i \tag{9.32}$$

and

$$\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) = [f(\boldsymbol{x}_{i1}\boldsymbol{\beta} + \boldsymbol{z}_{i1}\boldsymbol{b}_i, t_{i1}), \ldots, f(\boldsymbol{x}_{in_i}\boldsymbol{\beta} + \boldsymbol{z}_{in_i}\boldsymbol{b}_i, t_{in_i})]^{\mathrm{T}}.$$

The Laplace approximation (Sect. 3.7.2) is a second-order Taylor series expansion of $g(\cdot)$ about

$$\widehat{\boldsymbol{b}}_i = \arg\min_{\boldsymbol{b}_i} \; [-g(\boldsymbol{b}_i)]$$

where this minimization constitutes a penalized least squares problem. For a nonlinear model, numerical methods are required for this minimization, but the dimensionality, $q+1$, is typically small. With respect to (9.31), the second difficulty is how to maximize the likelihood as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$; again see Pinheiro and Bates (2000) and Bates (2011) for details.

In terms of the random effects, empirical Bayes estimates may be calculated, as with the GLMM. In the example that follows, we evaluate the MLEs using the procedure described in Lindstrom and Bates (1990) in which estimates of $\boldsymbol{b}_i$ are $\boldsymbol{\beta}$ are first obtained by minimizing the penalized least squares criteria (9.32), given estimates of $\boldsymbol{D}$ and σ_ϵ^2. Then a first-order Taylor series expansion of $\boldsymbol{f}_i$ about the current estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ is carried out, which results in a LMM. For such a model, the random effects may be integrated out analytically, and the subsequent (approximate) likelihood can be maximized with respect to $\boldsymbol{D}$ and σ_ϵ^2. This procedure is then iterated until convergence.

Approximate inference for $[\boldsymbol{\beta}, \boldsymbol{\alpha}]$ is carried out via asymptotic normality of the MLE:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \boldsymbol{I}_{\beta\alpha} \\ \boldsymbol{I}_{\alpha\beta} & \boldsymbol{I}_{\alpha\alpha} \end{bmatrix}^{-1} \right)$$

where $\boldsymbol{I}_{\beta\beta}$, $\boldsymbol{I}_{\beta\alpha}$, $\boldsymbol{I}_{\alpha\beta}$, and $\boldsymbol{I}_{\alpha\alpha}$ are the relevant information matrices.

Many approximation strategies have been suggested for nonlinear hierarchical models, but care is required since validity of the asymptotic distribution depends on the approximation used. For example, a historically popular approach (Beal and Sheiner 1982) was to carry out a first-order Taylor series about $\mathrm{E}[\boldsymbol{b}_i] = \mathbf{0}$ to give

$$\begin{aligned} y_{ij} &= f(\boldsymbol{x}_{ij}\boldsymbol{\beta}_i + \boldsymbol{z}_{ij}\boldsymbol{b}_i, t_{ij}) + \epsilon_{ij} \\ &\approx f(\boldsymbol{x}_{ij}\boldsymbol{\beta}_i, t_{ij}) + \boldsymbol{b}_i^{\mathrm{T}} \left. \frac{\partial f}{\partial \boldsymbol{b}_i} \right|_{\boldsymbol{b}_i = \mathbf{0}} + \epsilon_{ij}. \end{aligned}$$

This first-order estimator is inconsistent, however, and has bias even if n_i and m both go to infinity; see Demidenko (2004, Chap. 8).

Example: Pharmacokinetics of Theophylline

For these data, the one-compartment model with first-order absorption and elimination is a good starting point for analysis. This model was described in some detail in Sect. 6.16.3. The mean concentration at time point t_{ij} for subject i is

$$\frac{D_i k_{ai} k_{ei}}{Cl_i(k_{ai} - k_{ei})} \left[\exp(-k_{ei}t_{ij}) - \exp(k_{ai}t_{ij})\right], \tag{9.33}$$

where we have parameterized in terms of $[Cl_i, k_{ai}, k_{ei}]$ and D_i is the initial dose.

We first fit the above model to each individual, using nonlinear least squares; Fig. 9.9 gives the resultant 95% asymptotic confidence intervals. The between-individual variability is evident, particularly for $\log k_a$. Figure 9.10 displays the data along with the fitted curves. The general shape of the curve seems reasonable, but the peak is missed for a number of individuals (e.g., numbers 10, 1, 5, and 9).

Turning now to a NLMM, we assume that each of the parameters is treated as a random effect so that

$$\log k_{ei} = \beta_1 + b_{1i} \tag{9.34}$$

$$\log k_{ai} = \beta_2 + b_{2i} \tag{9.35}$$

$$\log Cl_i = \beta_3 + b_{3i} \tag{9.36}$$

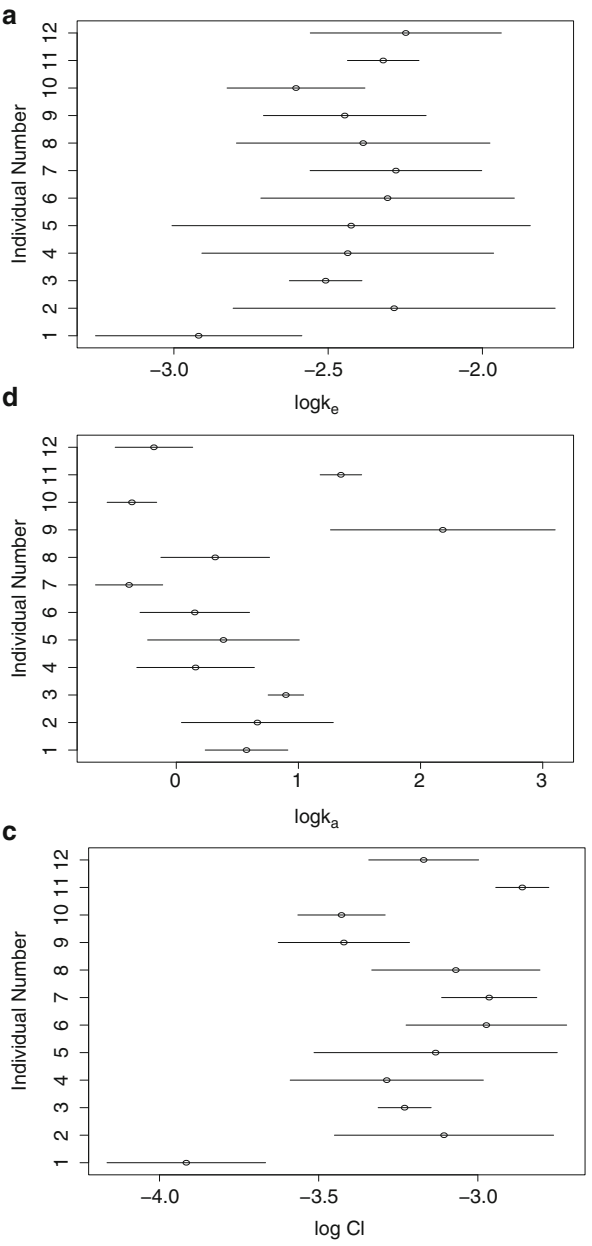

Fig. 9.9 95% confidence intervals for each of the three parameters and 12 individuals in the theophylline data. Obtained via individual fitting

with $\boldsymbol{b}_i \mid \boldsymbol{D} \sim \mathrm{N}_3(\boldsymbol{0}, \boldsymbol{D})$ where $\boldsymbol{b}_i = [b_{i1}, b_{i2}, b_{i3}]^{\mathrm{T}}$. The estimates resulting from the Lindstrom and Bates (1990) method described in the previous section are given in Table 9.13. The standard deviation of the random effects for $\log k_a$ is large, as we anticipated from examination of Fig. 9.9.

