38. Random Effects

This chapter includes well-known as well as state-of-the-art statistical modeling techniques for drawing inference on correlated data, which occur in a wide variety of studies (during quality control studies of similar products made on different assembly lines, community-based studies on cancer prevention, and familial research of linkage analysis, to name a few).

The first section briefly introduces statistical models that incorporate random effect terms, which are increasingly being applied to the analysis of correlated data. An effect is classified as a random effect when inferences are to be made on an entire population, and the levels of that effect represent only a sample from that population.

The second section introduces the linear mixed model for clustered data, which explicitly models complex covariance structure among observations by adding random terms into the linear predictor part of the linear regression model. The third section discusses its extension – generalized linear mixed models (GLMMs) – for correlated nonnormal data.

The fourth section reviews several common estimating techniques for GLMMs, including the EM and penalized quasi-likelihood approaches, Markov chain Newton-Raphson, the stochastic approximation, and the S-U algorithm. The fifth section focuses on some special topics related to hypothesis tests of random effects, including score tests for various models. The last section is a general discussion of the content of the chapter and some other topics relevant to random effects models.

38.1 Overview

Classical linear regression models are a powerful tool for exploring the dependence of a response (such as blood pressure) on explanatory factors (such as weight, height and nutrient intake). However, the normality assumption required for these response variables has severely limited its applicability. To accommodate a wide variety of independent nonnormal data, *Nelder* and *Wedderburn* [38.1] and *McCullagh* and *Nelder* [38.2] introduced generalized linear models (GLMs), a natural generalization of linear regression models. The GLMs allow responses to have nonGaussian distributions. Hence, data on counts and proportions can be conveniently fitted into this framework. In a GLM, the mean of a response is typically linked to linear predictors via a nonrandom function, termed the *link function*. For analytical convenience, the link function is often determined by

the response's distribution. As an example, for Poisson data, the link is routinely chosen as log, whereas for Bernoulli responses, the link is usually chosen to be logit.

In many applications, however, responses are correlated due to unobservable factors, such as circumstantial or genetic factors. Consider the problem of investigating the strength of the beams made by randomly selected manufacturers. Beams made at the same factory are likely to be correlated because the were made using the same manufacturing procedures. Other examples include a longitudinal study of blood pressure, where repeated observations taken from the same individuals are likely be correlated, and a familial study in cardiovascular disease, where the incidents of heart failure from family members are likely to be related. In the last two decades, random effects models have emerged as a major tool for analyzing these kinds of correlated data (see [38.3–7] among others).

Indeed, using random effects in the modeling of correlated data provides several benefits. First, it provides a framework for performing data modeling in unbalanced designs, especially when measurements are made at arbitrary irregularly spaced intervals over many observational studies (as opposed to ANOVA, which requires a balanced dataset). Secondly, random effects can be used to model subject-specific effects, and they offer a neat way to separately model within- and between-subject variations. Thirdly, the framework of random effects provides a systematic way to estimate or predict individual effects.

Though conceptually attractive, GLMMs are often difficult to fit because of the intractability of the underlying likelihood functions. Only under special circumstances, such as when both response and random effects are normally or conjugately distributed, will the associated likelihood function have a closed form. Cumbersome numerical integrations often have to be performed. To alleviate this computational burden, various modeling techniques have been proposed. For example, *Stiratelli* et al. [38.4] proposed an EM algorithm for fitting serial binary data; *Schall* [38.5] developed an iterative Newton-Raphson algorithm; *Zeger* and *Karim* [38.6] and *McCulloch* [38.7] considered Monte Carlo EM methods. All of these commonly used inferential procedures will be presented and discussed in this chapter.

The rest of this chapter is structured as follows. Section 38.2 introduces the linear mixed model for clustered data and Sect. 38.3 discusses its extension, generalized linear mixed models, for correlated nonnormal data. Section 38.4 reviews several common estimating techniques for GLMMs, including the EM approach, penalized quasi-likelihood, Markov chain Newton-Raphson, the stochastic approximation and the S–U algorithm. Section 38.5 focuses on some special topics related to hypothesis tests of random effects. Section 38.6 concludes this chapter with discussion and some other topics relevant to random effects models.

Throughout this chapter, $f(\cdot)$ and $F(\cdot)$ denote the probability density (or probability mass) function (with respect to some dominating measure, such as the Lebesgue measure) and the cumulative distribution function, respectively. If the context is clear, we do not use separate notations for random variables and their actual values.

38.2 Linear Mixed Models

A clustered data structure is typically characterized by a series of observations on each of a collection of observational clusters. Consider the problem of investigating whether the beam produced from iron is more resilient than that from an alloy. To do this, we measure the strength of the beams made of iron and alloy from randomly selected manufacturers. Each manufacturer may contribute multiple beams, in which case each manufacturer is deemed as a cluster, while each beam contributes to a unit of observation. Other examples include the measurements of products produced by a series of assembly lines, and blood pressure taken weekly on a group of patients, in which cases the clusters are assembly lines and patients respectively. Clustering typically induces dependence among observations. A linear mixed model [38.3] explicitly models the complex covariance structure among observations by adding random terms into the linear predictor part of a linear regression model. Thus, both random and fixed effects will be present in an LMM. In data analysis, the decision on whether a factor should be fixed or random is often made on the basis of which effects vary with clusters. That is, clusters are deemed to be a random sample of a larger population, and therefore any effects that are not constant for all clusters are regarded as random.

As an example, let's say that $\boldsymbol{Y}_i$ denotes the response vector for the ith of a total of m clusters, where n_i measurements of blood pressure were taken for the

ith patient. $\mathbf{X}_i$ the known covariate matrix ($n_i \times p$) associated with the observations, such as the patient's treatment assignment and the time when the observation was taken, $\boldsymbol{b}_i$ is the vector of random effects and $\mathbf{Z}_i$ is the known design matrix associated with the random effects. Usually, the columns of $\mathbf{Z}_i$ are vectors of ones and a subset of those of $\mathbf{X}_i$ for modeling random intercepts and slopes. A linear mixed model can thus be specified as

$$\boldsymbol{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \,, \tag{38.1}$$

where we typically assume that the random error vector $\boldsymbol{\epsilon}_i \sim MVN(0, \sigma^2\mathbf{I}_{n_i})$ and $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{b}_i$, which is assumed to have an expectation of zero for model identifiability. Here, $\mathbf{I}_{n_i}$ is an identity matrix of order n_i. In practice, we often assume $\boldsymbol{b}_i \sim MVN(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where its variance–covariance matrix is dependent on a fixed q-dimensional (a finite number) parameter, say, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, termed "variance components". These variance components convey information about the population that the clusters are randomly selected from and are often of interest to practitioners, aside from the fixed effects.

To encompass all data, we denote the concatenated collections of $\boldsymbol{Y}_i$'s, $\mathbf{X}_i$'s, $\boldsymbol{b}_i$'s and $\boldsymbol{\epsilon}_i$'s by $\boldsymbol{Y}$, $\mathbf{X}$, $\boldsymbol{b}$, $\boldsymbol{\epsilon}$. For example, $\boldsymbol{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_m')'$. We now denote a block diagonal matrix whose ith diagonal block is $\mathbf{Z}_i$ by $\mathbf{Z}$. In this case (38.1) can be rewritten compactly as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} + \boldsymbol{\epsilon} \,, \tag{38.2}$$

where $\boldsymbol{b} \sim MVN(0, \mathbf{D})$, $\boldsymbol{\epsilon} \sim MVN(0, \sigma^2\mathbf{I}_{\mathcal{N}})$ and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent. Here, $\mathbf{D}$ is a block diagonal matrix whose diagonal blocks are $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, and $\mathbf{I}_{\mathcal{N}}$ is an identity matrix of order $\mathcal{N}$, where $\mathcal{N}$ is the total number of observations (so $\mathcal{N} = \sum_{i=1}^{m} n_i$).

Indeed, model (38.2) accommodates a much more general data structure beyond clustered data. For example, with properly defined $\mathbf{Z}$ and random effects $\boldsymbol{b}$, model (38.2) encompasses crossed factor data [38.8] and Gaussian spatial data [38.9].

38.2.1 Estimation

Fitting model (38.1) or its generalized version (38.2) is customarily likelihood-based. A typical maximum likelihood estimation procedure is as follows.

First observe that $\boldsymbol{Y}$ is normally distributed, $\boldsymbol{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \sigma^2\mathbf{I}_{\mathcal{N}}$, so that the log-likelihood for the observed data is

$$\ell = -\frac{1}{2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\log|\mathbf{V}| - \frac{\mathcal{N}}{2}\log 2\pi \,. \tag{38.3}$$

Denote the collection of unknown parameters in the model by $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{\theta}', \sigma^2)'$. Setting $\partial\ell/\partial\boldsymbol{\gamma} = 0$ gives the maximum likelihood equation. Specifically, a direct calculation of $\partial\ell/\partial\boldsymbol{\beta}$ yields the ML equation for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{Y} \,. \tag{38.4}$$

Denote by θ_k the kth element of the variance components $(\boldsymbol{\theta}, \sigma^2)$, where we label $\theta_{q+1} = \sigma^2$. Equating $\partial\ell/\partial\theta_k = 0$ gives

$$-\frac{1}{2}\left[\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_k}\right) - (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})' \times \mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_k}\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})\right] = 0 \,, \tag{38.5}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. In practice, iterations are required between (38.4) and (38.5) to obtain the MLEs. Furthermore the asymptotic sampling variance is routinely obtained from the inverse of the information matrix, which is minus the expected value of the matrix of second derivatives of the log-likelihood (38.3).

It is, however, worth pointing out that the MLEs obtained from (38.4, 5) are biased, especially for the variance components when the sample size is small. This is because the estimating equation (38.5) for the variance components fails to account for the loss of degrees of freedom when the true $\boldsymbol{\beta}$ is replaced by its estimate, $\hat{\boldsymbol{\beta}}$. To address this issue, an alternative maximum likelihood procedure, called the restricted maximum likelihood procedure, has been proposed for estimating the variance components [38.10]. The key idea is to replace the original response $\boldsymbol{Y}$ by a linear transform, so that the resulting 'response' contains no information about $\boldsymbol{\beta}$. The variance components can then be estimated based on this transformed response variable.

More specifically, choose a vector $\boldsymbol{a}$ such that $\boldsymbol{a}'\mathbf{X} = 0$. For more efficiency we use the maximum number, $\mathcal{N} - p$, of linearly independent vectors $\boldsymbol{a}$ and write $\mathbf{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{\mathcal{N}-p})$, which has a full row rank of $\mathcal{N} - p$. The restricted MLE will essentially apply the MLE procedure on $\mathbf{A}'\boldsymbol{Y}$, in lieu of the original $\boldsymbol{Y}$.

To proceed, we note that $\mathbf{A}'\boldsymbol{Y} \sim MVN(0, \mathbf{A}'\mathbf{V}\mathbf{A})$. The ML equations for the variance components can now be derived in a similar way to those for the original

$\boldsymbol{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, namely by replacing $\boldsymbol{Y}$, $\mathbf{X}$ and $\mathbf{V}$ with $\mathbf{A}'\boldsymbol{Y}$, 0 and $\mathbf{AVA}'$ respectively in (38.5).

