Random Effects Models

YI LI
Department of Biostatistics
Harvard School of Public Health and Dana-Farber Cancer Institute
44 Binney Street, M232
Boston, MA 02115, U.S.A.
Tel: 617-632-5134   Fax: 617-632-2444
yili@jimmy.harvard.edu

Summary

This chapter includes well-known as well as state-of-the-art statistical modelling techniques for drawing inference on correlated data, wich emerge from a wide range of studies, for example, in quality control study of products made from various assembly lines, in community based studies on cancer prevention, and in familial research of linkage analysis.

The first section briefly introduces the statistical models incorporating random effect terms, which have become increasingly popular in analyzing 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.

After this introduction, the second section introduces the linear mixed model for clustered data, which explicitly models the complex covariance structure among observations by adding random terms into the linear predictor part of a linear regression model. The third section discusses its extension, generalized linear mixed models, for correlated non-normal data.

The fourth section reviews several commonly estimating techniques for the GLMMs, including the EM approach, penalized Quasi-likelihood, the 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. Specifically score tests under various models are presented. The last section ends this chapter with discussion and some other relevant topics in random effects models.

Keywords: Random effects; Clustered data; Linear Mixed Models; Generalized Linear Mixed Models; Maximum Likelihood Estimation; EM algorithm; PQL; Stochastic Approximation; S-U algorithm; Simulated Maximum Likelihood Estimation; Score test; SIMEX; Variance Components.

1 Introduction

Classical linear regression models are a powerful tool for exploring the dependence of a response (e.g. blood pressure) on explanatory factors (e.g. weight, height and nutrient intake). However the normality assumption required for the response variables has severely limited its applicability. To accommodate a wide variety of independent non-normal data, Nelder and Wedderburn (1972) and McCullagh and Nelder (1983) introduced *generalized linear models (GLMs)*, a natural generalization of linear regression models. The GLMs allow responses to have non-Gaussian distributions. Hence, data of counts and proportions can conveniently be fitted within this framework. Typically in a GLM, the mean of a response is linked to the linear predictors via a non-random function, termed *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 from randomly selected manufactures. The beams made from the same factory are likely to be correlated because of the same manufacture 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 dependent. Random effects models have emerged in the last two decades as a major tool for analyzing such correlated data; see, e.g. Laird and Ware (1982), Stiratelli et al. (1984); Schall (1991), Zeger and Karim (1991) and McCulloch (1997), among others.

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

Though conceptually attractive, the GLMMs are often difficult to fit, to a large extent, 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 close-form. Typically cumbersome numerical integrations have to be performed. To alleviate such computational burden, various modeling techniques have been proposed. For example, Stiratelli et al.(1984) proposed an EM algorithm for fitting serial binary data; Schall (1991) developed an iterative Newton-Raphson algorithm; Zeger and Karim (1991) and McCulloch (1997) considered Monte Carlo EM methods. All these commonly used inferential procedures will be presented and discussed in this chapter.

The rest of this chapter is structured as follows. Section 2 introduces the linear mixed model for clustered data and section 3 discusses its extension, generalized linear mixed models, for correlated non-normal data. Section 4 reviews several commonly estimating techniques for the GLMMs, including the EM approach, penalized Quasi-likelihood, the Markov chain Newton-Raphson, the Stochastic Approximation and the S-U algorithm. Section 5 focuses on some special topics related to hypothesis tests of random effects. Section 6 concludes this chapter with discussion and some other relevant topics in random effects models.

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

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 or an alloy metal is more resilient. For this purpose the strength of the beams made of iron and alloy by randomly selected manufactures is measured. Each manufacture may contribute multiple beams, in which case each manufacture is deemed as a cluster, while each beam contributes as 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. Typically clustering induces dependence among observations. A linear mixed model (Laird and Ware, 1982) 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 present in an LMM. In data analysis, the decision as to 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 as a random sample of a larger population and therefore any effects that are not constant for all clusters are regarded as random.

To fix ideas, denote by $\mathbf{Y}_i$ the response vector for the i-th of a total m clusters, for example, the n_i observations of blood pressure taken on the i-th patient, $\mathbf{X}_i$ the known covariate matrix ($n_i \times p$) associated with the observations, e.g. the patient's treatment assignment and the time when the observation was taken, $\mathbf{b}_i$ the vector of random effects and $\mathbf{Z}_i$ the known design matrix associated with the random effects. Usually the columns of $\mathbf{Z}_i$ are a vector 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

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where typically we 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 $\mathbf{b}_i$, which is assumed to have expectation 0 for model identifiability. Here, $\mathbf{I}_{n_i}$ is an identity matrix of order n_i. In practice, we often assume $\mathbf{b}_i \sim MVN(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, 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 the information about the population where the clusters are randomly selected from and are often of interest to practitioners, aside from the fixed effects.

To encompass all data, we denote by $\mathbf{Y}, \mathbf{X}, \mathbf{b}, \boldsymbol{\epsilon}$ the concatenated collections of $\mathbf{Y}_i$'s, $\mathbf{X}_i$'s, $\mathbf{b}_i$'s and $\boldsymbol{\epsilon}_i$'s. For example, $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_m')'$. Denote by $\mathbf{Z}$ a block diagonal matrix whose i-th diagonal block is $\mathbf{Z}_i$. Then (1) can be compactly rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{b} \sim MVN(0, \mathbf{D})$, $\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 \mathbf{I}_{\mathcal{N}})$ and $\mathbf{b}$ 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, i.e. $\mathcal{N} = \sum_{i=1}^m n_i$.

Indeed, model (2) accommodates a much more general data structure beyond clustered data. For example, with properly defined $\mathbf{Z}$ and random effects $\mathbf{b}$ model (2) encompasses crossed factor data (Breslow and Clayton, 1993) and Gaussian spatial data (Cressie, 1991).

2.1 Estimation

Fitting of model (1) or its generalized version (2) is customarily likelihood based. A typical maximum likelihood estimation procedure is as follows.

First observe that $\mathbf{Y}$ is normally distributed, $\mathbf{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, 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

$$\ell = -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\log|\mathbf{V}| - \frac{\mathcal{N}}{2}\log 2\pi. \tag{3}$$

Denote by $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{\theta}', \sigma^2)'$ the collection of unknown parameters in the model. 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 $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{4}$$

Denote by θ_k the k-th 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

$$-\frac{1}{2}\left[tr(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_k}) - (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_k}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right] = 0, \tag{5}$$

where $tr(\cdot)$ denotes the trace of a square matrix. In practice, iterations between (4) and (5) are required to obtain the MLEs. Furthermore the asymptotic sampling variance are 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 (3).

It is, however, worth pointing out that the MLEs obtained from (4) and (5) are biased, especially for the variance components when the sample size is small. This is because the estimating equation (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 (Harville, 1974). The key idea is to replace the original response $\mathbf{Y}$ 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 $\mathbf{a}$ such that $\mathbf{a}'\mathbf{X} = 0$. For more efficiency we use the maximum number, $\mathcal{N} - p$, of linearly independent vectors $\mathbf{a}$ and write $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_{\mathcal{N}-p})$, which has a full row rank $\mathcal{N} - p$. The restricted MLE essentially will apply the MLE procedure on $\mathbf{A}'\mathbf{Y}$, in lieu of the original $\mathbf{Y}$.