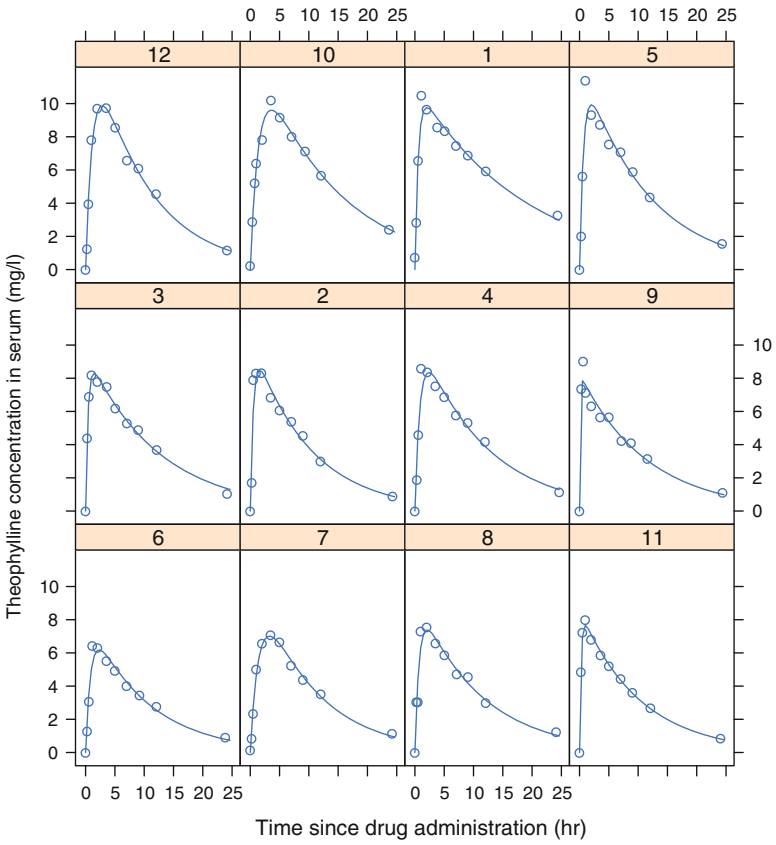

Fig. 9.10 Concentrations versus time for 12 individuals given the drug theophylline, along with individual nonlinear least squares fits

9.18 Bayesian Inference for the Nonlinear Mixed Model

The first two stages of the model are as in the likelihood formulation. We first discuss how hyperpriors may be specified, before discussing inference for functions of interest.

9.18.1 Hyperpriors

A Bayesian approach requires a prior distribution for $\boldsymbol{\beta}, \boldsymbol{\alpha}$. As with the LMM, a proper prior is required for the matrix $\boldsymbol{D}$. In contrast to the LMM, a proper prior is required for $\boldsymbol{\beta}$ also, to ensure the propriety of the posterior distribution.

Table 9.13 Comparison of likelihood and Bayesian NLMM estimation techniques for the theophylline data

		Likelihood		Bayes normal		Bayes lognorm		Bayes power	
PK label	Parameter	Est.	(s.e.)	Est.	(s.d.)	Est.	(s.d.)	Est.	(s.d.)
$\log\ k_e$	β_1	−2.43	(0.063)	−2.46	(0.077)	−2.43	(0.075)	−2.25	(0.083)
$\log\ k_a$	β_2	0.45	(0.20)	0.47	(0.19)	0.26	(0.23)	0.45	(0.22)
$\log\ Cl$	β_3	−3.21	(0.081)	−3.23	(0.082)	−3.22	(0.090)	−3.22	(0.092)
$\log\ k_e$	$\sqrt{D_{11}}$	0.13	(–)	0.19	(0.049)	0.22	(0.059)	0.23	(0.061)
$\log\ k_a$	$\sqrt{D_{22}}$	0.64	(–)	0.62	(0.15)	0.72	(0.19)	0.69	(0.18)
$\log\ Cl$	$\sqrt{D_{33}}$	0.25	(–)	0.25	(0.051)	0.30	(0.071)	0.29	(0.072)

For the likelihood summaries, we report the MLEs and the asymptotic standard errors, while for the Bayesian analysis, we report the mean and standard deviation of the posterior distribution. The three Bayesian models differ in the error models assumed at the first stage with normal, lognormal, and power models being considered

If parameters occur linearly, then proper priors are not required, but, as usual, the safest strategy is to specify proper priors.

For simplicity, we assume that random effects are associated with all parameters and as, in Sect. 8.6.3, parameterize the model as $\tau = \sigma_\epsilon^{-2}$, $\boldsymbol{W} = \boldsymbol{D}^{-1}$, and $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{b}_i$ for $i = 1, \ldots, m$, with the dimensionality of $\boldsymbol{\beta}_i$ being $k + 1$. The joint posterior is

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \tau, \boldsymbol{\beta}, \boldsymbol{W} \mid \boldsymbol{y}) \propto \prod_{i=1}^{m} \left[p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W}) \right] \pi(\boldsymbol{\beta}) \pi(\tau) \pi(\boldsymbol{W}).$$

We assume the priors

$$\boldsymbol{\beta} \sim \mathrm{N}_{k+1}(\boldsymbol{\beta}_0, \boldsymbol{V}_0), \quad \tau \sim \mathrm{Ga}(a_0, b_0), \quad \boldsymbol{W} \sim \mathrm{Wish}_{k+1}(r, \boldsymbol{R}^{-1}),$$

for further discussion of this specification, see Sect. 8.6.2. Closed-form inference is unavailable, but MCMC is almost as straightforward as in the LMM case. The INLA approach is not (at time of writing) available for the Bayesian analysis of nonlinear models. With respect to MCMC, the conditional distributions for $\boldsymbol{\beta}, \tau, \boldsymbol{W}$ are unchanged from the linear case. There is no closed-form conditional distribution for $\boldsymbol{\beta}_i$, which is given by

$$p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \tau, \boldsymbol{W}, \boldsymbol{y}) \propto p(\boldsymbol{y}_i \mid \boldsymbol{\beta}_i, \tau) \times p(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{W})$$

but a Metropolis–Hastings step can be used (to give a Metropolis within Gibbs algorithm, as described in Sect. 3.8.5).

9.18.2 Inference for Functions of Interest

We discuss prior choice and inferential summaries in the context of fitting a NLMM to the theophylline data. For these data, the parameterization

$$\boldsymbol{\beta}_i = [\log k_{ei}, \log k_{ai}, \log Cl_i]$$

was initially adopted, with random effects normal prior $\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} \sim_{iid} \mathrm{N}_3(\boldsymbol{\beta}, \boldsymbol{D})$. We assume independent normal priors for the elements of $\boldsymbol{\beta}$, centered at 0 and with large variances (recall that we need proper priors). For $\boldsymbol{D}^{-1}$, we assume a Wishart$(r, \boldsymbol{R}^{-1})$ distribution with diagonal $\boldsymbol{R}$ (see Sect. 8.6.2 and Appendix D for discussion of the Wishart distribution). We describe the procedure that is followed in order to choose the diagonal elements.