Caution must be exercised if the MLEs or the RMLEs of the variance components fall outside of the parameter space, as in the case of a negative estimate for a variance, in which case those solutions must be adjusted to yield estimates in the parameter space; see a more detailed discussion in *McCulloch* and *Searle* [38.11].

38.2.2 Prediction of Random Effects

A fixed effect differs from a random effect in that the former is considered to be constant and is often the main parameter we wish to estimate. In contrast, a random effect is considered to be an effect deriving from a population of effects. Consider again the aforementioned study of beam strength. Aside from the differences between the beams made from iron and alloy, there should be at least two sources of variability: (1) among beams produced by the same manufacturer; (2) between manufacturers. A simple random effects model can be specified as

$$E(Y_{ij}|b_i) = X_{ij}\beta + b_i \,,$$

where Y_{ij} is the strength of the j-th beam produced by the ith manufacturer and X_{ij} indicates whether iron or alloy was used to produce such a beam. Note that b_i is the effect on the strength of the beams produced by the i-th manufacturer, and this manufacturer was just the one among the selected manufacturers that happened to be labeled i in the study. The manufacturers were randomly selected as representative of the population of all manufacturers in the nation, and inferences about random effects were to be made about that population. Hence, estimating the variance components is of substantial interest for this purpose. On the other hand, one may wish to gain information about the performance of particular manufacturers. For instance, one may want to rank various manufacturers in order to select the best (or worst) ones. In these cases we are interested in predicting b_i.

In general the 'best' prediction of $\boldsymbol{b}$ in (38.2) based on observed response $\boldsymbol{Y}$ is required to minimize the mean squared error

$$\int (\hat{\boldsymbol{b}} - \boldsymbol{b})'\mathbf{G}(\hat{\boldsymbol{b}} - \boldsymbol{b}) f(\boldsymbol{Y}, \boldsymbol{b}) \mathrm{d}\boldsymbol{Y} \mathrm{d}\boldsymbol{b} \,, \tag{38.6}$$

where the predictor $\hat{\boldsymbol{b}}$ depends only on $\boldsymbol{Y}$, $f(\boldsymbol{Y}, \boldsymbol{b})$ is the joint density function of $\boldsymbol{Y}$ and $\boldsymbol{b}$, and $\mathbf{G}$ is a given nonrandom positive definite matrix. It can be shown for any given $\mathbf{G}$ that the minimizer is $E(\boldsymbol{b}|\boldsymbol{Y})$; in other words the conditional expectation of $\boldsymbol{b}$ given the observed response $\boldsymbol{Y}$.

If the variance components are known, an analytical solution exists based on the linear mixed model (38.2). That is, assuming $\boldsymbol{Y}$ and $\boldsymbol{b}$ follow a joint multinormal distribution, it follows that

$$E(\boldsymbol{b}|\boldsymbol{Y}) = \mathbf{DZ}'\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) \,.$$

Replacing $\boldsymbol{\beta}$ by its MLE

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{Y}$$

would yield the Best linear unbiased predictor (BLUP) of random effects [38.12]. Because $\mathbf{D}$ and $\mathbf{V}$ are usually unknown, they are often replaced by their MLEs or RMLEs when calculating the BLUP, namely

$$\hat{\boldsymbol{b}} = \hat{\mathbf{D}}\hat{\mathbf{Z}}'\hat{\mathbf{V}}^{-1}(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \,.$$

Extensive derivation for the variance of the BLUP when the variance components are known has been given by *Henderson* et al. [38.12]. The variance of the BLUP with unknown variance components is not yet fully available.

38.3 Generalized Linear Mixed Models

Nonnormal data frequently arise from engineering studies. Consider again the beam study, where we now change the response to be a binary variable, indicating whether a beam has satisfied the criteria of quality control. For such nonnormal data, statistical models can be traced back to as early as 1934, when *Bliss* [38.13] proposed the first probit regression model for binary data. However, it took another four decades before *Nelder* and *Wedderburn* [38.1] and *McCullagh* and *Nelder* [38.2] proposed generalized linear models (GLMs) that could unify models and modeling techniques for analyzing more general data (such as counted data and polytomous data). Several authors [38.3–5] have considered a natural generalization of the GLMs to accommodate correlated nonnormal data. Their approach was to add random terms to the linear predictor parts, and the resulting models are termed generalized linear mixed models (GLMMs).

As an example, let $\boldsymbol{Y_{ij}}$ denote the status (such as a pass or a fail from the quality assurance test) of the

jth beam from the i-th manufacturer. We might create a model such as

$$\begin{aligned} Y_{ij}|b_i &\overset{\text{iid}}{\sim} \text{Bernoulli}\left(\mu_{ij}^{\text{b}}\right) ; \\ & i = 1, \ldots, m,\ j = 1, \ldots, n_i , \\ \text{logit}\left(\mu_{ij}^{\text{b}}\right) &= \boldsymbol{X}_{ij}'\boldsymbol{\beta} + b_i , \\ b_i &\overset{\text{iid}}{\sim} N\left(0, \sigma_u^2\right) , \end{aligned}$$

where $\text{logit}(\mu) = \log[\mu/(1-\mu)]$ is the link function that bridges the conditional probability and the linear predictors. The normal assumption for the random effects b_i is reasonable because the logit link carries the range of parameter space of μ_{ij} from $[0, 1]$ into the whole real line. Finally, we use independent b_i's to model the independent cluster effects and the within-cluster correlations among observations.

It is straightforward to generalize the above formulation to accommodate more general data. Specifically, let $\boldsymbol{X}_{ij}$ be a $p\times 1$ covariate vector associated with response Y_{ij}. Conditional on an unobserved cluster-specific random variable $\boldsymbol{b}_i$ (an $r\times 1$ vector), Y_{ij} are independent and follow a distribution of exponentials, that is

$$Y_{ij}|\boldsymbol{b}_i \overset{\text{iid}}{\sim} f(Y_{ij}|\boldsymbol{b}_i) , \tag{38.7}$$

$$f(Y_{ij}|\boldsymbol{b}_i) = \exp\left\{[Y_{ij}\alpha_{ij} - h(\alpha_{ij})]/\tau^2 - c(Y_{ij}, \tau)\right\} . \tag{38.8}$$

The conditional mean of $Y_{ij}|\boldsymbol{b}_i$, μ_{ij}^{b}, is related to α_{ij} through the identity $\mu_{ij}^{\text{b}} = \partial h(\alpha_{ij})/\partial\alpha_{ij}$, the transformation of which is to be modeled as a linear model in both the fixed and random effects:

$$g\left(\mu_{ij}^{\text{b}}\right) = \boldsymbol{X}_{ij}'\boldsymbol{\beta} + \boldsymbol{Z}_{ij}'\boldsymbol{b}_i , \tag{38.9}$$

where $g(\cdot)$ is termed a *link function*, often chosen to be invertible and continuous, and $\boldsymbol{Z}_{ij}$ is an $r\times 1$ design vector associated with the random effect. The random effects $\boldsymbol{b}_i$ are mutually independent with a common underlying distribution $F(\cdot;\boldsymbol{\theta})$ [or density $f(\cdot;\boldsymbol{\theta})$], where the variance components $\boldsymbol{\theta}$ is an unknown scalar or vector.

Model (38.9) is comprehensive and encompasses a variety of models. For continuous outcome data, by setting

$$h(\alpha) = \frac{1}{2}\alpha^2, \quad c(y, \tau^2) = \frac{1}{2}y^2/\tau^2 - \frac{1}{2}\log(2\pi\tau^2)$$

and $g(\cdot)$ to be an identity function, model (38.9) reduces to a linear mixed model. For binary outcome data, let

$$h(\alpha) = \log[1 + \exp(\alpha)] .$$

Choosing $g(\mu) = \text{logit}(\mu)$ yields a logit random effects model, while choosing $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi(\cdot)$ is the CDF for a standard normal, gives a probit random effects model.

From (38.7) and (38.8) it is easy to construct the likelihood that the inference will be based on. That is,

$$\ell = \sum_{i=1}^{m} \log \int \prod_{j=1}^{n_i} f(Y_{ij}|\boldsymbol{b}_i;\boldsymbol{\beta}) f(\boldsymbol{b}_i;\boldsymbol{\theta}) \mathrm{d}\boldsymbol{b}_i ,$$

where the integration is over the r-dimensional random effect $\boldsymbol{b}_i$ and the summation results from independence across clusters.

We can also reformulate model (38.9) in a compact form that encompasses all of the data from all of the clusters. Using $\boldsymbol{Y}$, $\mathbf{X}$, $\mathbf{Z}$, $\boldsymbol{b}$ as defined in the previous section, we write

$$g[E(\boldsymbol{Y}|\boldsymbol{b})] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} . \tag{38.10}$$

Hence, the log-likelihood function can be rewritten as

$$\begin{aligned} \ell(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta}) &= \log L(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta}) \\ &= \log \int f(\boldsymbol{Y}|\boldsymbol{b};\boldsymbol{\beta}) f(\boldsymbol{b};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{b} , \end{aligned} \tag{38.11}$$

where $f(\boldsymbol{Y}|\boldsymbol{b};\boldsymbol{\beta})$ is the conditional likelihood for $\boldsymbol{Y}$ and $f(\boldsymbol{b};\boldsymbol{\theta})$ is the density function for $\boldsymbol{b}$, often assumed to have a mean of zero.

Model (38.10) is not a simple reformat – it accommodates more complex data structures than clustered data. For example, with a properly defined $\mathbf{Z}$ and random effects $\boldsymbol{b}$ it encompasses crossed factor data [38.8] and nonnormal spatial data [38.14]. Hence, for more generality, the inferential procedures that we encounter in Sect. 38.4 will be based on (38.10, 11).

The GLMM is advantageous when the objective is to make inferences about individuals rather than the population average. Within its framework, random effects can be estimated and each individual's profile or growth curve can be obtained. The best predictor of random effects minimizing (38.6) is $E(\boldsymbol{Y}|\boldsymbol{b})$, which is not necessarily linear in $\boldsymbol{Y}$. However, if we confine our interest to the predictors that are linear in $\boldsymbol{Y}$, or of the form

$$\hat{\boldsymbol{b}} = \boldsymbol{c} + \mathbf{Q}\boldsymbol{Y}$$

for some conformable vector $\boldsymbol{c}$ and matrix $\mathbf{Q}$, minimizing the mean squared error (38.6) with respect to $\boldsymbol{c}$ and $\mathbf{Q}$ leads to the best *linear* predictor

$$\hat{\boldsymbol{b}} = E(\boldsymbol{b}) + \text{cov}(\boldsymbol{b}, \boldsymbol{Y})\text{var}(\boldsymbol{Y})[\boldsymbol{Y} - E(\boldsymbol{Y})] , \tag{38.12}$$

which holds true without any normality assumptions [38.11].

For example, consider a beta-binomial model for clustered binary outcomes such that

$$Y_{ij}|b_i \sim \text{Bernoulli}(b_i)$$

and the random effect $b_i \sim \text{Beta}(\alpha, \eta)$, where $\alpha, \eta > 0$. Using (38.12) we obtain the best linear predictor for b_i,

$$\hat{b}_i = \frac{\alpha + \bar{Y}_i}{\alpha + \beta + 1} ,$$

where $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$.

38.4 Computing MLEs for GLMMs

A common theme when fitting a GLMM has been the difficulty involved with computating likelihood-based inference. Indeed, computing the likelihood itself is often challenging for GLMMs, mostly because of intractable integrals. This section presents various commonly used likelihood-based approaches to estimating the coefficients and variance components in GLMMs.