To proceed, we note that $\mathbf{A}'\mathbf{Y} \sim MVN(0, \mathbf{A}'\mathbf{V}\mathbf{A})$. The ML equations for the variance components can hence be derived similarly from those for the original $\mathbf{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, namely, by replacing $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{V}$ with $\mathbf{A}'\mathbf{Y}$, 0 and $\mathbf{A}\mathbf{V}\mathbf{A}'$ respectively in (5).

Caution must be exercised if the MLEs or the RMLEs of the variance components fall out of the parameter space, e.g. 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 (2001).

2.2 Prediction of Random Effects

A fixed effect differs from a random effect in that the former is considered as a constant and is often the main parameter we wish to estimate. In contrast, a random effect is considered as an effect coming 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 manufacture (2) among manufactures. 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 i-th manufacture and X_{ij} indicates whether iron or alloy was used for producing such a beam. Note b_i is the effect on the strength of beams produced in the i-th manufacture; and this manufacture was just the one among the selected manufactures that happened to be labeled i in the study. The manufactures had been randomly selected as a representation of the population of all manufactures in the nation, and the inferences about the 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 some particular manufactures. For instance, one may want to rank various manufactures in order to select the best (or worst) ones. In these cases we will be interested in predicting b_i.

In general the 'best' prediction of $\mathbf{b}$ in (2) based on observed response $\mathbf{Y}$ is required to minimize the mean squared error

$$\int (\hat{\mathbf{b}} - \mathbf{b})'\mathbf{G}(\hat{\mathbf{b}} - \mathbf{b})f(\mathbf{Y}, \mathbf{b})d\mathbf{Y}d\mathbf{b}, \tag{6}$$

where the predictor $\hat{\mathbf{b}}$ depends only on $\mathbf{Y}$, $f(\mathbf{Y}, \mathbf{b})$ is the joint density function of $\mathbf{Y}$ and $\mathbf{b}$, and $\mathbf{G}$ is a given non-random positive definite matrix. It can be shown for any given $\mathbf{G}$, the minimizer is $E(\mathbf{b}|\mathbf{Y})$, i.e. the conditional expectation of $\mathbf{b}$ given the observed response $\mathbf{Y}$.

If the variance components were known, an analytical solution exists based on the linear mixed model (2). That is, assuming $\mathbf{Y}$ and $\mathbf{b}$ follow a joint multi-normal distribution, it follows

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

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

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

would yield the *Best linear unbiased predictor* (BLUP) of random effects (Henderson et al., 1959). Because **D** and **V** are typically unknown, they are often replaced by their MLEs or RMLEs when calculating the BLUP, namely

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

Extensive derivation has been given by Henderson et al. (1959) for the variance of the BLUP when the variance components are known. The variance of the BLUP with unknown variance components are yet fully available.

3 Generalized Linear Mixed Models

Non-normal data frequently arise from engineering studies. Consider again the beam study, where now the response is a binary variable, indicating whether a beam has satisfied the criteria of quality control. For such non-normal data, statistical models can be traced back as early as 1934, when Bliss (1934) proposed the first probit regression model for binary data. It was not, however, until four decades later did Nelder and Wedderburn (1972) and McCullagh and Nelder (1983 1st ed., 1989 2nd ed.) propose Generalized Linear Models (GLMs) to unify the models and modeling techniques for analyzing more general data (e.g. counted data and polytomous data). Several authors (Laird and Ware, 1982; Stiratelli et al., 1984; Schall, 1991, among others) have considered a natural generalization of the GLMs to accommodate correlated non-normal data. Their approach was adding random terms to the linear predictor parts and the resulting new models are termed Generalized Linear Mixed Models (GLMMs).

As an example, let Y_{ij} denote the status (e.g. pass or fail the quality assurance test) of the j th beam from the i-th manufacture. We might create a model as

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

where $logit(\mu) = log\{\mu/(1-\mu)\}$ is the link function that links together 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 the 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 $\mathbf{X}_{ij}$ be a $p \times 1$ covariate vector associated with response Y_{ij}. Conditional on an unobserved cluster-specific random variable $\mathbf{b}_i$ (an $r \times 1$ vector), Y_{ij} are independent and follow a distribution of exponential family, that is

$$\begin{aligned} Y_{ij}|\mathbf{b}_i &\overset{iid}{\sim} f(Y_{ij}|\mathbf{b}_i) & (7) \\ f(Y_{ij}|\mathbf{b}_i) &= \exp\{[Y_{ij}\alpha_{ij} - h(\alpha_{ij})]/\tau^2 - c(Y_{ij}, \tau)\}. & (8) \end{aligned}$$

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

$$g(\mu^b_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\mathbf{b}_i, \tag{9}$$

where $g(\cdot)$ is coined a *link function*, often chosen as an invertible and continuous function, and $\mathbf{Z}_{ij}$ is an $r \times 1$ design vector associated with the random effect. The random effects $\mathbf{b}_i$ 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 (9) is comprehensive and encompasses a variety of models. For continuous outcome data, by setting

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

and $g(\cdot)$ to be an identity function, model (9) reduces to a linear mixed model. For binary outcome data, let

$$h(\alpha) = \log\{1 + \exp(\alpha)\}.$$

Choosing $g(\mu) = logit(\mu)$ yields 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 (7) and (8) it is easy to construct the likelihood that the inference will be based on. That is,

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

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

We can further reformulate model (9) in such a compact form that it encompasses all the data from all all clusters. With $\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{b}$ as defined in the previous section, we write

$$g\{E(\mathbf{Y}|\mathbf{b})\} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}. \tag{10}$$

Hence, the log likelihood function can be rewritten as

$$\ell(\mathbf{Y}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \log L(\mathbf{Y}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \log \int f(\mathbf{Y}|\mathbf{b}; \boldsymbol{\beta}) f(\mathbf{b}; \boldsymbol{\theta}) d\mathbf{b}, \tag{11}$$

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

Model (10) is not a simple reformat - it accommodates more complex data structure beyond clustered data. For example, with properly defined $\mathbf{Z}$ and random effects $\mathbf{b}$ it encompasses crossed factor data (Breslow and Clayton, 1993) and non-normal spatial data (Diggle et al., 1998). Hence, for more generality, the ensuing inferential procedures in Section 4 will be based on (10) and (11).

The GLMM is advantageous when the objective is to make inference 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 (6) is $E(\mathbf{Y}|\mathbf{b})$, not necessarily linear in $\mathbf{Y}$. But if we confine our interest to the predictors which are linear in $\mathbf{Y}$, or, of the form

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

for some conformable vector $\mathbf{c}$ and matrix $\mathbf{Q}$, minimizing the mean squared error (6) with respect to $\mathbf{c}$ and $\mathbf{Q}$ leads to the best *linear* predictor

$$\hat{\mathbf{b}} = E(\mathbf{b}) + cov(\mathbf{b}, \mathbf{Y}) var(\mathbf{Y}) \{\mathbf{Y} - E(\mathbf{Y})\}, \tag{12}$$

which holds true without any normality assumptions (McCulloch and Searle, 2001).

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

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

and the random effect $b_i \sim Beta(\alpha, \eta)$, where $\alpha, \eta > 0$. Using (12) we obtain the best linear predictor for b_i,

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

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

4 Computing MLEs for GLMMs

A common theme in fitting a GLMM has been the difficulty of computation of 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 the GLMMs.

4.1 The EM Approach

The EM algorithm (Dempster et al., 1977) is a widely used approach to calculating the MLEs with missing observation. 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. Often, imputations are made via conditional expectations.