Consider a generic univariate "natural" parameter θ (e.g., k_e, k_a, or Cl) for which we assume the lognormal prior LogNorm(β, D). Pharmacokineticists have insight into the coefficient of variation for θ, that is, $\mathrm{CV}(\theta) = \mathrm{sd}(\theta)/\mathrm{E}[\theta]$. Recall the first two moments of a lognormal

$$\begin{aligned}
\mathrm{E}[\theta] &= \exp(\beta + D/2) \\
\mathrm{var}(\theta) &= \mathrm{E}[\theta]^2[\exp(D) - 1] \\
\mathrm{sd}(\theta) &= \mathrm{E}[\theta]\sqrt{\exp(D) - 1} \\
&\approx \mathrm{E}[\theta]\sqrt{D}
\end{aligned}$$

so that

$$\mathrm{CV}(\theta) \approx \sqrt{D}.$$

We can therefore assign a prior for D by providing a prior estimate of $\sqrt{D}$. Under the Wishart parameterization, we have adopted $\mathrm{E}[\boldsymbol{D}^{-1}] = r\boldsymbol{R}^{-1}$. We take $r = 3$ (which is the smallest integer that gives a proper prior) and $\boldsymbol{R} = \mathrm{diag}(1/5, 1/5, 1/5)$ which gives $\mathrm{E}[D_{kk}^{-1}] = 15$ so that, for $k = 1, 2, 3$, $\mathrm{E}[\sqrt{D}_{kk}] \approx 1/\sqrt{15} = 0.26$, or an approximate prior expectation of the coefficient of variation of 26%, which is reasonable in this context (Wakefield et al. 1999).

For inference, again consider a generic parameter θ with prior LogNorm(β, D). The mode, median, and mean of the *population distribution* of θ are

$$\exp(\beta - \sqrt{D}), \qquad \exp(\beta), \qquad \exp(\beta + D/2),$$

respectively. Further, $\exp(\beta \pm 1.96\sqrt{D})$ is a 95% interval for θ in the population. Consequently, given samples from the posterior $p(\beta, D \mid \boldsymbol{y})$, one may simply convert to samples for any of these summaries.

In a pharmacokinetic context, interest often focuses on various functions of the natural parameters. As a first example, consider the terminal half-life which is given

by $t_{1/2} = k_e^{-1} \log 2$. In the parameterization adopted in the theophylline study, $\log k_e \sim \mathrm{N}(\beta_1, D_{11})$, and so the distribution of the log half-life is normal also:

$$\log t_{1/2} \sim \mathrm{N}[\log(\log 2) - \beta_1, D_{11}]$$

which simplifies inference since one can summarize the population distribution in the same way as was just described for a generic parameter θ. Other parameters of interest are not simple linear combinations, however. For example, the time to maximum is

$$t_{\max} = \frac{1}{k_a - k_e} \log\left(\frac{k_a}{k_e}\right)$$

and the maximum concentration is

$$\begin{aligned} \mathrm{E}[Y \mid t_{\max}] &= \frac{Dk_a}{V(k_a - k_e)} [\exp(-k_e t_{\max}) - \exp(-k_a t_{\max})] \\ &= \frac{D}{V}\left(\frac{k_a}{k_e}\right)^{k_a/(k_a - k_e)}. \end{aligned}$$

For such summaries, the population distribution may be examined by simulating parameter sets $[\log k_e, \log k_a, \log Cl]$ for new individuals from the population distribution, and then converting to the functions of interest.

As noted in Sect. 9.16, the parameterization $[\log k_e, \log k_a, \log Cl]$ that we have adopted is non-identifiable since the same likelihood values are achieved with the set $[\log k_a, \log k_e, \log Cl]$. For the theophylline data, we performed MCMC with two chains, and one of the chains "flipped" between the two non-identifiable regions in the parameter space, as illustrated in Fig. 9.11 (note that in panels (a) and (b), the vertical axes have the same scale). In this plot the three population parameters β_1, β_2, β_3 are plotted in the three rows. Here, the labeling of β_1 and β_2 is arbitrary. The parameter β_3 is unaffected by the flip-flop behavior because the mean log clearance is the same under each nonidentifiable set. In Fig. 9.11(a), the chain represented by the solid line corresponds to the smaller of the two rate constants and, after a small period of burn-in, remains in the region of the parameter space corresponding to the smaller constant. In contrast, the chain represented by the dotted line flips to the region corresponding to the larger rate constant at around (thinned) iteration number 200. In panel (b), we see that the dotted chain flips the other way, as it is required to do.

We now constrain the parameters by enforcing the known ordering on the rates: $k_{ai} > k_{ei} > 0$. To avoid the flip-flop problem, we use the parameterization

$$\theta_{1i} = \log k_{ei} = \beta_1 + b_{1i} \tag{9.37}$$

$$\theta_{2i} = \log(k_{ai} - k_{ei}) = \beta_2 + b_{2i} \tag{9.38}$$

$$\theta_{3i} = \log Cl_i = \beta_3 + b_{3i} \tag{9.39}$$

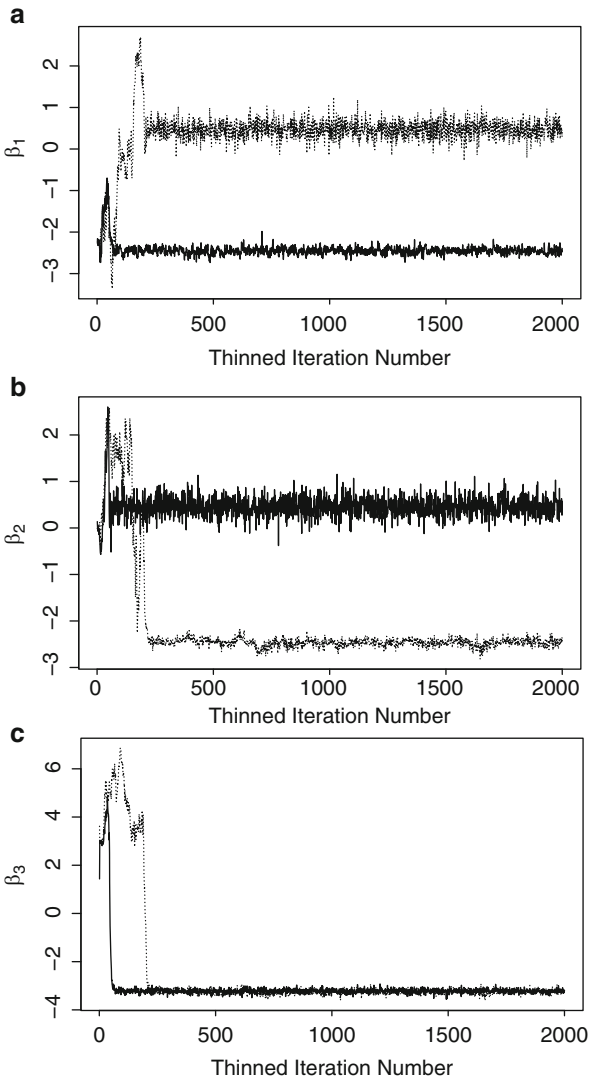

Fig. 9.11 Demonstration of flip-flop behavior for the theophylline data and the unconstrained parameterization given by (9.34)–(9.36): (**a**) β_1, (**b**) β_2, (**c**) β_3. Thinned realizations from two chains appear in each plot

with $\boldsymbol{b}_i = [b_{1i}, b_{2i}, b_{3i}]^{\mathrm{T}} \sim \mathrm{N}_3(\boldsymbol{0}, \boldsymbol{D})$. This is a different model to the model that does not prevent flip-flop since the prior inputs are different. In this case, we keep the same priors which correspond to assuming that the coefficient of variation for $k_a - k_e$ is around 26% which is clearly less meaningful, but in this example, k_a is considerably larger than k_e.