38.4.1 The EM Approach

The EM algorithm [38.15] is a widely used approach to calculating MLEs with missing observations. The basic idea behind its application to the random effects models is to treat the random terms as 'missing' data, and to impute the missing information based on the observed data. Imputations are often made via conditional expectations.

When drawing inference, our goal is to maximize the marginal likelihood of the observed data in order to obtain the MLEs for unknown $\boldsymbol{\beta}$ and variance components $\boldsymbol{\theta}$. If random effects ($\boldsymbol{b}$) were observed, we would be able to write the 'complete' data as ($\boldsymbol{Y}, \boldsymbol{b}$) with a joint log-likelihood

$$\ell(\boldsymbol{Y}, \boldsymbol{b}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \log f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) + \log f(\boldsymbol{b}; \boldsymbol{\theta}) . \tag{38.13}$$

However, since $\boldsymbol{b}$ is unobservable, directly computing (38.13) is not feasible. Instead, the EM algorithm adopts a two-step iterative process. The expectation step ("E" step) computes the expectation of (38.13) conditional on the observed data. That is,

$$\tilde{\ell} = E\{\ell(\boldsymbol{Y}, \boldsymbol{b}; \boldsymbol{\beta}, \boldsymbol{\theta})|\boldsymbol{Y}, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0\} ,$$

where $\boldsymbol{\beta}_0, \boldsymbol{\theta}_0$ are the current values, followed by the maximization step ("M" step), which maximizes $\tilde{\ell}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The E and M steps are iterated until convergence is achieved. Generally, the E step is computationally intensive, because it still needs to calculate a high-dimensional integral.

Indeed, since the conditional distribution of $\boldsymbol{b}|\boldsymbol{Y}$ involves the marginal distribution $f(\boldsymbol{Y})$, which is an intractable integral, a direct Monte Carlo simulation cannot fulfill the expectation step. In view of this difficulty, *McCulloch* [38.7] utilized the Metropolis–Hastings algorithm to make random draws from $\boldsymbol{b}|\boldsymbol{Y}$ without calculating the marginal density $f(\boldsymbol{Y})$.

The Metropolis–Hastings algorithm, dated back to the papers by *Metropolis* et al. [38.16] and *Hastings* [38.17], can be summarized as follows. Choose an auxiliary function $q(\boldsymbol{u}, \boldsymbol{v})$ such that $q(., \boldsymbol{v})$ is a pdf for all $\boldsymbol{v}$. This function is often called a *jumping distribution* from point $\boldsymbol{v}$ to $\boldsymbol{u}$. Draw $\boldsymbol{b}^*$ from $q(., \boldsymbol{b})$, where $\boldsymbol{b}$ is the current value of the Markov chain. Compute the ratio of importance

$$\omega = \frac{f(\boldsymbol{b}|\boldsymbol{Y})q(\boldsymbol{b}^*, \boldsymbol{b})}{f(\boldsymbol{b}^*|\boldsymbol{Y})q(\boldsymbol{b}, \boldsymbol{b}^*)} .$$

Set the current value of the Markov chain as $\boldsymbol{b}^*$ with probability $\min(1, \omega)$, and $\boldsymbol{b}$ has a probability $\max(0, 1 - \omega)$. It can be shown that, under mild conditions, the distribution of $\boldsymbol{b}$ drawn from such a procedure converges weakly to $f(\boldsymbol{b}|\boldsymbol{Y})$ (see, for example, [38.18]). Since the unknown density $f(\boldsymbol{Y})$ cancels out in the calculation of ω, the Metropolis–Hastings algorithm has successfully avoided computing $f(\boldsymbol{Y})$.

The ideal Metropolis–Hastings algorithm jumping rule is to sample the point directly from the target distribution. That is, in our case, $q(\boldsymbol{b}^*, \boldsymbol{b}) = f(\boldsymbol{b}^*|\boldsymbol{Y})$ for all $\boldsymbol{b}$. Then the ratio of importance, ω, is always 1, and the iterations of $\boldsymbol{b}^*$ are a sequence of independent draws from $f(\boldsymbol{b}^*|\boldsymbol{Y})$. In general, however, iterative simulation is applied to situations where direct sampling is not possible. Efficient jumping rules have been addressed by *Gelman* et al. [38.19].

We can now turn to the Monte Carlo EM algorithm, which takes the following form.

1. Choose initial values $\boldsymbol{\beta}^0$ and $\boldsymbol{\theta}^0$.

2. Denote the updated value at iteration s by $(\boldsymbol{\beta}^s, \boldsymbol{\theta}^s)$. Generate n values of $\boldsymbol{b}^1, \ldots, \boldsymbol{b}^n$ from $f(\boldsymbol{b}|\boldsymbol{Y}; \boldsymbol{\beta}^s, \boldsymbol{\theta}^s)$.
3. At iteration $s+1$, choose $\boldsymbol{\beta}^{s+1}$ to maximize $\frac{1}{n}\sum_{k=1}^{n} \log f(\boldsymbol{Y}|\boldsymbol{b}^k; \boldsymbol{\beta})$.
4. Find $\boldsymbol{\theta}^{s+1}$ to maximize $\frac{1}{n}\sum_{k=1}^{n} \log f(\boldsymbol{b}^k; \boldsymbol{\theta})$.
5. Repeat steps 2–4 until convergence.

While computationally intensive, this algorithm is relatively stable since the log marginal likelihood increases at each iteration step and it is convergent at a linear rate [38.15].

38.4.2 Simulated Maximum Likelihood Estimation

Implementation of the EM is often computationally intensive. A naive approach would be to numerically approximate the likelihood (38.11) and maximize it directly. For example, when the random effects ($\boldsymbol{b}$) follow a normal distribution, we may use Gaussian quadrature to evaluate (38.11) and its derivatives. However, this approach quickly fails when the dimensions of $\boldsymbol{b}$ are large. We now consider a simulation technique, namely, simulated maximum likelihood estimation, to approximate the likelihood directly and, further, to obtain the MLEs. The key idea behind this approach is to approximate (38.11) and its first- and second-order derivatives by Monte Carlo simulations while performing Newton-Raphson iterations.

We begin with the likelihood approximation. Following *Geyer* and *Thompson* [38.20] and *Gelfand* and *Carlin* [38.21], one notices that for any density function $h(\boldsymbol{b})$ with the same support as $f(\boldsymbol{b}; \boldsymbol{\theta})$,

$$L(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \int \frac{f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}; \boldsymbol{\theta})}{h(\boldsymbol{b})} h(\boldsymbol{b}) \mathrm{d}\boldsymbol{b} \,. \quad (38.14)$$

Hence, Monte Carlo simulations can be applied to evaluate $L(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta})$. Explicitly, if $\boldsymbol{b}^1, \ldots, \boldsymbol{b}^n$ are generated independently from $h(\boldsymbol{b})$ (termed an *importance sampling distribution*), (38.14) can be approximated by

$$1/n \sum_{i=1}^{n} \frac{f(\boldsymbol{Y}|\boldsymbol{b}^i; \boldsymbol{\beta}) f(\boldsymbol{b}^i; \boldsymbol{\theta})}{h(\boldsymbol{b}^i)} \quad (38.15)$$

with an accuracy of order $O_p(n^{-1/2})$. The optimal (in the sense that the Monte Carlo approximation has zero variance) importance sampling distribution is $f(\boldsymbol{b}|\boldsymbol{Y})$, evaluated at the MLEs [38.22]. However, since the MLEs are unknown and the conditional distribution cannot be evaluated, such an optimal distribution is never meaningful practically. Nevertheless, we can find a distribution (such as a normal distribution) to approximate $f(\boldsymbol{b}|\boldsymbol{Y})$.

More specifically, notice that

$$\begin{aligned} f(\boldsymbol{b}|\boldsymbol{Y}) &= c \times f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}; \boldsymbol{\theta}) \\ &= c \times \exp[-K(\boldsymbol{Y}, \boldsymbol{b})] \,, \end{aligned}$$

where c (which does not depend on $\boldsymbol{b}$) is used to ensure a proper density function. We use

$$\begin{aligned} h(\boldsymbol{b}; \boldsymbol{\beta}, \boldsymbol{\theta}) =& ||2\pi \hat{\boldsymbol{\Sigma}}||^{-1/2} \\ & \times \exp\left[-\frac{1}{2}(\boldsymbol{b}-\hat{\boldsymbol{b}})' \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{b}-\hat{\boldsymbol{b}})\right] \,, \end{aligned}$$

where $||\cdot||$ denotes the determinant of a square matrix, $\hat{\boldsymbol{b}} = \mathrm{argmin}_{\boldsymbol{b}}[K(\boldsymbol{Y}, \boldsymbol{b})]$ and $\hat{\boldsymbol{\Sigma}} = [\frac{\partial}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} K(\boldsymbol{Y}, \hat{\boldsymbol{b}})]^{-1}$, to approximate the conditional density $f(\boldsymbol{b}|\boldsymbol{Y})$ evaluated at $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Similarly, the derivatives of $L(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta})$ can also be approximated by Monte Carlo simulations.

Then the algorithm proceeds as follows:

1. Choose the initial values $\boldsymbol{\gamma}^0 = (\boldsymbol{\beta}^0, \boldsymbol{\theta}^0)$ for $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\theta})$.
2. Denote the current value at the sth step by $\boldsymbol{\gamma}^s$. Generate $\boldsymbol{b}^1, \ldots, \boldsymbol{b}^n$ based on $h(\boldsymbol{b}|\boldsymbol{\gamma}^s)$.
3. Calculate the approximate derivatives of the marginal likelihood function $L(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta})$ evaluated at $\boldsymbol{\gamma}^s$:

$$\mathcal{B}_\beta^s = \frac{1}{n}\sum_{k=1}^{n} \frac{f(\boldsymbol{b}^k; \boldsymbol{\theta}^s)}{h(\boldsymbol{b}^k; \boldsymbol{\gamma}^s)} \frac{\partial}{\partial \boldsymbol{\beta}} f(\boldsymbol{Y}|\boldsymbol{b}^k; \boldsymbol{\beta})|_{\boldsymbol{\beta}^s} \,,$$

$$\mathcal{B}_\theta^s = \frac{1}{n}\sum_{k=1}^{n} \frac{f(\boldsymbol{Y}|\boldsymbol{b}^k; \boldsymbol{\beta}^s)}{h(\boldsymbol{b}^k; \boldsymbol{\gamma}^s)} \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{b}^k; \boldsymbol{\theta})|_{\boldsymbol{\beta}^s} \,,$$

$$\mathcal{A}_{\beta\beta}^s = \frac{1}{n}\sum_{k=1}^{n} \frac{f(\boldsymbol{b}^k; \boldsymbol{\theta}^s)}{h(\boldsymbol{b}^k; \boldsymbol{\gamma}^s)} \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} f(\boldsymbol{Y}|\boldsymbol{b}^k; \boldsymbol{\beta})|_{\boldsymbol{\beta}^s} \,,$$

$$\mathcal{A}_{\theta\theta}^s = \frac{1}{n}\sum_{k=1}^{n} \frac{f(\boldsymbol{Y}|\boldsymbol{b}^k, \boldsymbol{\beta}^s)}{h(\boldsymbol{b}^k; \boldsymbol{\gamma}^s)} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{b}^k; \boldsymbol{\theta})|_{\boldsymbol{\theta}^s} \,,$$

$$\begin{aligned} \mathcal{A}_{\beta\theta}^s =& \frac{1}{n}\sum_{k=1}^{n} \frac{1}{h(\boldsymbol{b}^k; \boldsymbol{\gamma}^s)} \frac{\partial}{\partial \boldsymbol{\beta}} f(\boldsymbol{Y}|\boldsymbol{b}^k; \boldsymbol{\beta})|_{\boldsymbol{\beta}^s} \\ & \times \left[\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{b}^k; \boldsymbol{\theta})|_{\boldsymbol{\theta}^s}\right]' \,. \end{aligned}$$

4. Compute the updated value at the $(s+1)$th step

$$\boldsymbol{\gamma}^{s+1} = \boldsymbol{\gamma}^s - (\mathcal{A}^s)^{-1} \mathcal{B}^s \,,$$

$$\text{where } \mathcal{A}^s = \begin{pmatrix} \mathcal{A}_{\beta\beta}^s & \mathcal{A}_{\beta\theta}^s \\ \left(\mathcal{A}_{\beta\theta}^s\right)' & \mathcal{A}_{\theta\theta}^s \end{pmatrix}$$

$$\text{and } \mathcal{B}^s = \left(\mathcal{B}_\beta^{s\prime}, \mathcal{B}_\theta^{s\prime}\right)' \,.$$

5. Repeat steps 2–4 until convergent criteria are met. Upon convergence, set $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^s$ and the Hessian matrix $\mathcal{A} = \mathcal{A}^s$.