When drawing inference, our goal lies in maximizing 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 $\mathbf{b}$ were observed, we would be able to write the 'complete' data as $(\mathbf{Y}, \mathbf{b})$ with a joint log likelihood

$$\ell(\mathbf{Y}, \mathbf{b}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \log f(\mathbf{Y}|\mathbf{b}; \boldsymbol{\beta}) + \log f(\mathbf{b}; \boldsymbol{\theta}). \tag{13}$$

However, since $\mathbf{b}$ is unobservable, directly computing (13) is not feasible. Rather the EM algorithm adopts a two-step iterative process. The Expectation step ('E' step) computes the expectation of (13) conditional on the observed data. That is,

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

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 step and M step are iterated until convergence is achieved. Generally, the 'E' step is much computationally costly, where a need to calculate a high dimensional integral still exists.

Indeed, since the conditional distribution of $\mathbf{b}|\mathbf{Y}$ involves the marginal distribution $f(\mathbf{Y})$, which is an intractable integral, a direct Monte Carlo simulation is infeasible to fulfill the Expectation step. In view of this difficulty, McCulloch (1997) utilized the Metropolis-Hastings algorithm to make random draws from $\mathbf{b}|\mathbf{Y}$ without calculating the marginal density $f(\mathbf{Y})$.

The Metropolis-Hastings algorithm, dated back to the papers by Metropolis et al. (1953) and Hastings (1970), can be summarized as follows. Choose an auxiliary function $q(\mathbf{u}, \mathbf{v})$ such that $q(., \mathbf{v})$ is a pdf for all $\mathbf{v}$. This function is often called a *jumping distribution* from point $\mathbf{v}$ to $\mathbf{u}$. Draw $\mathbf{b}^*$ from $q(., \mathbf{b})$, where $\mathbf{b}$ is the current value of the Markov chain. Compute the ratio of importance

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

Set the current value of the Markov chain as $\mathbf{b}^*$ with probability $min(1, \omega)$ and $\mathbf{b}$ with probability $max(0, 1 - \omega)$. It can be shown under mild conditions the distribution of $\mathbf{b}$ drawn from such a procedure converges weakly to $f(\mathbf{b}|\mathbf{Y})$ (see, e.g. Carlin and Louis, 2000). Since the unknown density $f(\mathbf{Y})$ cancels out in the calculation of ω, the Metropolis-Hastings algorithm has successfully avoided computing $f(\mathbf{Y})$.

The ideal Metropolis-Hastings algorithm jumping rule is to sample the point directly from the target distribution. That is, in our case, $q(\mathbf{b}^*, \mathbf{b}) = f(\mathbf{b}^*|\mathbf{Y})$ for all $\mathbf{b}$. Then the ratio of importance, ω, is always 1, and the iterates $\mathbf{b}^*$ are a sequence of independent draws from $f(\mathbf{b}^*|\mathbf{Y})$. In general, however, iterative simulation is applied to the situations where direct sampling is not possible. Efficient jumping rules have been addressed by Gelman et al. (1995).

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 by $(\boldsymbol{\beta}^s, \boldsymbol{\theta}^s)$ the updated value at iteration s. Generate n values of $\mathbf{b}^1, \ldots, \mathbf{b}^n$ from $f(\mathbf{b}|\mathbf{Y}; \boldsymbol{\beta}^s, \boldsymbol{\theta}^s)$.

3. At iteration $s+1$, choose $\boldsymbol{\beta}^{s+1}$ to maximize $\frac{1}{n}\sum_{k=1}^n \log f(\mathbf{Y}|\mathbf{b}^k; \boldsymbol{\beta})$.

4. Find $\boldsymbol{\theta}^{s+1}$ to maximize $\frac{1}{n}\sum_{k=1}^n \log f(\mathbf{b}^k; \boldsymbol{\theta})$.

5. Repeat steps 2-4 until convergence.

While computationally intensive, this algorithm is relatively stable by increasing the log marginal likelihood at each iteration step and is convergent at a linear rate (Dempster et al., 1977).

4.2 Simulated Maximum Likelihood Estimation

Implementation of the EM is often computationally burdensome. A naive approach would numerically approximate the likelihood (11) and maximize it directly. For example, when the random effects $\mathbf{b}$ follow a normal distribution, we may use the Gaussian Quadrature to evaluate (11) and its derivatives. But this approach quickly fails when the dimension of $\mathbf{b}$ is 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 (11) and its first two order derivatives by Monte Carlo simulations while performing the Newton-Raphson iterations.

We begin with the likelihood approximation. Following Geyer and Thompson (1992) and Gelfand and Carlin (1993), one notices that for any density function $h(\mathbf{b})$ with the same support as $f(\mathbf{b}; \boldsymbol{\theta})$,

$$L(\mathbf{Y}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \int \frac{f(\mathbf{Y}|\mathbf{b}; \boldsymbol{\beta}) f(\mathbf{b}; \boldsymbol{\theta})}{h(\mathbf{b})} h(\mathbf{b}) d\mathbf{b}. \tag{14}$$

Hence, Monte Carlo simulations can be applied to evaluate $L(\mathbf{Y}; \boldsymbol{\beta}, \boldsymbol{\theta})$. Explicitly, if $\mathbf{b}^1, \ldots, \mathbf{b}^n$ are generated independently from $h(\mathbf{b})$ (termed *importance sampling distribution*), (14) can be approximated by

$$1/n \sum_{i=1}^{n} \frac{f(\mathbf{Y}|\mathbf{b}^i;\boldsymbol{\beta})f(\mathbf{b}^i;\boldsymbol{\theta})}{h(\mathbf{b}^i)} \tag{15}$$

with an accuracy order $O_p(n^{-1/2})$. The optimal (in the sense that the Monte Carlo approximation has 0 variance) importance sampling distribution is $f(\mathbf{b}|\mathbf{Y})$, evaluated at the MLEs (Robert and Casella, 1999). But, since the MLEs are unknown and the conditional distribution can not be evaluated, such an optimal distribution is never practically meaningful. Nevertheless we may find a distribution (e.g. normal distribution) to approximate $f(\mathbf{b}|\mathbf{Y})$.

More specifically, notice that

$$f(\mathbf{b}|\mathbf{Y}) = c \times f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})f(\mathbf{b};\boldsymbol{\theta}) = c \times \exp\{-K(\mathbf{Y},\mathbf{b})\},$$

where c (not depending on $\mathbf{b}$) is to ensure a proper density function. We use

$$h(\mathbf{b};\boldsymbol{\beta},\boldsymbol{\theta}) = ||2\pi\hat{\boldsymbol{\Sigma}}||^{-1/2} \exp\{-\frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{b}-\hat{\mathbf{b}})\},$$

where $||\cdot||$ denotes the determinant of a square matrix, $\hat{\mathbf{b}} = argmin_{\mathbf{b}}\{K(\mathbf{Y},\mathbf{b})\}$ and $\hat{\boldsymbol{\Sigma}} = \{\frac{\partial}{\partial\mathbf{b}\partial\mathbf{b}'}K(\mathbf{Y},\hat{\mathbf{b}})\}^{-1}$, to approximate the conditional density $f(\mathbf{b}|\mathbf{Y})$ evaluated at $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Similarly, the derivatives of $L(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\theta})$ 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 by $\boldsymbol{\gamma}^s$ the current value at the s-th step. Generate $\mathbf{b}^1,\ldots,\mathbf{b}^n$ based on $h(\mathbf{b}|\boldsymbol{\gamma}^s)$,

3. Calculate the approximate derivatives of marginal likelihood function $L(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\theta})$ evaluated at $\boldsymbol{\gamma}^s$.