We can convert to the original parameters via

$$\begin{aligned} k_{ei} &= \exp(\theta_{1i}) \\ k_{ai} &= \exp(\theta_{1i}) + \exp(\theta_{2i}) \\ Cl_i &= \exp(\theta_{3i}). \end{aligned}$$

Inference for the population distribution of k_{ei} and Cl_i is straightforward, but for k_{ai}, more work is required. However, the expectation of the population absorption rate is

$$\begin{aligned} \mathrm{E}[k_{ai}] &= \mathrm{E}[\,\exp(\theta_{1i}) + \exp(\theta_{2i})\,] \\ &= \exp\left(\beta_1 + \sqrt{D_{11}}/2\right) + \exp(\beta_1 + \sqrt{D_{11}}/2\Big). \end{aligned}$$

A full Bayesian analysis is postponed until later in the chapter (at the end of Sect. 9.20).

9.19 Generalized Estimating Equations

If interest lies in population parameters, then we may use the estimator $\widehat{\gamma}$ that satisfies

$$\boldsymbol{G}(\boldsymbol{\gamma}, \widehat{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{f}_i) = \boldsymbol{0}, \tag{9.40}$$

where $\boldsymbol{D}_i = \partial \boldsymbol{f}_i / \partial \boldsymbol{\gamma}$, $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\gamma}, \widehat{\boldsymbol{\alpha}})$ is the working covariance model, $\boldsymbol{f}_i = \boldsymbol{f}_i(\boldsymbol{\gamma})$, and $\widehat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. Sandwich estimation may be used to obtain an empirical estimate of the variance $\boldsymbol{V}_{\gamma}$:

$$\left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1} \left[\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{cov}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right] \left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1}. \tag{9.41}$$

We then have the usual asymptotic result: $\boldsymbol{V}_{\gamma}^{-1/2}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \to_d \mathrm{N}(\boldsymbol{0}, \mathbf{I})$.

GEE has not been extensively used in a nonlinear (non-GLM) setting. This is partly because in many settings (e.g., pharmacokinetics/pharmacodynamics), interest focuses on understanding between-individual variability, and explaining this in terms of individual-specific covariates, or making predictions for particular individuals. The interpretation of the parameters within a GEE implementation is also not straightforward. For a marginal GLM, there is a link function and a *linear predictor* which allows interpretation in terms of differences in averages between

Table 9.14 GEE estimates of marginal parameters for the theophylline data

PK label	Parameter	Est.	(s.e.)
$\log\ k_e$	γ_1	-2.52	(0.068)
$\log\ k_a$	γ_2	0.40	(0.17)
$\log\ Cl$	γ_3	-3.25	(0.076)

populations defined by covariates; see Sect. 9.11. Consider a nonlinear model over time. In a mixed model, the population mean parameters are averages of individual-level parameters. A marginal approach models the average response as a nonlinear function of time, and the parameters do not, in general, have interpretations as averages of parameters. Rather, parameters within a marginal nonlinear model determine a population-averaged curve. The parameters can be made a function of covariates such as age and gender, but the interpretation is less clear when compared to a mixed model formulation. For example, in (9.29), we model the individual-level log clearance as a function of a covariate x_i. We could include covariates in the marginal model in an analogous fashion, but it is not individual clearance we are modeling, and the subsequent analysis cannot be used in the same way to derive optimal doses as a function of x, for example. Obviously, GEE cannot provide estimates of between-individual variability or obtain predictions for individuals.

Example: Pharmacokinetics of Theophylline

GEE was implemented with mean model

$$\mathrm{E}[Y_{ij}] = f_i(\boldsymbol{\gamma}) = \frac{D_i \exp(\gamma_1 + \gamma_2)}{\exp(\gamma_3)[\exp(\gamma_2) - \exp(\gamma_1)]} \left[\exp(-\mathrm{e}^{\gamma_1} t_{ij}) - \exp(-\mathrm{e}^{\gamma_2} t_{ij})\right]. \tag{9.42}$$

As just discussed, the interpretation of the parameters for this model is not straightforward since we are simply modeling a population-averaged curve. So, for example, $k_e = \exp(\gamma_1)$ is the rate of elimination that defines the population-averaged curve and is *not* the average elimination rate in the population.

We use working independence ($\boldsymbol{W}_i = \mathbf{I}_{n_i}$) so that (9.40) is equivalent to a nonlinear least squares criteria, which allows the estimates to be found using standard software. The variance estimate (9.41) simplifies under working independence, and the most tedious part is evaluating the $n_i \times 3$ matrix of partial derivatives $\boldsymbol{D}_i = \partial \boldsymbol{f}_i / \partial \boldsymbol{\gamma}$. The estimates and standard errors are given in Table 9.14. It is not possible to directly compare these estimates with those obtained from a mixed model formulation.

9.20 Assessment of Assumptions for General Regression Models

Model checking proceeds as with the linear model with dependent data (Sect. 8.8) except that interpretation is not as straightforward since the properties of residuals are difficult to determine even when the model is correct. We focus on generalized and nonlinear mixed models. For both of these classes Pearson (stage one) residuals,

$$e_{ij} = \frac{Y_{ij} - \mathrm{E}[Y_{ij} \mid \boldsymbol{b}_i]}{\sqrt{\mathrm{var}(Y_{ij} \mid \boldsymbol{b}_i)}}$$

are straightforward to calculate.

With respect to mixed models, as with the LMM, there are assumptions at each of the stages, and one should endeavor to provide checks at each stage. If we are in the situation in which there are individuals with sufficient data to reliably estimate the parameters from these data alone, we should use the resultant estimates to provide checks. Residuals from individual fits can be used to assess whether the nonlinear model is appropriate and if the assumed variance model is appropriate. One may also construct normal QQ plots and bivariate plots of the estimated individual-level parameters to see if the second-stage normality assumption appears reasonable. In a nonlinear setting, there are few results availability on consistency of estimates, unless the model is correct, and so it is far more important to have random effects distributions that are approximately correctly specified.

If individual-level covariates are available, then the estimated parameters may be plotted against these to determine whether a second-stage regression model is appropriate (if we are in exploratory mode). In the pharmacokinetic context, one may model clearance as a function of weight, for example, via a loglinear model as in (9.29). Examining whether the spread of the random effects estimates changes with covariates is also an important step.

All of the above checks can be carried out based on the (shrunken) estimates obtained from random effects modeling, but caution is required as these estimates may be strongly influenced by the assumption of normality. If n_i is large, then this will be less problematic.