The covariance of the resulting $\hat{\boldsymbol{\gamma}}$ is approximated (ignoring the Monte Carlo error) by the inverse of the observed information matrix, given by

$$-\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\log L(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta})|_{\hat{\boldsymbol{\gamma}}} \doteq -\hat{L}^{-1}\mathcal{A}\,,$$

where $\hat{L}$ and $\mathcal{A}$ are the approximations of $L(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta})$ and the Hessian matrix evaluated at $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$, respectively.

38.4.3 Monte Carlo Newton-Raphson (MCNR)/ Stochastic Approximation (SA)

Monte Carlo Newton-Raphson and stochastic approximation are two similar approaches to finding the MLEs for the GLMMs. They both approximate the score function using simulated random effects and improve the precision of the approximation at each iteration step.

We first describe a typical (MCNR) algorithm. Consider the decomposition of the joint density of the response vector and random effects vector

$$f(\boldsymbol{Y},\boldsymbol{b};\boldsymbol{\gamma}) = f(\boldsymbol{Y};\boldsymbol{\gamma})f(\boldsymbol{b}|\boldsymbol{Y};\boldsymbol{\gamma})\,.$$

Hence

$$\frac{\partial\log f(\boldsymbol{Y},\boldsymbol{b};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}} = S(\boldsymbol{\gamma}) + \frac{\partial\log f(\boldsymbol{b}|\boldsymbol{Y};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\,, \quad (38.16)$$

where $S(\boldsymbol{\gamma}) = \partial\log f(\boldsymbol{Y};\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$, the score function of main interest. In view of

$$E\left(\frac{\partial\log f(\boldsymbol{b}|\boldsymbol{Y};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}|\boldsymbol{Y}\right) = 0\,,$$

(38.16) can be written in the format of a regression equation

$$\frac{\partial\log f(\boldsymbol{Y},\boldsymbol{b};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}} = S(\boldsymbol{\gamma}) + \text{error}\,,$$

where the "error" term substitutes $\partial\log f(\boldsymbol{b}|\boldsymbol{Y};\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$, a mean zero term. Thus, inserting values of $\boldsymbol{b}\sim f(\boldsymbol{b}|\boldsymbol{Y})$ into $\partial\log f(\boldsymbol{Y},\boldsymbol{b};\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$ yields "data" to perform such a regression.

The MCNR algorithm is typically implemented as follows. Denote by $\boldsymbol{\gamma}^{(s)}$ the value of the estimate of $\boldsymbol{\gamma}$ at iteration step s. Generate via the Metropolis-Hastings algorithm a sequence of realized values $\boldsymbol{b}^{(s,1)},\ldots,\boldsymbol{b}^{(s,n)}\sim f(\boldsymbol{b}|\boldsymbol{Y};\boldsymbol{\gamma}^{(s)})$. At the $(s+1)$th step, compute

$$\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} - a_s\hat{E}\left(\frac{\partial\log f(\boldsymbol{Y},\boldsymbol{b};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(s)}}\right)\,. \quad (38.17)$$

Here a_s is a constant, incorporating information about the expectation of the derivative of $\partial\log f(\boldsymbol{Y},\boldsymbol{b};\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$ at the root, an unknown quantity. In practice, a_s is often set to be the inverse of a Monte Carlo estimate of the expectation based on the realized values of $\boldsymbol{b}^{(s,1)},\ldots,\boldsymbol{b}^{(s,n)}$.

The SA differs from the MCNR in that the SA uses a single simulated value of random effects in (38.17), that is

$$\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} - a_s\frac{\partial\log f(\boldsymbol{Y},\boldsymbol{b}^{(s)};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(s)}}\,,$$

and a_s is chosen to gradually decrease to zero. *Ruppert* and *Gu* and *Kong* have recommended that

$$a_s = \frac{e}{(s+\kappa)^\alpha}\left[\hat{E}\left(\frac{\partial^2\log f(\boldsymbol{Y},\boldsymbol{b};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right)\right]^{-1}\,,$$

where $e=3$, $\kappa=50$ and $\alpha=0.75$ as chosen by *McCulloch* and *Searle* [38.11]. The multiplier a_s decreases the step size as the iterations increase in the SA and eventually serves to eliminate the stochastic error involved in the Metropolis-Hastings steps. *McCulloch* and *Searle* [38.11] stated that the SA is advantageous in that it can use all of the simulated data to calculate estimates and only uses the simulated values one at a time; however, the detailed implementations of both methods are yet to be settled on in the literature.

38.4.4 S–U Algorithm

The S–U algorithm is a technique for finding the solution of an estimating equation that can be expressed as the expected value of a full data estimating equation, where the expectation is taken with respect to the missing data, given the observed data. This algorithm alternates between two steps: a simulation step wherein the missing values are simulated based on the conditional distributions given the observed data, and an updating step wherein parameters are updated without performing a numerical maximization. An attractive feature of this approach is that it is sequential – the number of Monte Carlo replicates does not have to be specified in advance, and the values of previous Monte Carlo replicates do not have to be stored or regenerated for later use. In the following, we will apply this approach in order to solve the maximum likelihood equations.

Differentiating the log-likelihood (38.26) with respect to the unknown parameters, $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\theta})$, gives

$$\boldsymbol{S}_b(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{f(\boldsymbol{Y}; \boldsymbol{\gamma})} \times \int \boldsymbol{S}_b(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{b} \,, \tag{38.18}$$

$$\boldsymbol{S}_t(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{1}{f(\boldsymbol{Y}; \boldsymbol{\gamma})} \times \int \boldsymbol{S}_t(\boldsymbol{b}; \boldsymbol{\theta}) f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{b} \,, \tag{38.19}$$

where $f(\boldsymbol{Y}; \boldsymbol{\gamma})$ is the marginal likelihood of the observed data set, and $\boldsymbol{S}_b(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta})$, $\boldsymbol{S}_t(\boldsymbol{b}; \boldsymbol{\theta})$ are conditional scores when treating $\boldsymbol{b}$ as observed constants, that is $\boldsymbol{S}_b(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) = \partial \log f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$, and $\boldsymbol{S}_t(\boldsymbol{b}; \boldsymbol{\theta}) = \partial \log f(\boldsymbol{b}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

Some algebra gives the second derivatives of the log-likelihood, which are needed in the algorithm. More specifically,

$$\boldsymbol{S}_{bb}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\boldsymbol{S}_b^{\otimes 2}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \frac{1}{f(\boldsymbol{Y}; \boldsymbol{\gamma})} \times \int \{\boldsymbol{S}_{bb}(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) + \boldsymbol{S}_b^{\otimes 2}(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta})\} f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{b} \,, \tag{38.20}$$

$$\boldsymbol{S}_{bt}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}'} = -\boldsymbol{S}_b(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{S}_t'(\boldsymbol{\beta}, \boldsymbol{\theta}) + \frac{1}{f(\boldsymbol{Y}; \boldsymbol{\gamma})} \times \int \boldsymbol{S}_b(\boldsymbol{\beta}, \boldsymbol{\theta}) \boldsymbol{S}_t'(\boldsymbol{b}; \boldsymbol{\theta}) f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{b} \,, \tag{38.21}$$

$$\boldsymbol{S}_{tt}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\boldsymbol{S}_t^{\otimes 2}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \frac{1}{f(\boldsymbol{Y}; \boldsymbol{\gamma})} \times \int \left\{\boldsymbol{S}_{tt}(\boldsymbol{b}; \boldsymbol{\theta}) + \boldsymbol{S}_t^{\otimes 2}(\boldsymbol{b}; \boldsymbol{\theta})\right\} f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{b} \,, \tag{38.22}$$

where $\boldsymbol{S}_{bb}(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta})$, $\boldsymbol{S}_{tt}(\boldsymbol{b}; \boldsymbol{\theta})$ are conditional information when treating $\boldsymbol{b}$ as observed constants, that is $\boldsymbol{S}_{bb}(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta}) = \partial^2 \log f(\boldsymbol{Y}|\boldsymbol{b}; \boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'$, and $\boldsymbol{S}_{tt}(\boldsymbol{b}; \boldsymbol{\theta}) = \partial^2 \log f(\boldsymbol{b}; \boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$. Here for a column vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}'$.

Hence, one can use the importance sampling scheme [38.23] to approximate these functions and their derivatives. We proceed as follows.

Having obtained the approximants $\hat{\boldsymbol{\gamma}}_1 = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\theta}}_1), \ldots, \hat{\boldsymbol{\gamma}}_j = (\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\theta}}_j)$ to $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$, the true MLE, at the jth S-step of the algorithm, we simulate $\boldsymbol{b}^{(j,l)}, l = 1, \ldots, n$, independently from $f(\boldsymbol{b}; \hat{\boldsymbol{\theta}}_j)$. Denote $w^{(j,l)}$ by

$$w^{(j,l)} = f\left(\boldsymbol{Y}|\boldsymbol{b}^{(j,l)}; \hat{\boldsymbol{\beta}}_j\right)$$

and let

$$\bar{w}_j = \frac{1}{j \cdot n} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \,.$$

As $j \to \infty$, the law of large numbers gives that $\bar{w}_j$ is asymptotically equal to $f(\boldsymbol{Y}; \hat{\boldsymbol{\gamma}})$ provided that $\hat{\boldsymbol{\gamma}}_j \xrightarrow{p} \hat{\boldsymbol{\gamma}}$.