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

$$\begin{aligned}\mathcal{A}^s_{\theta\theta} &= \frac{1}{n}\sum_{k=1}^{n}\frac{f(\mathbf{Y}|\mathbf{b}^k,\boldsymbol{\beta}^s)}{h(\mathbf{b}^k;\boldsymbol{\gamma}^s)}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f(\mathbf{b}^k;\boldsymbol{\theta})|_{\boldsymbol{\theta}^s},\\ \mathcal{A}^s_{\beta\theta} &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{h(\mathbf{b}^k;\boldsymbol{\gamma}^s)}\frac{\partial}{\partial\boldsymbol{\beta}}f(\mathbf{Y}|\mathbf{b}^k;\boldsymbol{\beta})|_{\boldsymbol{\beta}^s}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}f(\mathbf{b}^k;\boldsymbol{\theta})|_{\boldsymbol{\theta}^s}\right\}',\end{aligned}$$

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

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

where $\mathcal{A}^s = \begin{pmatrix} \mathcal{A}^s_{\beta\beta} & \mathcal{A}^s_{\beta\theta} \\ (\mathcal{A}^s_{\beta\theta})' & \mathcal{A}^s_{\theta\theta}\end{pmatrix}$ and $\mathcal{B}^s = (\mathcal{B}^{s}_{\beta}{}', \mathcal{B}^{s}_{\theta}{}')'$.

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

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

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

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

The Monte Carlo Newton-Raphson and Stochastic Approximation (Moyeed and Baddley, 1991; Rupert, 1991; Gu and Kong, 1998) are two similar approaches to finding the MLEs for the GLMMs. They both approximate the score function using the simulated random effects and improve the precision of 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

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

Hence

$$\frac{\partial \log f(\mathbf{Y}, \mathbf{b}; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = S(\boldsymbol{\gamma}) + \frac{\partial \log f(\mathbf{b}|\mathbf{Y}; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}, \tag{16}$$

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

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

(16) can be written in a format of a regression equation

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

where the 'error' term substitutes $\partial \log f(\mathbf{b}|\mathbf{Y}; \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$, a mean 0 term. Thus, inserting values of $\mathbf{b} \sim f(\mathbf{b}|\mathbf{Y})$ into $\partial \log f(\mathbf{Y}, \mathbf{b}; \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$ yields 'data' for performing 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 $\mathbf{b}^{(s,1)}, \ldots, \mathbf{b}^{(s,n)} \sim f(\mathbf{b}|\mathbf{Y}; \boldsymbol{\gamma}^{(s)})$. At the $(s+1)$-th step, compute

$$\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} - a_s \hat{E}\left[\frac{\partial \log f(\mathbf{Y}, \mathbf{b}; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(s)}}\right]. \tag{17}$$

Here a_s is a constant, incorporating information about the expectation of the derivative of $\partial \log f(\mathbf{Y}, \mathbf{b}; \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$ 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 $\mathbf{b}^{(s,1)}, \ldots, \mathbf{b}^{(s,n)}$.

The SA differs from the MCNR in that the SA uses a single simulated value of random effects in (17), that is

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

and a_s is chosen to gradually decrease to 0. Ruppert (1991) and Gu and Kong (1998) have recommended

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

where $e = 3, \kappa = 50$ and $\alpha = 0.75$ as chosen by McCulloch and Searle (2001). 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 (2001) 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 implementation of both methods have yet been 'settled' in the literature.

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, i.e. 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 to solve the maximum likelihood equations.

Differentiating the log likelihood (26) with respect to the unknown parameters, $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\theta})$, gives

$$\begin{aligned} \mathbf{S}_b(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{f(\mathbf{Y};\gamma)} \int \mathbf{S}_b(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta}) f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta}) f(\mathbf{b};\boldsymbol{\theta}) d\mathbf{b}, \qquad (18) \\ \mathbf{S}_t(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{1}{f(\mathbf{Y};\gamma)} \int \mathbf{S}_t(\mathbf{b};\boldsymbol{\theta}) f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta}) f(\mathbf{b};\boldsymbol{\theta}) d\mathbf{b} \qquad (19) \end{aligned}$$

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

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

$$\mathbf{S}_{bb}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\mathbf{S}_b^{\otimes 2}(\boldsymbol{\beta}, \boldsymbol{\theta})$$

$$
\begin{aligned}
& +\frac{1}{f(\mathbf{Y};\gamma)}\int\{\mathbf{S}_{bb}(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})+\mathbf{S}_b^{\otimes 2}(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})\}f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})f(\mathbf{b};\boldsymbol{\theta})d\mathbf{b}, \qquad (20)\\
\mathbf{S}_{bt}(\boldsymbol{\beta},\boldsymbol{\theta}) &= \frac{\partial^2\ell}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}'}=-\mathbf{S}_b(\boldsymbol{\beta},\boldsymbol{\theta})\mathbf{S}_t'(\boldsymbol{\beta},\boldsymbol{\theta})\\
& +\frac{1}{f(\mathbf{Y};\gamma)}\int\mathbf{S}_b(\boldsymbol{\beta},\boldsymbol{\theta})\mathbf{S}_t'(\mathbf{b};\boldsymbol{\theta})f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})f(\mathbf{b};\boldsymbol{\theta})d\mathbf{b} \qquad (21)\\
\mathbf{S}_{tt}(\boldsymbol{\beta},\boldsymbol{\theta}) &= \frac{\partial^2\ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}=-\mathbf{S}_t^{\otimes 2}(\boldsymbol{\beta},\boldsymbol{\theta})\\
& +\frac{1}{f(\mathbf{Y};\gamma)}\int\{\mathbf{S}_{tt}(\mathbf{b};\boldsymbol{\theta})+\mathbf{S}_t^{\otimes 2}(\mathbf{b};\boldsymbol{\theta})\}f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})f(\mathbf{b};\boldsymbol{\theta})d\mathbf{b}, \qquad (22)
\end{aligned}
$$

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

Hence, one can apply the importance sampling scheme (Tanner and Wong, 1987) to approximate these functions and their derivatives. We proceed as follows.

Having obtained 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 j-th S-step of the algorithm, we simulate $\mathbf{b}^{(j,l)}, l=1,\ldots,n$, independently from $f(\mathbf{b};\hat{\boldsymbol{\theta}}_j)$. Denote $w^{(j,l)}$ by

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

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 $f(\mathbf{Y};\hat{\boldsymbol{\gamma}})$ provided that $\hat{\boldsymbol{\gamma}}_j\xrightarrow{p}\hat{\boldsymbol{\gamma}}$.