Example: Pharmacokinetics of Theophylline

We present some diagnostics for the theophylline data. We first carry out individual fitting using nonlinear least squares (which is possible here since $n_i = 11$), and Fig. 9.12 gives normal QQ plots of the $\log k_e$, $\log k_a$, and $\log Cl$ parameters. There is at least one outlying individual here, but there is nothing too worrying in these plots.

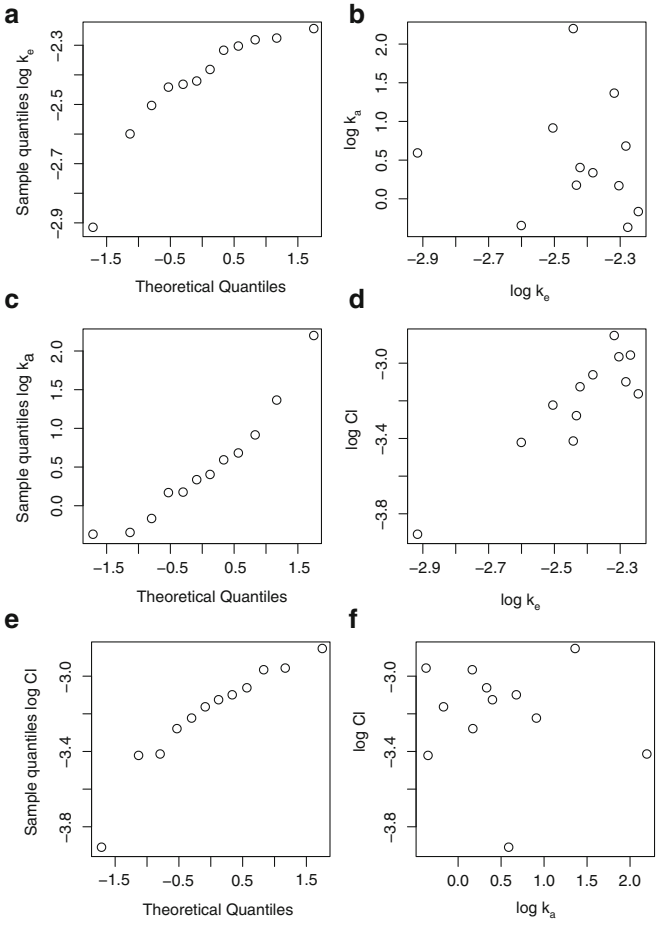

Fig. 9.12 Normal QQ plots (*left column*) and scatterplots (*right column*) of the parameter estimates from individual nonlinear least square fits for the theophylline data. (**a**) QQ plot for $\log k_e$, (**b**) $\log k_a$ versus $\log k_e$, (**c**) QQ plot for $\log k_a$, (**d**) $\log Cl$ versus $\log k_e$, (**e**) QQ plot for $\log Cl$, (**f**) $\log Cl$ versus $\log k_a$

In the following, a number of mixed models are fitted in an exploratory fashion in order to demonstrate some of the flexibility of NLMMs. We first fit a mixed model using MLE and the nonlinear form (9.33). The error terms were assumed to be normal on the concentration scale, with constant variance. Plots of the Pearson residuals versus fitted value and versus time are displayed in Figs. 9.13(a) and (b).

Figure 9.13(b) suggests the variance changes with time (or that the model is inadequate for time points close to 0), and we carry out another analysis with the model

$$\log y_{ij} = \log(\mu_{ij}) + \delta_{ij} \tag{9.43}$$

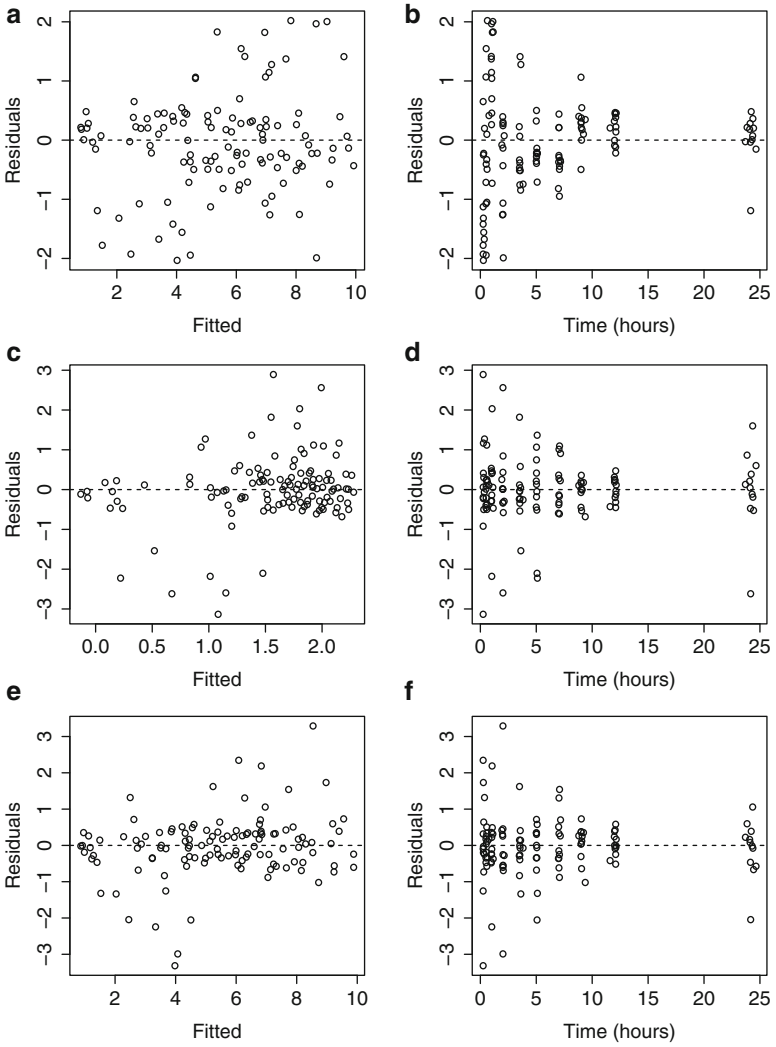

Fig. 9.13 Residuals obtained from various NLMM fits to the theophylline data: (**a**) normal model: residuals against fitted values (**b**) normal model: residuals against time, (**c**) lognormal model: residuals against fitted values (**d**) lognormal model: residuals against time, (**e**) power model: residuals against fitted values (**f**) power model: residuals against time

with μ_{ij} again given by (9.33) and $\delta_{ij} \mid \sigma_\delta^2 \sim_{iid} \mathrm{N}(0, \sigma_\delta^2)$. This lognormal model has (approximately) a constant coefficient of variation. To fit this model, the responses at time 0 were removed since $\mu_{ij} = 0$ for $t_{ij} = 0$. This time, we adopt the parameterization that prevents flip-flop, that is, the model with (9.37)–(9.39). This model produced the Bayesian summaries given in Table 9.13 which are

reasonably consistent with those in the normal model. Unfortunately, the residual plot in Fig. 9.13(d) shows only a slight improvement over the normal model in panel (b).