We write

$$\bar{\boldsymbol{S}}_{b,j} = \frac{1}{jn\bar{w}_j} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \boldsymbol{S}_b\left(\boldsymbol{Y}|\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j\right) \,,$$

$$\bar{\boldsymbol{S}}_{t,j} = \frac{1}{jn\bar{w}_j} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \boldsymbol{S}_t\left(\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\theta}}_j\right) \,,$$

$$\bar{\boldsymbol{S}}_{bb,j} = -\bar{\boldsymbol{S}}_{b,j}^{\otimes 2} + \frac{1}{jn\bar{w}_j} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \times \left[\boldsymbol{S}_{bb}\left(\boldsymbol{Y}|\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j\right) + \boldsymbol{S}_b^{\otimes 2}\left(\boldsymbol{Y}|\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j\right)\right] \,,$$

$$\bar{\boldsymbol{S}}_{tt,j} = -\bar{\boldsymbol{S}}_{t,j}^{\otimes 2} + \frac{1}{jn\bar{w}_j} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \times \left[\boldsymbol{S}_{tt}\left(\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j\right) + \boldsymbol{S}_t^{\otimes 2}\left(\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j\right)\right] \,,$$

$$\bar{\boldsymbol{S}}_{bt,j} = -\bar{\boldsymbol{S}}_{b,j} \bar{\boldsymbol{S}}_{t,j}' + \frac{1}{jn\bar{w}_j} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \times \left[\boldsymbol{S}_b\left(\boldsymbol{Y}|\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j\right) \boldsymbol{S}_t'\left(\boldsymbol{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j\right)\right] \,,$$

With j sufficiently large, $\bar{\boldsymbol{S}}_{b,j}, \bar{\boldsymbol{S}}_{t,j}, \bar{\boldsymbol{S}}_{bb,j}, \bar{\boldsymbol{S}}_{bt,j}, \bar{\boldsymbol{S}}_{tt,j}$ provide good estimates for (38.18, 22).

Denote $\boldsymbol{S}_j = (\boldsymbol{S}_{b,j}', \boldsymbol{S}_{t,j}')'$ and

$$\boldsymbol{H}_j = \begin{pmatrix} \bar{\boldsymbol{S}}_{bb,j} & \bar{\boldsymbol{S}}_{bt,j} \\ \bar{\boldsymbol{S}}_{bt,j}' & \bar{\boldsymbol{S}}_{tt,j} \end{pmatrix} \,.$$

Then, at the jth U-step, the updated value for $\hat{\boldsymbol{\gamma}}$ is

$$\boldsymbol{\gamma}^{(j+1)} = \boldsymbol{\gamma}^{(j)} - a_j \boldsymbol{H}_j^{-1} \boldsymbol{S}_j \,,$$

where the tuning parameter a_j can be chosen as discussed in the previous section. Note that each of the quantities required at this step, such as $\bar{\boldsymbol{S}}_j$, $\bar{\boldsymbol{S}}_{\boldsymbol{\beta},j}$, and so on, can be calculated recursively so that the past values of these intermediate variables never need to be stored.

Following *Satten* and *Datta* [38.24], as $j \to \infty$, $\hat{\boldsymbol{\gamma}}_j$ almost surely converges to $\hat{\boldsymbol{\gamma}}$. Denote the S–U estimate

by $\hat{\boldsymbol{\gamma}}_{\mathrm{SU}}$. The total sampling variance of $\hat{\boldsymbol{\gamma}}_{\mathrm{SU}}$ around $\boldsymbol{\gamma}_0$ is the sum of the variance of $\hat{\boldsymbol{\gamma}}_{\mathrm{SU}}$ around $\hat{\boldsymbol{\gamma}}$ due to the S–U algorithm and the sampling variance of $\hat{\boldsymbol{\gamma}}$ around $\boldsymbol{\gamma}_0$ [38.25]. In most cases, the S–U algorithm should be iterated until the former is negligible compared to the latter. In theory, the starting value for the S–U algorithm is arbitrary. However, a poor starting value might cause instability at the beginning of this algorithm. Hence, in the next section, we consider several approximate methods that generate a starting value sufficiently close to the true zero of the estimating equations.

38.4.5 Some Approximate Methods

In view of the cumbersome and often intractable integrations required for a full likelihood analysis, several techniques have been made available for approximate inference in the GLMMs and other nonlinear variance component models.

The penalized quasi-likelihood (PQL) method introduced by *Green* [38.26] for semiparametric models was initially exploited as an approximate Bayes procedure to estimate regression coefficients. Since then, several authors have used the PQL to draw approximate inferences based on random effects models: *Schall* [38.5] and *Breslow* and *Clayton* [38.8] developed iterative PQL algorithms, *Lee* and *Nelder* [38.27] applied the PQL directly to hierarchical models. We present the PQL from the likelihood perspective below.

Consider the GLMM (38.10). For notational simplicity, we write the integrand of the likelihood function

$$f(\boldsymbol{Y}|\boldsymbol{b};\boldsymbol{\beta})f(\boldsymbol{b};\boldsymbol{\theta}) = \exp[-K(\boldsymbol{Y},\boldsymbol{b})]\,. \tag{38.23}$$

More generally, if one only specifies the first two conditional moments of $\boldsymbol{Y}$ given $\boldsymbol{b}$ in lieu of a full likelihood specification, $f(\boldsymbol{Y}|\boldsymbol{b};\boldsymbol{\beta})$ in (38.23) can be replaced by the quasi-likehood function $\exp[ql(\boldsymbol{Y}|\boldsymbol{b};\boldsymbol{\beta})]$, where

$$ql(\boldsymbol{Y}|\boldsymbol{b};\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\int_{Y_{ij}}^{\mu_{ij}^{b}} \frac{Y_{ij}-t}{V(t)}\mathrm{d}\,t\,.$$

Here $\mu_{ij}^{\mathrm{b}} = E(Y_{ij}|\boldsymbol{b};\boldsymbol{\beta})$ and $V(\mu_{ij}^{\mathrm{b}}) = \mathrm{var}(Y_{ij}|\boldsymbol{b};\boldsymbol{\beta})$.

Next evaluate the marginal likelihood. We temporarily assume that $\boldsymbol{\theta}$ is known. For any fixed $\boldsymbol{\beta}$, expanding $K(\boldsymbol{Y},\boldsymbol{b})$ around its mode $\hat{\boldsymbol{b}}$ up to the second-order term, we have

$$\begin{aligned} L(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta}) &= \int \exp[-K(\boldsymbol{Y},\boldsymbol{b})]\mathrm{d}\,\boldsymbol{b} \\ &= \left|\left|2\pi\left[K''(\boldsymbol{Y},\tilde{\boldsymbol{b}})\right]^{-1}\right|\right|^{1/2}\exp\left[-K(\boldsymbol{Y},\hat{\boldsymbol{b}})\right], \end{aligned}$$

where $K''(\boldsymbol{Y},\boldsymbol{b})$ denotes the second derivative of $K(\boldsymbol{Y},\boldsymbol{b})$ with respect to $\boldsymbol{b}$, and $\tilde{\boldsymbol{b}}$ lies in the segment joining zero and $\hat{\boldsymbol{b}}$. If $K''(\boldsymbol{Y},\boldsymbol{b})$ does not vary too much as $\boldsymbol{b}$ changes (for instance, $K''(\boldsymbol{Y},\boldsymbol{b}) = \textit{constant}$ for normal data), maximizing the marginal likelihood (38.11) is equivalent to maximizing

$$\mathrm{e}^{-K(\boldsymbol{Y},\hat{\boldsymbol{b}})} = f(\boldsymbol{Y}|\hat{\boldsymbol{b}},\boldsymbol{\beta})f(\hat{\boldsymbol{b}};\boldsymbol{\theta})\,.$$

Or, equivalently, $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{b}}(\boldsymbol{\theta})$ are obtained by jointly maximizing $f(\boldsymbol{Y}|\boldsymbol{b};\boldsymbol{\beta})f(\boldsymbol{b};\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{b}$ with $\boldsymbol{\theta}$ being held constant. If $\boldsymbol{\theta}$ is unknown, it can be estimated by maximizing the approximate profile likelihood of $\boldsymbol{\theta}$,

$$\left|\left|2\pi\left\{K''[\boldsymbol{Y},\hat{\boldsymbol{b}}(\boldsymbol{\theta})]\right\}^{-1}\right|\right|^{1/2}\exp\left\{-K[\boldsymbol{Y},\hat{\boldsymbol{b}}(\boldsymbol{\theta})]\right\}.$$

A more detailed discussion can be found in *Breslow* and *Clayton* [38.8].

As no closed-form solution is available, the PQL is often performed through an iterative process. In particular, *Schall* [38.5] derived an iterative algorithm where the random effects follow normal distributions. Specifically, with the current estimated values of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{b}$, a working 'response' $\tilde{\boldsymbol{Y}}$ is constructed by the first-order Taylor expansion of $g(\boldsymbol{Y})$ around $\boldsymbol{\mu}^{\mathrm{b}}$, or explicitly,

$$\begin{aligned} \tilde{\boldsymbol{Y}} &= g(\boldsymbol{\mu}^{\mathrm{b}}) + g'(\boldsymbol{\mu}^{\mathrm{b}})(\boldsymbol{Y}-\boldsymbol{\mu}^{\mathrm{b}}) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}b + g'(\boldsymbol{\mu}^{\mathrm{b}})(\boldsymbol{Y}-\boldsymbol{\mu}^{\mathrm{b}})\,, \end{aligned} \tag{38.24}$$

where $g(\cdot)$ is defined in (38.9).

Viewing the last term in (38.24) as a random error, (38.24) suggests fitting a linear mixed model on $\tilde{\boldsymbol{Y}}$ to obtain the updated values of $\boldsymbol{\beta}, \boldsymbol{b}$ and $\boldsymbol{\theta}$, which are used to recalculate the working 'response'. The iteration continues until convergence. Computationally, the PQL is easy to implement; it only requires repeatedly calling in existing software, for example, SAS 'PROC MIXED'. The PQL procedure yields exact MLEs for normally distributed data and for some cases when the conditional distribution of $\boldsymbol{Y}$ and the distribution of $\boldsymbol{b}$ are conjugate.

Other approaches, such as the Laplace method and the Solomon-Cox approximation, have also received much attention. The Laplace method (see for example *Liu* and *Pierce* [38.28]) differs from the PQL only in that the former obtains $\hat{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta})$ by maximizing the integrand $\mathrm{e}^{-K(\boldsymbol{Y},\boldsymbol{b})}$ with $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ being held fixed, and subsequently estimates $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})$ by jointly maximizing

$$\left|\left|2\pi\left[K''(\boldsymbol{Y},\hat{\boldsymbol{b}})\right]^{-1}\right|\right|^{1/2}\exp\left[-K(\boldsymbol{Y},\hat{\boldsymbol{b}})\right].$$

On the other hand, with the assumption of $E(\boldsymbol{b}) = 0$, the Solomon-Cox technique approximates

the integral $\int f(\boldsymbol{Y}|\boldsymbol{b})f(\boldsymbol{b})\mathrm{d}\boldsymbol{b}$ by expanding the integrand $f(\boldsymbol{Y}|\boldsymbol{b})$ around $\boldsymbol{b}=0$; see *Solomon* and *Cox* [38.29].

In general, none of these approximate methods produce consistent estimates,h except in some special cases, for example with normal data. Moreover, these methods are essentially based on normal approximation, and they often do not perform well for sparse data, such as binary data, and when the cluster size is relatively small [38.30].

38.5 Special Topics: Testing Random Effects for Clustered Categorical Data

It is useful to test for correlation within clusters and the heterogeneity among clusters when (or prior to) fitting random effects models. Tests have been proposed that are based on score statistics for the null hypothesis that variance components are zero for clustered continuous, binary and Poisson outcomes within the random effects model framework [38.31, 32]. However, literature that deals with tests for clustered polytomous data is scarce.