We write

$$
\begin{aligned}
\bar{\mathbf{S}}_{b,j} &= \frac{1}{j\cdot n\cdot\bar{w}_j}\sum_{j'=1}^{j}\sum_{l=1}^{n}w^{(j',l)}\mathbf{S}_b(\mathbf{Y}|\mathbf{b}^{(j',l)};\hat{\boldsymbol{\beta}}_j),\\
\bar{\mathbf{S}}_{t,j} &= \frac{1}{j\cdot n\cdot\bar{w}_j}\sum_{j'=1}^{j}\sum_{l=1}^{n}w^{(j',l)}\mathbf{S}_t(\mathbf{b}^{(j',l)};\hat{\boldsymbol{\theta}}_j),\\
\bar{\mathbf{S}}_{bb,j} &= -\bar{\mathbf{S}}_{b,j}^{\otimes 2}+\frac{1}{j\cdot n\cdot\bar{w}_j}\sum_{j'=1}^{j}\sum_{l=1}^{n}w^{(j',l)}\{\mathbf{S}_{bb}(\mathbf{Y}|\mathbf{b}^{(j',l)};\hat{\boldsymbol{\beta}}_j)+\mathbf{S}_b^{\otimes 2}(\mathbf{Y}|\mathbf{b}^{(j',l)};\hat{\boldsymbol{\beta}}_j)\},
\end{aligned}
$$

$$
\begin{aligned}
\bar{\mathbf{S}}_{tt,j} &= -\bar{\mathbf{S}}_{t,j}^{\otimes 2} + \frac{1}{j \cdot n \cdot \bar{w}_j} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \{ \mathbf{S}_{tt}(\mathbf{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j) + \mathbf{S}_t^{\otimes 2}(\mathbf{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j) \}, \\
\bar{\mathbf{S}}_{bt,j} &= -\bar{\mathbf{S}}_{b,j} \bar{\mathbf{S}}_{t,j}' + \frac{1}{j \cdot n \cdot \bar{w}_j} \sum_{j'=1}^{j} \sum_{l=1}^{n} w^{(j',l)} \{ \mathbf{S}_b(\mathbf{Y} | \mathbf{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j) \mathbf{S}_t'(\mathbf{b}^{(j',l)}; \hat{\boldsymbol{\beta}}_j) \},
\end{aligned}
$$

With j sufficiently large, $\bar{\mathbf{S}}_{b,j}, \bar{\mathbf{S}}_{t,j}, \bar{\mathbf{S}}_{bb,j}, \bar{\mathbf{S}}_{bt,j}, \bar{\mathbf{S}}_{tt,j}$ provide good estimates for (18) - (22).

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

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

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

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

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 $\bar{\mathbf{S}}_j, \bar{\mathbf{S}}_{\boldsymbol{\beta},j}$, etc., can be calculated recursively so that the past values of these intermediate variables never need to be stored.

Following Satten and Datta (2000), as $j \to \infty$, $\hat{\boldsymbol{\gamma}}_j$ converges to $\hat{\boldsymbol{\gamma}}$ almost surely. Denote by $\hat{\boldsymbol{\gamma}}_{su}$ the S-U estimate. The total sampling variance of $\hat{\boldsymbol{\gamma}}_{su}$ around $\boldsymbol{\gamma}_0$ is the sum of the variance of $\hat{\boldsymbol{\gamma}}_{su}$ around $\hat{\boldsymbol{\gamma}}$ due to the S-U algorithm and the sampling variance of $\hat{\boldsymbol{\gamma}}$ around $\boldsymbol{\gamma}_0$ (Satten, 1996). 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.

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(1987) for semiparametric models has initially been exploited as an approximate Bayes procedure to estimate regression coefficients. Since then, several authors have explored the PQL to draw approximate inferences based on random effects models: Schall (1991) and Breslow and Clayton(1993) developed iterative PQL algorithms, Lee and Nelder (1996) applied the PQL directly to hierarchical models. We present below the PQL from the likelihood perspective.

Consider the GLMM (10). For notational simplicity we write the integrand of the likelihood function

$$f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})f(\mathbf{b};\boldsymbol{\theta}) = \exp\{-K(\mathbf{Y},\mathbf{b})\}. \tag{23}$$

More generally, if one only specifies the first two conditional moments of $\mathbf{Y}$ given $\mathbf{b}$ in lieu of a full likelihood specification, $f(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})$ in (23) can be replaced by the quasi-likehood function $\exp\{ql(\mathbf{Y}|\mathbf{b};\boldsymbol{\beta})\}$, where

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

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

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

$$L(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\theta}) = \int \exp\{-K(\mathbf{Y},\mathbf{b})\}d\mathbf{b} = ||2\pi\{K''(\mathbf{Y},\tilde{\mathbf{b}})\}^{-1}||^{1/2}\exp\{-K(\mathbf{Y},\hat{\mathbf{b}})\},$$

where $K''(\mathbf{Y},\mathbf{b})$ denotes the second derivative of $K(\mathbf{Y},\mathbf{b})$ with respect to $\mathbf{b}$, and $\tilde{\mathbf{b}}$ lies in the segment joining 0 and $\hat{\mathbf{b}}$. If $K''(\mathbf{Y},\mathbf{b})$ does not vary too much as $\mathbf{b}$ changes (for instance, $K''(\mathbf{Y},\mathbf{b}) = \mathit{constant}$ for normal data), maximizing the marginal likelihood (11) is equivalent to maximizing

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

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

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

A more detailed discussion can be found in Breslow and Clayton (1993).

As no close-form solution is available, the PQL is often performed through an iterative process. In particular, Schall (1991) derived an iterative algorithm when the random effects follow normal distributions. Specifically, with the current estimated values of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\mathbf{b}$, a working 'response' $\tilde{\mathbf{Y}}$ is constructed by the first order Taylor expansion of $g(\mathbf{Y})$ around $\boldsymbol{\mu}^b$, or explicitly,

$$\tilde{\mathbf{Y}} = g(\boldsymbol{\mu}^b) + g'(\boldsymbol{\mu}^b)(\mathbf{Y} - \boldsymbol{\mu}^b) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + g'(\boldsymbol{\mu}^b)(\mathbf{Y} - \boldsymbol{\mu}^b), \tag{24}$$

where $g(\cdot)$ is defined in (9).

Viewing the last term in (24) as a random error, (24) suggests fitting a linear mixed model on $\tilde{\mathbf{Y}}$ to obtain the updated values of $\boldsymbol{\beta}, \mathbf{b}$ and $\boldsymbol{\theta}$, which are used to recalculate the working 'response'. The iteration shall continue until convergence. Computationally, the PQL is easy to implement, which 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 $\mathbf{Y}$ and the distribution of $\mathbf{b}$ are conjugate.

Other approaches, such as the Laplace method and the Solomon-Cox approximation have also received much attention. The Laplace method (see, e.g. Liu and Pierce (1993)) differs from the PQL only in that the former obtains $\hat{\mathbf{b}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ by maximizing the integrand $e^{-K(\mathbf{Y},\mathbf{b})}$ with $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ being held fixed, and subsequently estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ by jointly maximizing

$$||2\pi\{K''(\mathbf{Y}, \hat{\mathbf{b}})\}^{-1}||^{1/2} \exp\{-K(\mathbf{Y}, \hat{\mathbf{b}})\}.$$

On the other hand, with the assumption of $E(\mathbf{b}) = 0$, the Solomon-Cox technique approximates the integral $\int f(\mathbf{Y}|\mathbf{b})f(\mathbf{b})d\mathbf{b}$ by expanding the integrand $f(\mathbf{Y}|\mathbf{b})$ around $\mathbf{b} = 0$; see Solomon and Cox (1992).

In general, none of these aforementioned approximate methods produces consistent estimates, with exception in some special cases, e.g. normal data. Moreover, these methods are essentially based on normal approximation, and they typically do not perform well for sparse data, e.g. binary data, and when the cluster size is relatively small (Breslow and Lin, 1995; Lin and Breslow, 1996).

5 Special Topics: Testing Random Effects for Clustered Categorical Data

When (or prior to) fitting random effects models, it is of substantial interest to test for the correlation within clusters and the heterogeneity among clusters. Such tests have been proposed by using score statistics for the null hypothesis that variance components are zero for clustered continuous, binary and Poisson outcomes within the random effects model framework (Commenges, et al., 1994; Lin, 1997). However, very few literatures have dealt with tests for clustered polytomous data.