The next model considered was $y_{ij} = \mu_{ij} + \epsilon_{ij}$ with the *power* model

$$\epsilon_{ij} \mid \mu_{ij}, \sigma_0^2, \sigma_1^2, \gamma \sim_{ind} \mathrm{N}\left(0, \sigma_0^2 + \sigma_1^2 \mu_{ij}^{\gamma}\right) \tag{9.44}$$

with μ_{ij} given by (9.33) and $0 < \gamma \leq 2$. This model has two components of variance and may be used when an assay method displays constant measurement at low concentrations with the variance increasing with the mean for larger concentrations. See Davidian and Giltinan (1995, Sect. 2.2.3) for further discussion of variance models.

The joint prior on $[\sigma_0, \sigma_1, \gamma]$ can be difficult to specify since there is dependence between σ_1 and γ in particular. For simplicity, uniform priors on the range [0,2] were placed on σ_0 and σ_1. The parameter γ controls the strength of the mean–variance relationship, and, considering the second component only, the constant coefficient of variation model corresponds to $\gamma = 2$. In the pharmacokinetics literature, fixing $\gamma = 1$ or 2 is not uncommon. A uniform prior on [0,2] was specified for γ also. Figures 9.13(e) and (f) show the residual plots for this model, and we see some improvement over the other two error models, though there is still some misspecification evident at low time points in panel (f). Further analyses for these data might examine other absorption models (since the kinetics may be nonlinear, which could explain the poor fit at low times).

Posterior summaries for the power variance model are given in Fig. 9.14. The strong dependence between σ_1 and γ is evident in panel (f). There is a reasonable amount of uncertainty in the posterior for γ, but the median is 0.71. The parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{D}$ are given in Table 9.13 and are similar to those from the normal and lognormal error models. Following the procedure described in Sect. 9.18.2, samples for the population medians for k_e, k_a, and Cl were generated, and these are displayed in Fig. 9.15, with notable skewness in the posteriors for k_a and Cl.

9.21 Concluding Remarks

The modeling of generalized linear and nonlinear dependent data is inherently more difficult than the modeling of linear dependent data due to mathematical tractability, the required computations to perform inference and parameter interpretation. Conceptually, however, the adaption of mixed (conditional) and GEE (marginal) models to the generalized linear and nonlinear scenarios is straightforward. With respect to parameter interpretation, the clear distinction between marginal and conditional models is critical and needs to be recognized.

There is little theory on the consistency of estimators in the face of model misspecification for GLMMs and NLMMs. This suggests that one should be more

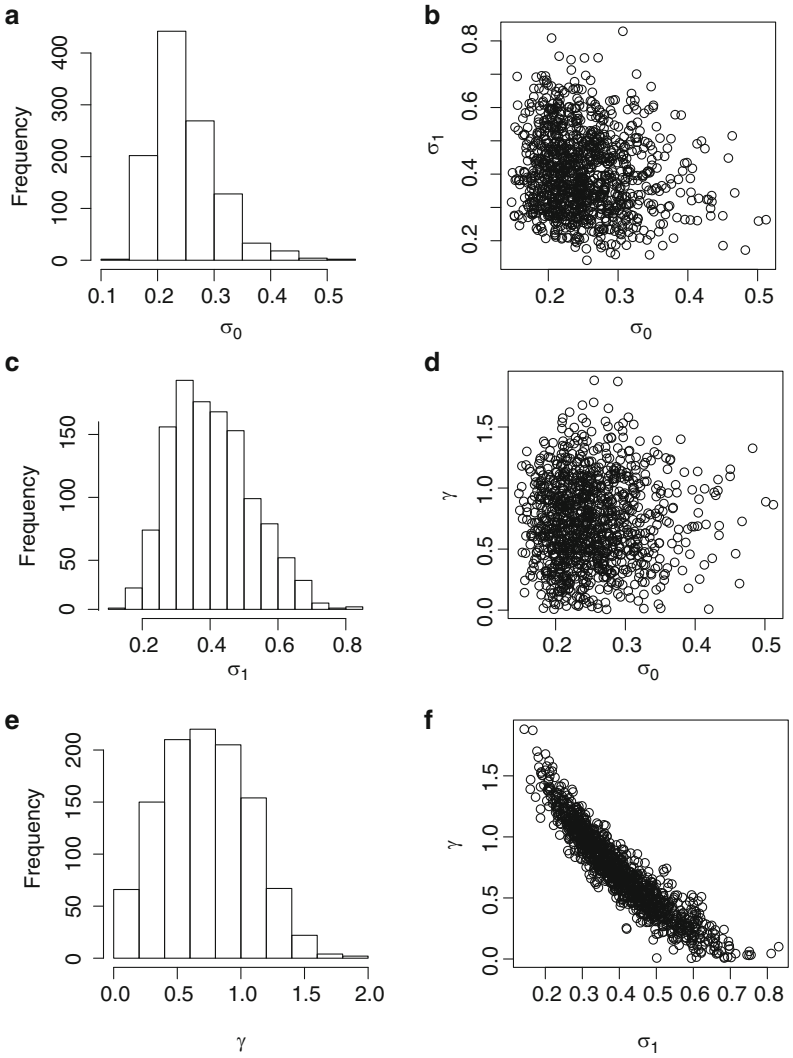

Fig. 9.14 Posterior summaries for the two-component power error model (9.44) fitted to the theophylline data. Posterior marginals for σ_0, σ_1, γ in the left common and bivariate plots in the right column

cautious in interpretation of the results from GLMMs and NLMMs, when compared to LMMs, and model checking should be carefully carried out. The effects of model misspecification with mixed models have attracted a lot of interest. Heagerty and Kurland (2001) illustrate the bias that is introduced when the random effects variances are a function of covariates. McCulloch and Neuhaus (2011) show that misspecification of the assumed random effects distribution has less impact on prediction of random effects. Sensitivity analyses, with respect to the random effects