A recent article by *Li* and *Lin* [38.33] investigated tests for within-cluster correlation for clustered polytomous and censored discrete time-to-event data by deriving score tests for the null hypothesis that variance components are zero in random effects models. Since the null hypothesis is on the boundary of the parameter space, unlike the Wald and likelihood ratio tests whose asymptotic distributions are mixtures of chi-squares, the score tests are advantageous because their asymptotic distributions are still chi-square. Another advantage of the score tests is that no distribution needs to be assumed for the random effects except for their first two moments. Hence they are robust to mis-specifying the distributions of the random effects. Further, the Wald tests and the LR tests involve fitting random effects models that involve numerical integration, in contrast with the score tests, which only involve fitting standard models under the null hypothesis using existing standard software, and do not require numerical integration.

A common problem in the analysis of clustered data is the presence of covariate measurement errors. For example, in flood forecasting studies, the radar measurements of precipitation are 'highly susceptible' to error due to improper electronic calibration [38.34]; in AIDS studies, CD4 counts are often measured with error [38.35]. Valid statistical inference needs to account for measurement errors in covariates. *Li* and *Lin* [38.33] have extended the score tests for variance components to the situation where covariates are measured with errors. They applied the SIMEX method [38.36] to correct for measurement errors and develop SIMEX score tests for variance components. These tests are an extension of the SIMEX score test of *Lin* and *Carroll* [38.37] to clustered polytomous data with covariate measurement error.

Random effects-generalized logistic models and cumulative probability models have been proposed to model clustered nominal and ordinal categorical data [38.38, 39]. This section focuses on the score tests for the null hypothesis that the variance components are zero in such models to test for the within-cluster correlation.

38.5.1 The Variance Component Score Test in Random Effects-Generalized Logistic Models

Suppose that, for the jth ($j=1,\ldots,n_i$) subject in the i-th ($i=1,\ldots,m$) cluster, a categorical response Y_{ij} belongs to one of N categories indexed by $1,\ldots,N$. Conditional on the cluster-level random effect b_i, the observations Y_{ij} are independent and the conditional probability $P_{ij,k}=P(Y_{ij}=k|b_i)$ depends on the $p\times 1$ covariate vector $\boldsymbol{X}_{ij}$ through a generalized logistic model

$$\log\left(\frac{P_{ij,k}}{P_{ij,N}}\right)=\alpha_k+\boldsymbol{X}'_{ij}\boldsymbol{\beta}_k+b_i=\boldsymbol{X}'_{ij,k}\boldsymbol{\beta}+b_i\,,\quad k=1,\ldots,N-1 \tag{38.25}$$

where $\boldsymbol{\beta}_k$ is a $p\times 1$ vector of fixed effects, $b_i\sim F(b_i;\theta)$ for some distribution function F that has zero mean and a variance θ, $\boldsymbol{X}'_{ij,k}=\boldsymbol{e}'_k\otimes(1,\boldsymbol{X}'_{ij})$; $\otimes$ denotes a Kronecker product, $\boldsymbol{e}_k$ is an $(N-1)\times 1$ vector with the kth component equal to 1 and the rest of the components set to zero, and $\boldsymbol{\beta}=(\alpha_1,\boldsymbol{\beta}'_1,\cdots,\alpha_{N-1},\boldsymbol{\beta}'_{N-1})'$.

The marginal log-likelihood function for $(\boldsymbol{\beta},\theta)$ is

$$\ell(\boldsymbol{\beta},\theta)=\sum_{i=1}^{m}\log\int\exp[\ell_i(\boldsymbol{\beta},b_i)]\mathrm{d}\,F(b_i;\theta)\,, \tag{38.26}$$

where $\ell_i(\boldsymbol{\beta},b_i)=\sum_{j=1}^{n_i}\sum_{k=1}^{N}y_{ij,k}\log P_{ij,k}$, $y_{ij,k}=I(Y_{ij}=k)$ and $I(\cdot)$ is an indicator function. The magnitude of θ measures the degree of the within-cluster correlation. We are interested in testing $H_0:\theta=0$

vs. $H_1 : \theta > 0$, where $H_0 : \theta = 0$ corresponds to no within-cluster correlation. Since the null hypothesis is on the boundary of the parameter space, neither the likelihood ratio test nor the Wald test follows a chi-square distribution asymptotically [38.40].

Li and *Lin* [38.33] considered a score test for H_0 and showed that it still follows a chi-square distribution asymptotically. Specifically, they showed that the score statistic of θ evaluated under $H_0 : \theta = 0$ is

$$U_\theta(\boldsymbol{\beta}) = \left.\frac{\partial \ell(\boldsymbol{\beta}, \theta)}{\partial \theta}\right|_{\theta=0} = \sum_{i=1}^{m} \frac{1}{2}\left[\frac{\partial^2 \ell_i(\boldsymbol{\beta}, b_i)}{\partial b_i^2} + \left(\frac{\partial \ell_i(\boldsymbol{\beta}, b_i)}{\partial b_i}\right)^2\right]\Bigg|_{b_i=0} \tag{38.27}$$

$$= \frac{1}{2}\sum_{i=1}^{m}\left\{\left[\sum_{j=1}^{n_i}(\widetilde{Y}_{ij} - \widetilde{P}_{ij})\right]^2 - \sum_{j=1}^{n_i}\widetilde{P}_{ij}(1-\widetilde{P}_{ij})\right\}, \tag{38.28}$$

where

$$\widetilde{Y}_{ij} = \sum_{k=1}^{N-1} y_{ij,k} = I(Y_{ij} \le N-1)$$

and

$$\widetilde{P}_{ij} = \sum_{k=1}^{N-1} \exp(X'_{ij,k}\,\boldsymbol{\beta}) \Big/ \left[1 + \sum_{k=1}^{N-1} \exp(X'_{ij,k}\boldsymbol{\beta})\right]$$

is the mean of $\widetilde{Y}_{ij}$ under H_0. It is interesting to note that the form of (38.28) resembles the variance component score statistic for clustered binary data [38.31]. It can be shown that under $H_0 : \theta = 0$, $E[U_\theta(\boldsymbol{\beta})] = 0$ and $m^{-1/2}U_\theta(\boldsymbol{\beta})$ is asymptotically normal $MVN(0, I_{\theta\theta})$, where $I_{\theta\theta}$ is given in (38.30).

To study the properties of $U_\theta(\boldsymbol{\beta})$ under $H_1 : \theta > 0$, they expanded $E(\widetilde{Y}_{ij}|b_i)$ as a quadratic function of b_i, and showed that, under $H_1 : \theta > 0$,

$$E[U_\theta(\boldsymbol{\beta})] \approx \frac{1}{2}\sum_{i=1}^{m}\left[\sum_{j=1}^{n_i}\sum_{k\neq j}^{n_i} a_{ij}a_{ik} + \frac{1}{2}\sum_{j=1}^{n_i} a_{ij}\{a'_{ij}\}^2\right]\theta\,,$$

where $a_{ij} = \widetilde{P}_{ij}(1-\widetilde{P}_{ij})$ and $a'_{ij} = 1 - 2\widetilde{P}_{ij}$. As a result, $E[U_\theta(\boldsymbol{\beta})]$ is an increasing function of θ. Hence the test is consistent and one would expect a large value of $U_\theta(\boldsymbol{\beta})$ for a large value of θ.

Since $\boldsymbol{\beta}$ is unknown under H_0 and needs to be estimated, the score statistic for testing H_0 is

$$S = U_\theta(\hat{\boldsymbol{\beta}})/\widetilde{I}_{\theta\theta}^{1/2}(\hat{\boldsymbol{\beta}})\,, \tag{38.29}$$

where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under H_0 and can be easily obtained by fitting the generalized logistic model $\log(P_{ij,k}/P_{ij,N}) = \boldsymbol{X}'_{ij,k}\boldsymbol{\beta}$, (using SAS PROC CATMOD for example), and $\widetilde{I}_{\theta\theta} = I_{\theta\theta} - I_{\theta\boldsymbol{\beta}'}I^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}'}I_{\boldsymbol{\beta}\theta}$ is the efficient information of θ evaluated under $H_0 : \theta = 0$. Using L'Hôpital's rule, some calculations show that

$$I_{\theta\theta} = E\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] = \frac{1}{4}\sum_{i=1}^{m}\left[\sum_{j=1}^{n_i}\widetilde{P}_{ij}\widetilde{Q}_{ij}(1-6\widetilde{P}_{ij}\widetilde{Q}_{ij}) + 2\left(\sum_{j=1}^{n_i}\widetilde{P}_{ij}\widetilde{Q}_{ij}\right)^2\right], \tag{38.30}$$

$$I_{\boldsymbol{\beta}\boldsymbol{\beta}'} = \sum_{i=1}^{m} E\left(\frac{\partial \ell_i}{\partial \boldsymbol{\beta}}\frac{\partial \ell_i}{\partial \boldsymbol{\beta}'}\right) = \sum_{i=1}^{m}\mathbf{X}'_i\boldsymbol{\Sigma}_i\mathbf{X}_i\,, \tag{38.31}$$

$$I_{\theta\boldsymbol{\beta}'} = \sum_{i=1}^{m} E\left(\frac{\partial \ell_i}{\partial \theta}\frac{\partial \ell_i}{\partial \boldsymbol{\beta}'}\right) = \frac{1}{2}\sum_{i=1}^{m}\mathbf{P}'_i\{\mathbf{I}_{N-1}\otimes\mathbf{G}_i\}\mathbf{X}_i\,, \tag{38.32}$$

where the expectations are taken under H_0; $\mathbf{I}_{N-1}$ denotes an $(N-1)\times(N-1)$ identity matrix, and $\boldsymbol{X}_i = (\mathbf{X}'_{i1}, \ldots, \mathbf{X}'_{in_i})'$, where $\boldsymbol{X}_{ij} = (\boldsymbol{X}_{ij,1}, \ldots, \boldsymbol{X}_{ij,N-1})'$, $\tilde{Q}_{ij} = 1 - \tilde{P}_{ij}$, and $\boldsymbol{\Sigma}_i = (\boldsymbol{\Sigma}_{i,rl})$, which is an $(N-1)\times(N-1)$ block matrix whose (r,l)-th block is

$$\boldsymbol{\Sigma}_{i,rr} = \mathrm{diag}[P_{i1,r}(1-P_{i1,r}), \ldots, P_{in_i,r}(1-P_{in_i,r})]$$
$$\boldsymbol{\Sigma}_{i,rl} = \mathrm{diag}[-P_{i1,r}P_{i1,l}, \ldots, -P_{in_i,r}P_{in_i,l}]\,,\ r \neq l\,,$$

$\mathbf{G}_i = \mathrm{diag}(2\tilde{P}_{ij}^2 - 3\tilde{P}_{ij} + 1, \ldots, 2\tilde{P}_{in_i}^2 - 3\tilde{P}_{in_i} + 1)$ and $\mathbf{P}_i = (\boldsymbol{P}'_{i,1}, \ldots, \boldsymbol{P}'_{i,N-1})'$, where $\boldsymbol{P}_{i,r} = (P_{ij,r}, \ldots, P_{in_i,r})'$. Standard asymptotic calculations show that S is asymptotically $N(0,1)$ under H_0 and one rejects H_0 if S is large and the test is one-sided.