A recent article by Li and Lin (2003) has considered tests for the 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 on the random effects needs to be assumed except for their first two moments. Hence they are robust to misspecification of the distributions of the random effects. Further the Wald tests and the LR tests require fitting random effects models which involve numerical integration, in contrast with the score tests, which only require 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 errors because of improper electronic calibration (Collier, 1996); in AIDS studies, CD4 counts are often measured with error (Tsiatis, et al., 1995). Valid statistical inference needs to account for measurement errors in covariates. Li and Lin (2003) have extended the score tests for variance components to the situation where covariates are measured with errors. They applied the SIMEX method (Cook and Stefanski, 1994) 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 (1999) 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 (Harville and Mee, 1984; Hedeker and Gibbons, 1994). 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.

5.1 The Variance Component Score Test in Random Effects Generalized Logistic Models

Suppose for the jth ($j = 1, \ldots, n_i$) subject in the ith ($i = 1, \ldots, m$) 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 $\mathbf{X}_{ij}$ through a generalized logistic model

$$\log\left(\frac{P_{ij,k}}{P_{ij,N}}\right) = \alpha_k + \mathbf{X}'_{ij}\boldsymbol{\beta}_k + b_i = \mathbf{X}'_{ij,k}\boldsymbol{\beta} + b_i, \qquad k = 1, \ldots, N-1 \qquad (25)$$

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 mean 0 and variance θ, $\mathbf{X}'_{ij,k} = \mathbf{e}'_k \otimes (1, \mathbf{X}'_{ij})$, $\otimes$ denotes a Kronecker product, $\mathbf{e}_k$ is an $(N-1) \times 1$ vector with the kth component equal to 1 and the rest components equal to 0, and $\boldsymbol{\beta} = (\alpha_1, \boldsymbol{\beta}'_1, \cdots, \alpha_{N-1}, \boldsymbol{\beta}'_{N-1})'$.

The marginal loglikelihood function for $(\boldsymbol{\beta}, \theta)$ is

$$\ell(\boldsymbol{\beta}, \theta) = \sum_{i=1}^{m} \log \int \exp\{\ell_i(\boldsymbol{\beta}, b_i)\} dF(b_i; \theta), \qquad (26)$$

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 (Self and Liang, 1987).

Li and Lin (2003) 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

$$\begin{aligned} U_\theta(\boldsymbol{\beta}) &= \left.\frac{\partial \ell(\boldsymbol{\beta}, \theta)}{\partial \theta}\right|_{\theta=0} = \sum_{i=1}^{m} \frac{1}{2}\left[\frac{\partial^2 \ell_i(\boldsymbol{\beta}, b_i)}{\partial b_i^2} + \left\{\frac{\partial \ell_i(\boldsymbol{\beta}, b_i)}{\partial b_i}\right\}^2\right]\Bigg|_{b_i=0} & (27) \\ &= \frac{1}{2}\sum_{i=1}^{m}\left[\left\{\sum_{j=1}^{n_i}(\widetilde{Y}_{ij} - \widetilde{P}_{ij})\right\}^2 - \sum_{j=1}^{n_i}\widetilde{P}_{ij}(1-\widetilde{P}_{ij})\right], & (28) \end{aligned}$$

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 (28) resembles the variance component score statistic for clustered binary data (Commenges, et al., 1994). 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 (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

$$S = U_\theta(\widehat{\boldsymbol{\beta}}) \Big/ \widetilde{I}_{\theta\theta}^{1/2}(\widehat{\boldsymbol{\beta}}), \qquad (29)$$

where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under H_0 and can be easily obtained by fitting the generalized logistic model $\log(P_{ij,k}/P_{ij,N}) = \mathbf{X}'_{ij,k}\boldsymbol{\beta}$, (e.g., using SAS PROC CATMOD), 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

$$\begin{aligned}
I_{\theta\theta} &= E\left\{\left(\frac{\partial \ell}{\partial \theta}\right)^2\right\} = \frac{1}{4}\sum_{i=1}^{m}\left\{\sum_{j=1}^{n_i}\widetilde{P}_{ij}\widetilde{Q}_{ij}(1-6\widetilde{P}_{ij}\widetilde{Q}_{ij}) + 2\left(\sum_{j=1}^{n_i}\widetilde{P}_{ij}\widetilde{Q}_{ij}\right)^2\right\}, \quad (30)\\
I_{\boldsymbol{\beta}\boldsymbol{\beta}'} &= \sum_{i=1}^{m} E\left\{\frac{\partial \ell_i}{\partial \boldsymbol{\beta}}\frac{\partial \ell_i}{\partial \boldsymbol{\beta}'}\right\} = \sum_{i=1}^{m}\mathbf{X}_i'\boldsymbol{\Sigma}_i\mathbf{X}_i, \quad (31)\\
I_{\theta\boldsymbol{\beta}'} &= \sum_{i=1}^{m} E\left\{\frac{\partial \ell_i}{\partial \theta}\frac{\partial \ell_i}{\partial \boldsymbol{\beta}'}\right\} = \frac{1}{2}\sum_{i=1}^{m}\mathbf{P}_i'\{\mathbf{I}_{N-1}\otimes\mathbf{G}_i\}\mathbf{X}_i, \quad (32)
\end{aligned}$$

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

$$\begin{aligned}
\boldsymbol{\Sigma}_{i,rr} &= diag\{P_{i1,r}(1-P_{i1,r}), \ldots, P_{in_i,r}(1-P_{in_i,r})\}\\
\boldsymbol{\Sigma}_{i,rl} &= diag\{-P_{i1,r}P_{i1,l}, \ldots, -P_{in_i,r}P_{in_i,l}\},\ r \neq l,
\end{aligned}$$

$\mathbf{G}_i = \mathrm{diag}(2\tilde{P}_{ij}^2 - 3\tilde{P}_{ij} + 1, \ldots, 2\tilde{P}_{in_i}^2 - 3\tilde{P}_{in_i} + 1)$ and $\mathbf{P}_i = (\mathbf{P}_{i,1}', \ldots, \mathbf{P}_{i,N-1}')'$, where $\mathbf{P}_{i,r} = (P_{ij,r}, \ldots, P_{in_i,r})'$. 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 $\log(P_{ij,k}/P_{ij,N}) = \mathbf{X}_{ij,k}'\boldsymbol{\beta}$, which is model (25) under H_0, using standard software, e.g., SAS PROC CATMOD. Hence calculations of S do not involve any numerical integration. Secondly, it is the locally most powerful test. Finally it is robust as no distribution is assumed for the random effect b_i. We discuss an application of the test based on (25) in Section 5.4.