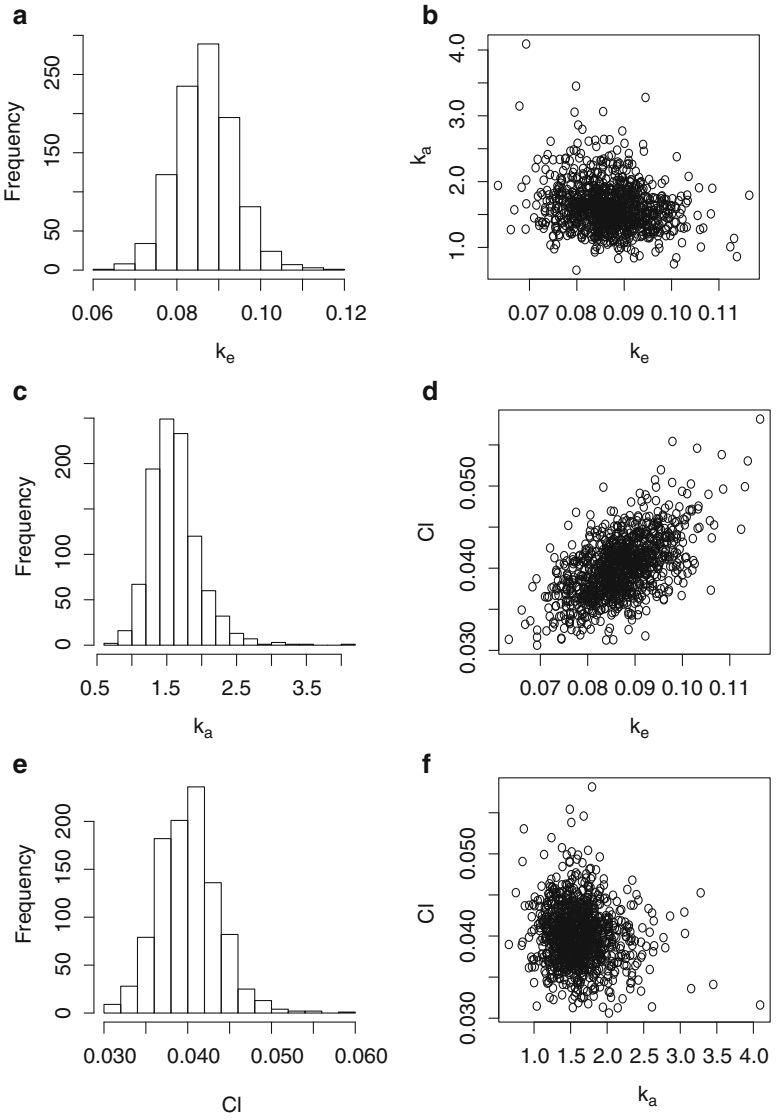

Fig. 9.15 Posterior distributions from the power model (9.44) fitted to the theophylline data: **(a)** population median k_e, **(b)** population median k_a versus population median k_e, **(c)** population median k_a, **(d)** population median k_e versus population median Cl, **(e)** population median Cl, **(f)** population median Cl versus population median k_a

distribution, for example, can be useful. The Bayesian approach, with computation via MCMC, is ideally suited to this endeavor. If the number of observations per unit, or the number of units, is small, then the MCMC route is appealing because

one does not have to rely on asymptotic inference. Model checking is difficult in this situation, however.

We have not discussed REML in the context of GLMs; Smyth and Verbyla (1996) show how REML may be derived from a conditional likelihood approach in the context of GLMs with dispersion parameters and canonical link functions.

The modeling of dependent binary data is a difficult enterprise since binary observations contain little information, and there is no obvious choice of multivariate binary distribution. Logistic mixed models are intuitively appealing but are restrictive in the dependence structure they impose on the data. Care in computation is required, and the use of adaptive Gauss–Hermite for MLE, or MCMC for Bayes, is recommended. As always, GEE has desirable robustness properties for large numbers of clusters. In the GLM context, we emphasize the fitting of both types of model in a complimentary fashion. We have illustrated how marginal inference may be carried out with the logistic mixed model, which allows direct comparison of results with GEE.

9.22 Bibliographic Notes

Liang and Zeger (1986) and Zeger and Liang (1986) popularized GEE by considering GLMs with dependence within units (in the context of longitudinal data). Prentice (1988) proposed using a second set of estimating equations for $\boldsymbol{\alpha}$. Gourieroux et al. (1984) considered the quadratic exponential model. Zhao and Prentice (1990) discussed the use of this model for multivariate binary data and Prentice and Zhao (1991) for general responses (to give the approach labelled GEE2). Qaqish and Ivanova (2006) describe an algorithm for detecting when an arbitrary set of logistic contrasts correspond to a valid set of joint probabilities and for computing them if they provide a legal set. Fitzmaurice et al. (2004) is a very readable account of the modeling of longitudinal data with GLMs, from a frequentist (GEE and mixed model) perspective.

An extensive treatment of Bayesian multilevel modeling is described in Gelman and Hill (2007). We have concentrated on inverse gamma priors for random effects variances, but a popular alternative is the half-normal prior; see Gelman (2006) for further details. Fong et al. (2010) describe how the INLA computational approach may be used for GLMMs, including a description of its shortcomings, in terms of accuracy, for the analysis of binary data. Models and methods of analysis for spatial data are reviewed in Gelfand et al. (2010).

Davidian and Giltinan (1995) is an extensive and excellent treatment of nonlinear modeling with dependent responses, mostly from a non-Bayesian perspective. Pinheiro and Bates (2000) is also excellent and covers mixed models (again primarily from a likelihood perspective) and is particularly good on computation.

9.23 Exercises

9.1 Consider the model

$$\mathrm{E}[Y \mid b] = \frac{\exp(\boldsymbol{\beta x} + b)}{1 + \exp(\boldsymbol{\beta x} + b)},$$

with $b \mid \sigma_0^2 \sim_{iid} \mathrm{N}(0, \sigma_0^2)$. Prove that

$$\mathrm{E}[Y] \approx \frac{\exp\left[\boldsymbol{\beta x}/(c^2\sigma_0^2 + 1)^{1/2}\right]}{1 + \exp\left[\boldsymbol{\beta x}/(c^2\sigma_0^2 + 1)^{1/2}\right]}$$

where $c = 16\sqrt{3}/(15\pi)$.
[Hint: $G(z) \approx \Phi(cz)$ where $G(z) = (1 + \mathrm{e}^{-z})^{-1}$ is the CDF of a logistic random variable, and $\Phi(\cdot)$ is the CDF of a normal random variable.]

9.2 Show that if each response is on the whole real line, then the density (9.19), with $c_i = 0$, corresponds to the multivariate normal model.

9.3 With respect to Table 9.11, show that, for a model for two binary responses parameterized in terms of the marginal means and marginal odds ratio, the likelihood is given by (9.26).

9.4 Sommer (1982) contains details of a study on 275 children in Indonesia. This study examined, among other things, the association between the risk of respiratory infection and xerophthalmia (dry eye syndrome), which may be caused by vitamin A deficiency. These data are available in the `R` package `epicalc` and are named `Xerop`.

Consider the marginal model for the jth observation on the ith child

$$\log\left(\frac{\mathrm{E}[Y_{ij}]}{1 - \mathrm{E}[Y_{ij}]}\right) = \gamma_0 + \gamma_1\,\mathtt{gender}_{ij} + \gamma_2\,\mathtt{hfora}_{ij} + \gamma_3\,\mathtt{cos}_{ij} + \gamma_4\,\mathtt{sin}_{ij} + \gamma_5\,\mathtt{xero}_{ij} + \gamma_6\,\mathtt{age}_{ij} + \gamma_7\,\mathtt{age}_{ij}^2 \tag{9.45}$$

where:

- Y_{ij} is the absence/presence of respiratory infection.
- $\mathtt{gender}_{ij}$ is the gender (0 = male, 1 = female).
- $\mathtt{hfora}_{ij}$ is the height-for-age.
- $\mathtt{cos}_{ij}$ is the cosine of time of measurement i, j (time is in number of quarters).
- $\mathtt{sin}_{ij}$ is the sine of time of measurement i, j (time is in number of quarters).
- $\mathtt{xero}_{ij}$ is the absence/presence (0/1) of xerophthalmia.
- $\mathtt{age}_{ij}$ is the age.

See Example 9.4 of Diggle et al. (2002) for more details on this model.

(a) Interpret each of the coefficients in (9.45).
(b) Provide parameter estimates and standard errors from a GEE analysis.

(c) Consider a GLMM logistic analysis with a normally distributed random intercept and the conditional version of the regression model (9.45). Interpret the coefficients of this model.
(d) Provide parameter estimates and standard errors from the GLMM analysis.
(e) Summarize the association between respiratory infection and xeropthalmia and age.