The score test S for $H_0 : \theta = 0$ has several attractive features. First, it can be easily obtained by fitting the generalized logistic model $\log(P_{ij,k}/P_{ij,N}) = \boldsymbol{X}'_{ij,k}\boldsymbol{\beta}$, which is model (38.25) under H_0, using standard software, such as SAS PROC CATMOD. Hence calculations of S do not involve any numerical integration. Secondly, it is the most powerful test locally. Finally it is robust, as no distribution is assumed for the random effect b_i. We discuss an application of the test based on (38.25) in Sect. 38.5.4.

38.5.2 The Variance Component Score Test in Random Effects Cumulative Probability Models

A widely used model for clustered ordinal data is the cumulative probability random effects model obtained by modeling the cumulative probabilities $r_{ij,k} = P(Y_{ij} \le k)$ as

$$g(r_{ij,k}) = \alpha_k + \boldsymbol{X}'_{ij}\boldsymbol{\beta}_x + b_i = \boldsymbol{X}'_{ij,k}\boldsymbol{\beta} + b_i \,, \quad k = 1, \ldots, N-1 \,, \tag{38.33}$$

where $g(\cdot)$ is a link function, $\boldsymbol{X}_{ij,k} = (\boldsymbol{e}'_k, \boldsymbol{X}'_{ij})'$, $\boldsymbol{\beta} = (\alpha_1, \cdots, \alpha_{N-1}, \boldsymbol{\beta}'_x)$, and $b_i \sim F(., \theta)$ for some distribution function F with zero mean and variance θ. For $g(\cdot) = \text{logit}(\cdot)$ and $g(\cdot) = \log[-\log(1-\cdot)]$, we have proportional odds and complementary log-log models. Define $o_{ij,k} = I(Y_{ij} \le k)$. Denote $\boldsymbol{r}_{ij} = (r_{ij,1}, \ldots, r_{ij,N-1})'$, $\boldsymbol{R}_i = (\boldsymbol{r}'_{i1}, \ldots, \boldsymbol{r}'_{in_i})'$ and define $\boldsymbol{o}_{ij}$, $\boldsymbol{O}_i$ similarly. Some calculations show that the score statistic of θ under $H_0 : \theta = 0$ is

$$U_\theta(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}\Big[(\boldsymbol{O}_i - \boldsymbol{R}_i)'\boldsymbol{\Gamma}_i^{-1}\boldsymbol{H}_i bf1_i bf1'_i \boldsymbol{H}_i \boldsymbol{\Gamma}_i^{-1} \times (\boldsymbol{O}_i - \boldsymbol{R}_i) - bf1'_i \tilde{\boldsymbol{W}}_i bf1_i\Big] \,, \tag{38.34}$$

where $\boldsymbol{1}_i$ is an $n_i(N-1)\times 1$ vector of ones; the weight matrices of $\boldsymbol{H}_i$, $\boldsymbol{\Gamma}_i$ and $\tilde{\boldsymbol{W}}_i$ are given in Appendix A.2 of *Li* and *Lin* [38.33]. Though seemingly complicated, (38.34) essentially compares the empirical variance of the weighted responses to its nominal variance.

The score statistic for testing $H_0 : \theta = 0$ is $S = U_\theta(\hat{\boldsymbol{\beta}}) \,/\, \tilde{I}^{1/2}_{\theta\theta}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under H_0, and it can be easily obtained by fitting the standard cumulative probability model $g(r_{ij,k}) = \boldsymbol{X}'_{ij,k}\boldsymbol{\beta}$, and $\tilde{I}_{\theta\theta}(\hat{\boldsymbol{\beta}})$ is the efficient information of θ. Computing the information matrices is tedious since the calculations involve the third and fourth cumulants of a multinomial distribution. The explicit expressions of the information matrices are given in *Li* and *Lin* [38.33].

Standard asymptotic calculations show that the score statistic S follows $N(0,1)$ asymptotically below H_0, and has the same optimality and robustness properties stated at the end of Sect. 38.5.1. It can be easily calculated by fitting the standard cumulative probability model $g(r_{ij,k}) = \boldsymbol{X}'_{ij,k}\boldsymbol{\beta}$ using existing software, such as SAS PROC CATMOD, and does not require any numerical integration. Again a one-sided test is used and H_0 is rejected for a large value of S. An application of score test based on (38.33) is presented in Sect. 38.5.4.

38.5.3 Variance Component Tests in the Presence of Measurement Errors in Covariates

Li and *Lin* [38.33] extended the variance component score tests to the situation when covariates are measured with error. To proceed, we denote a vector of unobserved covariates (such as the true precipitation level or the true CD4 count) by $\boldsymbol{X}_{ij}$ and $\boldsymbol{C}_{ij}$ denotes other accurately measured covariates (such as rainfall location or patients' gender).

The random effects cumulative probability model (38.33) and the random effects generalized logistic model (38.25) can be written in a unified form

$$g(p_{ij,k}) = \alpha_k + \boldsymbol{X}'_{ij}\boldsymbol{\beta}_{x,k} + \boldsymbol{C}'_{ij}\boldsymbol{\beta}_{c,k} + b_i \,, \tag{38.35}$$

where b_i follows some distribution $F(., \theta)$ with mean 0 and variance θ. For the random effects cumulative probability model (38.33), $p_{ij,k} = r_{ij,k}$ and $\boldsymbol{\beta}_{x,1} = \ldots = \boldsymbol{\beta}_{x,N-1}$ and $\boldsymbol{\beta}_{c,1} = \ldots = \boldsymbol{\beta}_{c,N-1}$. For the random effects generalized logistic model (38.25), $p_{ij,k} = P_{ij,k}/P_{ij,N}$ and $g(\cdot) = \log(\cdot)$.

Suppose the observed covariates $\mathbf{W}_{ij}$ (such as radar measurements of rainfall or observed CD4 counts) measure $\boldsymbol{X}_{ij}$ (such as the true precipitation amount or the true CD4 counts) with error. It is customary to postulate a nondifferential additive measurement error model for $\mathbf{W}_{ij}$ [38.41],

$$\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij} \,, \tag{38.36}$$

where the $\mathbf{U}_{ij}$ are independent measurement errors following $MVN(0, \boldsymbol{\Sigma}_u)$. Suppose that the measurement error covariance $\boldsymbol{\Sigma}_u$ is known or is estimated as $\hat{\boldsymbol{\Sigma}}_u$, using replicates or validation data for example. We are interested in testing for no within-cluster correlation $H_0 : \theta = 0$ in the random effects measurement error models (38.35) and (38.36). *Li* and *Lin* [38.33] have proposed using the SIMEX method by extending the results in the previous two sections to construct score tests for H_0 to account for measurement errors.

Simulation extrapolation (SIMEX) is a simulation-based functional method for inference on model parameters in measurement error problems [38.36], where no distributional assumption is made about the unobserved covariates $\mathbf{X}_{ij}$. We first briefly describe parameter estimation in random effects measurement error models (38.35, 36) using the SIMEX method, then discuss how to use the SIMEX idea to develop SIMEX score tests for $H_0 : \theta = 0$.

The SIMEX method involves two steps: the simulation step and the extrapolation step. In the simulation step, data $\mathbf{W}_{ij}^*$ is generated by adding to $\mathbf{W}_{ij}$ a random error following $N(0, \eta\boldsymbol{\Sigma}_u)$ for some constant $\eta > 0$. Naive parameter estimates are then calculated by fitting (38.35) with $\boldsymbol{X}_{ij}$ replaced by $\mathbf{W}_{ij}^*$. This gives the naive estimates if the measurement error covariance is $(1+\eta)\boldsymbol{\Sigma}_u$. This procedure is repeated for a large number B of times (for example $B = 100$), and the mean of the resulting B naive parameter estimates is calculated. One does this for a series of values of η (such as $\eta = 0.5, 1, 1.5, 2$). In the extrapolation step, a regression (such as a quadratic) model is fitted to the means of these naive estimates as a function of η, and is extrapolated to $\eta = -1$, which corresponds to zero measurement error variance. These extrapolated estimates give the SIMEX estimates for the model parameters. For details of the SIMEX method, see *Cook* and *Stefanski* [38.36] and *Carroll* et al. [38.41]. The SIMEX idea can be utilized to construct score tests for $H_0 : \theta = 0$ in the random effects measurement error models (38.35) and (38.36) by extending the results in Sects. 38.5.1 and 38.5.2. The resulting SIMEX score tests are an extension of the work of *Lin* and *Carroll* [38.37] to random effects measurement error models for clustered polytomous data.

In the absence of measurement error, the score statistics for testing $H_0 : \theta = 0$ under (38.35) take the same form $U_\theta(\hat{\boldsymbol{\beta}}) \,/\, \widetilde{I}_{\theta\theta}^{1/2}(\hat{\boldsymbol{\beta}})$, where $U_\theta(\hat{\boldsymbol{\beta}})$ is given in (38.34) for random effects cumulative probability models and in (38.28) for random effects generalized logistic models. The denominator $\widetilde{I}_{\theta\theta}(\hat{\boldsymbol{\beta}})$ is in fact the variance of $U_\theta(\hat{\boldsymbol{\beta}}|)$. The main idea of the SIMEX variance component score test is to treat the score statistic in the numerator $U_\theta(\cdot)$ as if it were a parameter estimator and use the SIMEX variance method (*Carroll* et al. [38.41]) to calculate the variance of this 'estimator'. Specifically, in the SIMEX simulation step, one simply calculates naive score statistics using the score formulae (38.34) and (38.28) by replacing $\boldsymbol{X}_{ij}$ with the simulated data $\boldsymbol{W}_{ij}^*$. The rest of the steps parallel those in the standard SIMEX method for parameter estimation. Denoting the results by $U_{\text{simex}}(\cdot)$ and $\widetilde{I}_{\theta\theta,\text{simex}}$ respectively, the SIMEX score statistic is simply

$$S_{\text{simex}} = U_{\text{simex}} / \widetilde{I}_{\theta\theta,\text{simex}}^{1/2} , \qquad (38.37)$$

which follows $N(0, 1)$ asymptotically when the true extrapolation function is used. Since the true extrapolation function is unknown in practice, an approximation (such as a quadratic) is used. The simulation study reported by *Li* and *Lin* [38.33] shows that the SIMEX score tests perform well. The theoretical justification for the SIMEX score tests can be found in *Lin* and *Carroll* [38.37].

The SIMEX score test possesses several important advantages. First, it can be easily calculated by fitting standard cumulative probability models using available software such as SAS PROC CATMOD. Secondly, it is robust in the sense that no distribution needs to be assumed for the frailty b_i and for the unobserved covariates $\boldsymbol{X}$.

38.5.4 Data Examples

To illustrate the variance component score tests for clustered polytomous data, we examine data from a longitudinal study on the efficacy of steam inhalation for treating common cold symptoms, conducted by *Macknin* et al. [38.42]. This study included 30 patients with colds of recent onset. At the time of enrolment, each patient went through two 20 min steam inhalation treatments spaced 60–90 minutes apart. Assessment of subjective response was made on an individual daily score card by the patient from day 1 (baseline) to day 4. On each day, the severity of nasal drainage was calibrated into four ordered categories (no symptoms, mild, moderate and severe symptoms). The study examined whether the severity improved following the treatment, and tested whether the observations over time for each subject were likely to be correlated.