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

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

$$g(r_{ij,k}) = \alpha_k + \mathbf{X}_{ij}'\boldsymbol{\beta}_x + b_i = \mathbf{X}_{ij,k}'\boldsymbol{\beta} + b_i, \qquad k = 1, \ldots, N-1 \quad (33)$$

where $g(\cdot)$ is a link function, $\mathbf{X}_{ij,k} = (\mathbf{e}'_k, \mathbf{X}'_{ij})'$, $\boldsymbol{\beta} = (\alpha_1, \cdots, \alpha_{N-1}, \boldsymbol{\beta}'_x)$, and $b_i \sim F(., \theta)$ for some distribution function F with mean 0 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 $o_{ij,k} = I(Y_{ij} \leq k)$. Denote by $\mathbf{r}_{ij} = (r_{ij,1}, \ldots, r_{ij,N-1})'$, $\mathbf{R}_i = (\mathbf{r}'_{i1}, \ldots, \mathbf{r}'_{in_i})'$ and $\mathbf{o}_{ij}$, $\mathbf{O}_i$ similarly. Some calculations show that the score statistic of θ under $H_0 : \theta = 0$ is

$$U_\theta(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}\left\{(\mathbf{O}_i - \mathbf{R}_i)'\boldsymbol{\Gamma}_i^{-1}\mathbf{H}_i\mathbf{1}_i\mathbf{1}'_i\mathbf{H}_i\boldsymbol{\Gamma}_i^{-1}(\mathbf{O}_i - \mathbf{R}_i) - \mathbf{1}'_i\tilde{\mathbf{W}}_i\mathbf{1}_i\right\}, \tag{34}$$

where $\mathbf{1}_i$ is an $n_i(N-1) \times 1$ vector of ones, the weight matrices of $\mathbf{H}_i, \boldsymbol{\Gamma}_i$ and $\tilde{\mathbf{W}}_i$ are given in Appendix A.2 of Li and Lin (2003). Though seemingly complicated, (34) essentially compares the empirical variance of the weighted responses to its nominal variance.

The score statistic for testing $H_0 : \theta = 0$ is $S = U_\theta(\hat{\boldsymbol{\beta}}) \Big/ \widetilde{I}_{\theta\theta}^{1/2}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under H_0 and can be easily obtained by fitting the standard cumulative probability model $g(r_{ij,k}) = \mathbf{X}'_{ij,k}\boldsymbol{\beta}$, and $\widetilde{I}_{\theta\theta}(\hat{\boldsymbol{\beta}})$ 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 (2003).

Standard asymptotic calculations show that the score statistic S follows $N(0,1)$ asymptotically under H_0, and has the same optimality and robustness properties stated at the end of Section 5.1. It can be easily calculated by fitting the standard cumulative probability model $g(r_{ij,k}) = \mathbf{X}'_{ij,k}\boldsymbol{\beta}$ using existing software, e.g., 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 (33) is presented in Section 5.4.

5.3 The Variance Component Tests in the Presence of Measurement Errors in Covariates

Li and Lin (2003) extended the variance component score tests to the situation when covariates are measured with error. To proceed, we denote by $\mathbf{X}_{ij}$ a vector of unobserved covariates (e.g., the true precipitation level or the true CD4 count) and $\mathbf{C}_{ij}$ other accurately measured covariates (e.g. rainfall location or patients' gender).

The random effects cumulative probability model (33) and the random effects generalized logistic model (25) can be written in a unified form

$$g(p_{ij,k}) = \alpha_k + \mathbf{X}'_{ij}\boldsymbol{\beta}_{x,k} + \mathbf{C}'_{ij}\boldsymbol{\beta}_{c,k} + b_i, \tag{35}$$

where b_i follows some distribution $F(.,\theta)$ with mean 0 and variance θ. For the random effects cumulative probability model (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 (25), $p_{ij,k} = P_{ij,k}/P_{ij,N}$ and $g(\cdot) = \log(\cdot)$.

Suppose the observed covariates $\mathbf{W}_{ij}$ (e.g., the radar measurements of rainfall or the observed CD4 counts) measure $\mathbf{X}_{ij}$ (e.g. the true precipitation amount or the true CD4 counts) with error. It is customary to postulate a non-differential additive measurement error model for $\mathbf{W}_{ij}$ (Carroll, et al., 1995),

$$\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij}, \tag{36}$$

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 $\hat{\boldsymbol{\Sigma}}_u$, e.g., using replicates or validation data. We are interested in testing for no within-cluster correlation $H_0 : \theta = 0$ in the random effects measurement error models (35) and (36). Li and Lin (2003) 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 (Cook and Stefanski, 1994), 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 (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 in two steps: the simulation step and the extrapolation step. In the simulation step, one generates data $\mathbf{W}^*_{ij}$ by adding to $\mathbf{W}_{ij}$ a random error following $N(0, \eta\boldsymbol{\Sigma}_u)$ for some constant $\eta > 0$. One then calculates naive parameter estimates by fitting

(35) with $\mathbf{X}_{ij}$ replaced by $\mathbf{W}^*_{ij}$. This would give the naive estimates if the measurement error covariance is $(1+\eta)\boldsymbol{\Sigma}_u$. This procedure is repeated for a large number B times (e.g., $B = 100$), and the means of the resulting B naive parameter estimates is calculated. One does this for a series of values of η (e.g, $\eta = 0.5, 1, 1.5, 2$). In the extrapolation step, a regression (e.g. quadratic) model is fit 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 of the model parameters. For details of the SIMEX method, see Cook and Stefanski (1994) and Carroll, et al., (1995). The SIMEX idea can be utilized to construct score tests for $H_0 : \theta = 0$ in the random effects measurement error models (35) and (36) by extending the results in Sections 5.1 and 5.2. The resulting SIMEX score tests are an extension of the work of Lin and Carroll (1999) to random effects measurement error models for clustered polytomous.

In the absence of measurement error, the score statistics for testing $H_0 : \theta = 0$ under (35) take the same form $U_\theta(\hat{\boldsymbol{\beta}}) \Big/ \widetilde{I}^{1/2}_{\theta\theta}(\hat{\boldsymbol{\beta}})$, where $U_\theta(\hat{\boldsymbol{\beta}})$ is given in (34) for random effects cumulative probability models and in (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 (Section 4.3.5 of Carroll, et al., 1995) to calculate the variance of this "estimator". Specifically, in the SIMEX simulation step, one simply calculates naive score statistics using the score formulae (34) and (28) by replacing $\mathbf{X}_{ij}$ with the simulated data $\mathbf{W}^*_{ij}$. The rest of the steps parallel those in the standard SIMEX method for parameter estimation. Denoting the results by $U_{simex}(\cdot)$ and $\widetilde{I}_{\theta\theta,simex}$ respectively, the SIMEX score statistic is simply

$$S_{simex} = U_{simex} / \widetilde{I}^{1/2}_{\theta\theta,simex}, \tag{37}$$

which follows $N(0,1)$ asymptotically when the true extrapolation function is used. Since the true extrapolation function is unknown in practice, an approximation (e.g., quadratic) is used. The simulation study reported by Li and Lin (2003) shows that the SIMEX score tests perform well. The theoretical justification of the SIMEX score tests can be found in

Lin and Carroll (1999).

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 $\mathbf{X}$.

5.4 Data Examples

To illustrate the variance component score tests for clustered polytomous data, we examine data from a longitudinal study on efficacy of steam inhalation for treating common cold symptoms conducted by Macknin, et al., (1990). This study included 30 patients with colds of recent onset. At the time of enrollment, each patient went through two 20-minute 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 4 ordered categories (no symptom, mild, moderate and severe symptom). One was interested in examining whether the severity improved following the treatment, and testing whether the observations over time for within each subject were likely to be correlated.

Li and Lin (2003) considered models (25) and (33) with the time from the baseline as a covariate. They first assumed a random effects logistic model (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 (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 time effect, they fitted (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 the 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 score test results.

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 as a random effect, if the scientists are interested in making inferences on the entire population of machines and randomly choose 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 is measuring 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).

Few difficulties arise from setting up the likelihood function for drawing inference based on a random effects model. The major obstacle lies in computation, as, for practitioners, the constant theme 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. But its computational burden is substantial and the convergence rate is often slow. Laplace approximation greatly reduces the computational load, but the resulting estimates are in general 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. But 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 a full use of the simulated data and obtain the estimates recursively. However, the detailed implementation of both methods have yet been finalized in the literature.