9.5 On the book website, you will find data on illiteracy and race collected during the US 1930 census. Wakefield (2009b) provides more information on these data. *Illiterate* is defined as being unable to read and over 10 years of age. For each of the $i = 1, \ldots, 49$ states that existed in 1930, the data consist of the number of illiterate individuals Y_{ij} and the total population aged 10 years and older N_{ij} by race, coded as native-born White ($j = 1$), foreign-born White ($j = 2$), and Black ($j = 3$). Let p_{ij} be the probability of being illiterate for an individual residing in state i and of race j. An additional binary state-level variable $x_i = 0/1$ describes whether Jim Crow laws were absent/present in state $i = 1, \ldots, 49$. These laws enforced racial segregation in all public facilities.

The association between illiteracy and race, state, and Jim Crow laws will be examined using logistic regression models. In particular, interest focuses on whether illiteracy in 1930 varied by race, varied across states, and was associated with the presence/absence of Jim Crow laws:

(a) Calculate the empirical logits of the p_{ij}'s, and provide informative plots that graphically display the association between illiteracy and state, race, and Jim Crow laws.
(b) First consider the native-born White data only (Y_{i1}, N_{i1}), $i = 1, \ldots, 49$, with the following models:

- *Binomial:* $Y_{i1} \mid p_{i1} \sim \text{Binomial}(N_{i1}, p_{i1})$, with the logistic model

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) = \gamma_1 \tag{9.46}$$

  for $i = 1, \ldots, 49$.
- *Quasi-Likelihood:* Model (9.46) with

$$\text{E}[Y_{i1}] = N_{i1}p_{i1}, \qquad \text{var}(Y_{i1}) = \kappa \times N_{i1}p_{i1}(1-p_{i1}).$$

- *GEE:* Model (9.46) with $\text{E}[Y_{i1}] = N_{i1}p_{i1}$ and working independence.
- *GLMM*

$$\log\left(\frac{p_{i1}}{1-p_{i1}}\right) = \beta_1 + b_{i1} \tag{9.47}$$

  with $b_{i1} \mid \sigma_1^2 \sim_{iid} \text{N}(0, \sigma_1^2)$, $i = 1, \ldots, 49$.

(i) Give careful definitions of $\exp(\gamma_1)$ in the GEE model and $\exp(\beta_1)$ in the GLMM.
(ii) Fit the binomial model to the native-born White data and give a 95% confidence interval for the odds of native-born White illiteracy. Is this model appropriate?
(iii) Fit the quasi-likelihood and GEE models to the native-born White data and give 95% confidence interval for the odds of native-born White illiteracy in each case. How does the GEE approach differ from quasi-likelihood here? Which do you prefer?
(iv) Fit the GLMM model to the data using a likelihood approach and give a 95% confidence interval for the odds of native-born White illiteracy along with an estimate of the between-state variability in logits. Are the results consistent with the GEE analysis?

(c) Now consider data on all three races. Using GEE fit, *separate* models to the data of each race. Give a 95% confidence interval for the odds ratios comparing illiteracy between foreign-born Whites and native-born Whites, and comparing Blacks with native-born Whites. Is there any problem with this analysis?
(d) Use GEE to fit a model to all three races simultaneously and compare your answer with the previous part. Which analysis is the most appropriate and why?
(e) Fit the GLMM

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_j + b_{ij} \tag{9.48}$$

with $b_{ij} \mid \sigma_j^2 \sim_{ind} \mathrm{N}(0, \sigma_j^2)$, $j = 1, 2, 3$, using likelihood-based methods. Give 95% confidence intervals for the odds ratios comparing illiteracy between foreign-born Whites and native-born Whites, and comparing Blacks with native-born Whites. Are your conclusions the same as with the GEE analysis? Does this model require refinement?
(f) The state-level Jim Crow law indicator will now be added to the analysis. Consider the model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{0j} + \gamma_{1j} x_i \tag{9.49}$$

Give interpretations of each of $\exp(\gamma_{0j}), \exp(\gamma_{1j})$ for $j = 1, 2, 3$. Fit this model using GEE and interpret and summarize the results in a clear fashion.
(g) Consider Bayesian fitting of the GLMM:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j} x_i + b_{ij} \tag{9.50}$$

where $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \mathrm{N}_3(\,\mathbf{0}, \boldsymbol{D}\,)$ with $\boldsymbol{b}_i = [b_{i1}, b_{i2}, b_{i3}]^{\mathrm{T}}$ and

$$\boldsymbol{D} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_2\sigma_1 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_3\sigma_1 & \rho_{23}\sigma_3\sigma_2 & \sigma_3^2 \end{bmatrix}$$

is a 3×3 variance–covaraince matrix for the random effects $\boldsymbol{b}_i$. Assume improper flat priors for β_{0j}, β_{1j}, $j = 1, 2, 3$, and the Wishart prior $\boldsymbol{W} = \boldsymbol{D}^{-1} \sim \text{Wishart}(r, \boldsymbol{S})$ parameterized so that $\mathrm{E}[\boldsymbol{W}] = r\boldsymbol{S}$, with $r = 3$ and

$$\boldsymbol{S} = \begin{bmatrix} 30.45 & 0 & 0 \\ 0 & 30.45 & 0 \\ 0 & 0 & 30.45 \end{bmatrix}.$$

Carry out a Bayesian analysis using this model and interpret and summarize the results in a clear fashion.

(h) Write a short summary of what you have found, concentrating on the particular substantive questions of interest stated in the introduction.

9.6 For the theophylline data considered in this chapter, reproduce the results in Table 9.14 by coding up the nonlinear GEE model with working independence. These data are available as `Theoph` in the `R` package.

9.7 Throughout this chapter, mixed models with clustering induced by normally distributed random effects have been considered. In this question, a non-normal random effects distribution will be considered. Suppose, for paired binary observations, that the data-generating mechanism is the following:

$$Y_{ij} \mid \mu_{ij} \sim_{ind} \text{Bernoulli}(\mu_{ij}),$$

for $i = 1, \ldots, n, j = 1, 2$, with

$$\mu_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{ij} + b_i)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + b_i)}$$

$$b_i = \begin{cases} -\gamma & \text{with probability } 1/2 \\ \gamma & \text{with probability } 1/2. \end{cases}$$

and $X_{ij} \sim_{iid} \text{Unif}(-10, 10)$. The parameters $\beta_1 \in \mathbb{R}$ and $\gamma > 0$ are unknown, and all b_i are independent and identically distributed. For simplicity, assume $\beta_0 = 0$ throughout:

(a) For $0 \leq \beta_1 \leq 1$ and $0 \leq \gamma \leq 5$, calculate the correlation between the outcomes Y_{ij} and $Y_{ij'}$ within cluster i, averaged over the distribution of clusters.

(b) For $\beta_1 = 1$ and $0 \leq \gamma \leq 5$, calculate the numerical value of the true slope parameter estimated by a GEE logistic regression model of y on x, with working independence within clusters. Compare this value to the true β_1.

(c) Consider a study with paired observations and binary outcomes (e.g., a matched-pairs case-control study as described in Sect. 7.10.3). The true data-generating mechanism is as above with $\beta_1 = 1, \gamma = 5$. First plot y versus x for all observations and add a smoother. This plot seems to indicate that there are low-, medium-, and high-risk subjects, depending on levels of x.

(d) In truth, of course, there are not three levels of risk. For some example data, give a plot that illustrates this and write an explanation of what your plot shows. The plot should use only observed, and not latent, variables.