Li and *Lin* [38.33] considered models (38.25) and (38.33) with the time from the baseline as a covariate. They first assumed a random effects logistic model (38.25), and obtained a variance component score statistic 5.32 (p-value < 0.001), which provided strong evidence for within-subject correlation over time. Similar results were found when they fitted a random effects proportional odds model (38.33) (score statistic = 9.70, p-value < 0.001). In these two tests they assumed no distribution for the random effect b_i.

To further examine the effect of time, they fitted (38.33) by further assuming that the random effect b_i followed $N(0, \theta)$. The MLE of the coefficient of time was -0.33 (SE $= 0.21$), which suggested that the severity improved following the treatment but that improvement was not statistically significant (p-value $= 0.11$). The estimated variance component was 2.31 (SE $= 0.45$). This result was consistent with the test results.

38.6 Discussion

Central to the idea of mixed modeling is the idea of fixed and random effects. Each effect in a model must be classified as either a fixed or a random effect. Fixed effects arise when the levels of an effect constitute the entire population of interest. For example, if an industrial experiment focused on the effectiveness of three brands of a machine, *machine* would be a fixed effect only if the experimenter's interest did not go beyond the three machine brands. On the other hand, an effect is classified as a random effect when one wishes to make inferences on an entire population, and the levels in the experiment represent only a sample from that population. Consider an example of psychologists comparing test results between different groups of subjects. Depending on the psychologists' particular interest, the group effect might be either fixed or random. For example, if the groups are based on the sex of the subject, *sex* would be a fixed effect. But if the psychologists are interested in the variability in test scores due to different teachers, they might choose a random sample of teachers as being representative of the total population of teachers, and *teacher* would be a random effect. Returning to the machine example presented earlier, *machine* would also be considered to be a random effect if the scientists were interested in making inferences on the entire population of machines and randomly chose three brands of machines for testing.

In summary, what makes a random effect unique is that each level of a random effect contributes an amount that is viewed as a sample from a population of random variables. The estimate of the variance associated with the random effect is known as the variance component because it measures the part of the overall variance contributed by that effect. In mixed models, we combine inferences about means (of fixed effects) with inferences about variances (of random effects).

A few difficulties arise from setting up the likelihood function to draw inference based on a random effects model. The major obstacle lies in computation, as, for practitioners, the main issue focuses on how to handle the intractable MLE calculations. This chapter reviews some commonly used approaches to estimating the regression coefficients and the variance components in the (generalized) linear mixed models. We note that the EM algorithm can yield maximum likelihood estimates, which are consistent and most efficient under regularity conditions. However, its computational burden is substantial, and the convergence rate is often slow. Laplace approximation greatly reduces the computational load, but the resulting estimates are generally biased. The simulated maximum likelihood estimation is considerably less computationally burdensome compared to the EM. For example, the rejection sampling is avoided, saving much computing time. However, its obvious drawback is the local convergence – a 'good' initial point is required to achieve the global maximizer. The so-called SA and S–U algorithms seem to be promising, as they make full use of the simulated data and obtain the estimates recursively. However, the detailed implementation of both methods have yet to be finalized in the literature.

It is worth briefly discussing marginal models, another major tool for handling clustered data. In a marginal model, the marginal mean of the response vector is modeled as a function of explanatory variables [38.43]. Thus, in contrast to the random effect models, the coefficients in a marginal model have population average interpretations. This type of model is typically fitted via the so-called generalized estimating equation (GEE). An appealing feature is that, for the right mean structure, even when the covariance structure of the response is mis-specified, the GEE acquires consistent estimates. However, the GEE method faces several difficulties, which may easily be neglected. First, the GEE estimator's efficiency becomes problematic when the variance function is mis-specified. Secondly, the consistency of the estimator is only guaranteed under noninformative censoring; informative censoring generally leads to biased estimates. More related discussion can be found in *Diggle* et al. [38.43].

Lastly, we point out other active research areas in mixed modeling, including evaluating the model's goodness of fit, choosing the best distribution for the random effects and selecting the best collection of covariates for a model. Readers are referred to some recent articles on these topics (such as [38.44–47]).

References

38.1 J. A. Nelder, R. W. Wedderburn: Generalized linear models, J. R. Stat. Soc. A **135**, 370–384 (1972)

38.2 P. McCullagh, J. A. Nelder: *Generalized Linear Models*, 2 edn. (Chapman Hall, London 1989) 1st edition, 1983

38.3 N. M. Laird, J. H. Ware: Random-effects models for longitudinal data, Biometrics **38**, 963–974 (1982)
38.4 R. Stiratelli, N. M. Laird, J. H. Ware: Random effects models for serial observations with binary response, Biometrics **40**, 961–971 (1984)
38.5 R. Schall: Estimation in generalized linear models with random effects, Biometrika **78**, 719–727 (1991)
38.6 S. L. Zeger, M. R. Karim: Generalized linear model with random effects: a Gibbs sampling approach, J. Am. Stat. Assoc. **86**, 79–86 (1991)
38.7 C. E. McCulloch: Maximum likelihood algorithms for generalized linear mixed models, J. Am. Stat. Assoc. **92**, 162–170 (1997)
38.8 N. E. Breslow, D. G. Clayton: Approximate inference in generalized linear mixed models, J. Am. Stat. Assoc. **88**, 9–25 (1993)
38.9 N. A. Cressie: *Statistics for Spatial Data* (Wiley, New York 1991)
38.10 D. A. Harville: Bayesian inference for variance components using only error contrasts, Biometrika **61**, 383–385 (1974)
38.11 C. E. McCulloch, S. R. Searle: *Generalized, Linear, and Mixed Models* (Wiley, New York 2001)
38.12 C. R. Henderson, O. Kempthorne, S. R. Searle, C. N. von Krosigk: Estimation of environmental, genetic trends from records subject to culling, Biometrics **15**, 192–218 (1959)
38.13 C. Bliss: The method of probits, Science **79**, 38–39 (1934)
38.14 P. J. Diggle, J. A. Tawn, R. A. Moyeed: Model-based geostatistics, J. R. Stat. Soc. C-AP **47**, 299–326 (1998)
38.15 A. P. Dempster, N. M. Laird, D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B **39**, 1–22 (1977)
38.16 N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth: Equation of state calculations by fast computing machines, J. Chem. Phys. **21**, 1087–1092 (1953)
38.17 W. Hastings: Monte Carlo sampling methods using Markov chains and their applications, Biometrika **57**, 97–109 (1970)
38.18 B. P. Carlin, T. A. Louis: *Bayes and Empirical Bayes Methods for Data Analysis* (Chapman Hall, New York 2000)
38.19 A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin: *Bayesian Data Analysis* (Chapman Hall, London 1995)
38.20 C. J. Geyer, E. A. Thompson: Constrained Monte Carlo maximization likelihood for dependent data, J. R. Stat. Soc. B **54**, 657–699 (1992)
38.21 A. E. Gelfand, B. P. Carlin: Maximum likelihood estimation for constrained- or missing-data problems, Can. J. Stat. **21**, 303–311 (1993)
38.22 C. P. Robert, G. Casella: *Monte Carlo Statistical Methods* (Springer, Berlin Heidelberg 1999)
38.23 M. A. Tanner, W. H. Wong: The calculation of posterior distributions by data augmentation, J. Am. Stat. Assoc. **82**, 528–549 (1987)
38.24 G. Satten, S. Datta: The S-U algorithm for missing data problems, Comp. Stat. **15**, 243–277 (2000)
38.25 G. Satten: Rank-based inference in the proportional hazards model for interval censored data, Biometrika **83**, 355–370 (1996)
38.26 P. J. Green: Penalized likelihood for general semi-parametric regression models, Int. Stat. Rev. **55**, 245–259 (1987)
38.27 Y. Lee, J. A. Nelder: Hierarchical generalized linear models, J. R. Stat. Soc. B **58**, 619–678 (1996)
38.28 Q. Liu, D. A. Pierce: Heterogeneity in Mantel-Haenszel-type models, Biometrika **80**, 543–556 (1993)
38.29 P. J. Solomon, D. R. Cox: Nonlinear component of variance models, Biometrika **79**, 1–11 (1992)
38.30 X. Lin, N. E. Breslow: Bias correction in generalized linear mixed models with multiple components of dispersion, J. Am. Stat. Assoc. **91**, 1007–1016 (1996)
38.31 D. Commenges, L. Letenneur, H. Jacqmin, J. Moreau, J. Dartigues: Test of homogeneity of binary data with explanatory variables, Biometrics **50**, 613–20 (1994)
38.32 X. Lin: Variance component testing in generalized linear models with random effects, Biometrika **84**, 309–326 (1997)
38.33 Y. Li, X. Lin: Testing random effects in uncensored/censored clustered data with categorical responses, Biometrics **59**, 25–35 (2003)
38.34 C. G. Collier: *Applications of Weather Radar Systems: A Guide to Uses of Radar in Meteorology and Hydrology* (Wiley, New York 1996)
38.35 A. A. Tsiatis, V. Degruttola, M. S. Wulfsohn: Modeling the relationship of survival to longitudinal data measured with error: applications to survival, CD4 counts in patients with AIDS, J. Am. Stat. Assoc. **90**, 27–37 (1995)
38.36 J. R. Cook, L. A. Stefanski: Simulation-extrapolation estimation in parametric measurement error models, J. Am. Stat. Assoc. **89**, 1314–1328 (1994)
38.37 X. Lin, R. J. Carroll: SIMEX variance component tests in generalized linear mixed measurement error models, Biometrics **55**, 613–619 (1999)
38.38 D. A. Harville, R. W. Mee: A mixed-model procedure for analyzing ordered categorical data, Biometrics **40**, 393–408 (1984)
38.39 D. Hedeker, R. Gibbons: A random-effects ordinal regression model for multilevel analysis, Biometrics **50**, 933–945 (1994)
38.40 S. G. Self, K. Y. Liang: Asymptotic properties of maximum likelihood estimators, likelihood ratio tests under nonstandard conditions, J. Am. Stat. Assoc. **82**, 605–610 (1987)
38.41 R. J. Carroll, D. Ruppert, L. A. Stefanski: *Measurement Error in Nonlinear Models* (Chapman Hall, London 1995)
38.42 M. L. Macknin, S. Mathew, S. V. Medendorp: Effect of inhaling heated vapor on symptoms of the

common cold, J. Am. Med. Assoc. **264**, 989–991 (1990)

38.43 P.J. Diggle, K.Y. Liang, S.L. Zeger: *Analysis of longitudinal data* (Oxford Univ. Press, New York 1994)

38.44 B. Zheng: Summarizing the goodness of fit of generalized linear models for longitudinal data, Stat. Med. **19**, 1265–1275 (2000)

38.45 G. Verbeke, E. Lesaffre: A linear mixed-effects model with heterogeneity in the random-effects population, J. Am. Stat. Assoc. **91**, 217–221 (1996)

38.46 P.J. Lindsey, J.K. Lindsey: Diagnostic tools for random effects in the repeated measures growth curve model, Comput. Stat. Data Anal. **33**, 79–100 (2000)

38.47 E.A. Houseman, L.M. Ryan, B.A. Coull: Cholesky residuals for assessing normal errors in a linear model with correlated outcomes, J. Am. Stat. Assoc. **99**, 383–394 (2004)