It is worth briefly discussing marginal models, another major tools for handling clustered data. In a marginal model, the marginal mean of the response vector is modeled as a function of explanatory variables (Zeger et al. 1995). Thus, as opposed to the random effect models, the coefficients in a marginal model have population average interpretations. This type of models are typically fitted via the so-called generalized estimating equation (GEE). An appealing feature is that, under the right mean structure, even when the covariance structure of the response is misspecified, 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 misspecified. 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 Zeger et al. (1995).

Last, we point out other active research areas in mixed modeling include evaluating goodness of fit of the model, choosing the best distribution for the random effects and selecting the best collection of covariates in a model. Readers are referred to some recent articles on these topics, e.g. Zheng (2000), Verbeke and Lesaffre (1996), Lindsey and Lindsey (2000) and Houseman et al. (2004).

REFERENCES

C. Bliss: The method of probits, *Science.* **79** (1934) 38-39

N.E. Breslow, D.G. Clayton: Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association.* **88** (1993) 9-25

B. P. Carlin, T. A. Louis: *Bayes and Empirical Bayes Methods for Data Analysis.* (Chapman and Hall, London, New York, 2000)

R. J. Carroll, D. Ruppert, L. A. Stefanski: *Measurement Error in Nonlinear Models.* (Chapman and Hall, London, 1995)

S. le Cessie, H.C. van Houwelingen: Testing the fit of a regression model via score tests in random effects models Biometrics **51** (1995) 600-614

C. G. Collier: *Applications of Weather Radar Systems: A Guide to Uses of Radar in Meteorology and Hydrology, 2nd ed.* (John Wiley, New York, 1996)

D. Commenges, L. Letenneur, H. Jacqmin, J. Moreau, J. Dartigues: (1994) Test of homogeneity of binary data with explanatory variables, *Biometrics.* **50** (1994) 613-20

J. R. Cook, L. A. Stefanski: Simulation-extrapolation estimation in parametric measurement error models, *Journal of the American Statistical Association.* **89** (1994) 1314-1328

N. A. Cressie: *Statistics for spatial data,* (John Wiley and Sons, New York, Chichester, 1991)

A.P. Dempster, N.M. Laird, D.B. Rubin: Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. B. **39** (1977) 1-22

P. J. Diggle, J. A. Tawn, R. A. Moyeed RA: Model-based geostatistics *J ROY STAT SOC C-AP.* **47** (1998) 299-326

A. E. Gelfand, B.P. Carlin: Maximum likelihood estimation for constrained- or missing-data problems, Canadian Journal of Statistics. **21**(1993) 303-311

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis.* (Chapman and Hall, London, 1995)

C. J. Geyer, E. A. Thompson: Constrained Monte Carlo maximization likelihood for dependent data, *Journal of the Royal Statistical Society.* B. **54** (1992) 657-699

R. Gray: Tests for variation over groups in survival data. *Journal of the American Statistical Association.* **90** (1995) 198-203

P. J. Green: Penalized likelihood for general semi-parametric regression models, *International Statistical Review.* **55** (1987) 245-259

D. A. Harville: Bayesian inference for variance components using only error contrasts, *Biometrika.* **61** (1974), 383-385

D. A. Harville, R. W. Mee: A mixed-model procedure for analyzing ordered categorical data, *Biometrics.* **40** (1984) 393-408

W. Hastings: Monte carlo sampling methods using markov chains and their applications, *Biometrika.* **57** (1970) 97-109

D. Hedeker, R. Gibbons: A random-effects ordinal regression model for multilevel analysis, *Biometrics.* **50** (1994) 933-945

C. R. Henderson, O. Kempthorne, S. R. Searle, C. N. von Krosigk: Estimation of environmental and genetic trends from records subject to culling, *Biometrics.* **15** (1959) 192-218

E. A. Houseman, L.M. Ryan, B. A. Coull: Cholesky residuals for assessing normal errors in a linear model with correlated outcomes, *JASA.* **99** (2004) 383-394

N. M. Laird, J. H. Ware: Random-effects models for longitudinal data, *Biometrics.* **38** (1982) 963-974

Y. Lee, J. A. Nelder: Hierarchical Generalized Linear Models, *J.R. Statist. Soc.* B. **58** (1996) 619-678

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

Y. Li, X. Lin: Testing random effects in uncensored/Censored clustered Data with categorical responses, *Biometrics.* **59** (2003) 25-35

X. Lin, N. E. Breslow: Bias Correction in generalized linear mixed models with multiple components of dispersion, *Journal of the American Statistical Association.* **91** (1995) 1007-1016

X. Lin: Variance component testing in generalized linear models with random effects, *Biometrika.* **84** (1997) 309-326

X. Lin, R. J. Carroll: SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics.* **55** (1999) 613-619

P. J. Lindsey, J. K. Lindsey: Diagnostic tools for random effects in the repeated measures growth curve model, *Computational Statistics and Data Analysis.* **33** (2000) 79-100

Q. Liu, D. A. Pierce: Heterogeneity in Mantel-Haenszel-type models *Biometrika.* **80** (1993) 543-556

M. L. Macknin, S. Mathew, S. V. Medendorp: Effect of inhaling heated vapor on symptoms of the common cold, *J. Am. Med. Assoc.* **264** (1990) 989-991

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

C. E. McCulloch: Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association.* **92** (1997) 162-170

C. E. McCulloch, S. R. Searle: *Generalized, linear, and mixed models,* (John Wiley and Sons, New York, Chichester, 2001)

C. A. McGilchrist: REML estimation for survival models with frailty, *Biometrics.* **49** (1993) 221-225

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth: Equation of state calculations by fast computing machines, *J. CHEM. PHYS.* **21** (1953) 1087-1092

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

C. P. Robert, G. Casella:*Monte Carlo statistical methods.* (Springer-Verlag Inc, Berlin, New York, 1999)

R. Schall: Estimation in generalized linear models with random effects, *Biometrika.* **78** (1991) 719-727

G. Satten: Rank-based Inference in the Proportional Hazards Model for Interval Censored Data, *Biometrika.* **83** (1996) 355-370

G. Satten, S. Datta: The S-U Algorithm for Missing Data Problems, *Computational Statistics.* **15** (2000) 243-277

S. G. Self, K. Y. Liang: Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association.* **82**(1987), 605-610

P. J. Solomon, D. R. Cox : Nonlinear component of variance models, *Biometrika.* **79** (1992) 1-11

R. Stiratelli, N. M. Laird, J. H. Ware: Random-effects models for serial observations with binary response, *Biometrics.* **40** (1984) 961-971

M. A. Tanner, W. H. Wong: The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association.* **82** (1987) 528-549

A. A. Tsiatis, V. Degruttola, M. S. Wulfsohn: Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS, *Journal of the American Statistical Association.* **90** (1995) 27-37

G. Verbeke, E. Lesaffre: A linear mixed-effects model with heterogeneity in the random-effects population, *Journal of the American Statistical Association,* **91** (1996) 217-221

S. L. Zeger, K. Y. Liang, P. S. Albert: Models for longitudinal data: a generalized estimating equation approach, *Biometrics.* **44** (1988) 1049-1060

S. L. Zeger, M. R. Karim: Generalized linear model with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association.* **86** (1991) 79-86

B. Zheng (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data, *Statistics in Medicine.* **19** (2000) 1265-1275