Chapter 7

Maximum likelihood for GLMMs

7.1 Introduction

As noted in Chapter 1, creation of a GLMM by incorporating random factors in the linear predictor of a GLM leads to difficult-to-handle likelihoods. This is first laid out more carefully in a simple example and then general approaches to maximum likelihood are described.

7.2 A simple example

To fix ideas consider the following logit-normal example:

$$
\begin{aligned}
Y_{ij}|\mathbf{u} &\sim \text{indep. Bernoulli}(p_{ij}), \qquad i = 1, 2, \ldots, q;\ j = 1, 2, \ldots, n, \\
\text{logit}(p_{ij}) &= \beta x_{ij} + u_i, \\
u_i &\sim \text{indep. } \mathcal{N}(0, \sigma^2).
\end{aligned} \tag{7.1}
$$

In this scenario there are q clusters, each with n observations, a logit link and a single random and single fixed factor. The random effects, u_i, are assumed to be i.i.d. normally distributed.

The example is so simplified it is a stretch to come up with a realistic situation it might reflect, but here is an attempt. Suppose we record $Y_{ij} = 1$ if a subject's blood pressure decreases on day j of treatment with a blood pressure drug at dose x_{ij}, and is 0 otherwise. There are q individuals tested, each at n different doses. Since the intercept is zero, when the dose is 0 and for u_i equal to its mean of zero, the probability of a decrease is 0.5. The interpretation of u_i is the person-specific propensity to decrease or increase blood pressure in response to treatment (a type of individual-specific placebo effect).

Since the model is specified conditionally, it is easiest to derive the likelihood

through a conditioning argument as follows:

$$
\begin{aligned}
\text{Likelihood} &= \Pr\{\mathbf{Y}=\mathbf{y}|\beta,\sigma^2\} \\
&= \int \Pr\{\mathbf{Y}=\mathbf{y}|\beta,\sigma^2,\mathbf{u}\}f(\mathbf{u}|\sigma^2)d\mathbf{u} \\
&= \int \Pr\{\mathbf{Y}=\mathbf{y}|\beta,\mathbf{u}\}f(\mathbf{u}|\sigma^2)d\mathbf{u} \\
&= \int \prod_{i,j} \Pr\{Y_{ij}=y_{ij}|\beta,\mathbf{u}\}f(\mathbf{u}|\sigma^2)d\mathbf{u} \\
&= \prod_i \int \prod_j \Pr\{Y_{ij}=y_{ij}|\beta,u_i\}f(u_i|\sigma^2)du_i \\
&= \prod_i \int e^{\beta\sum_j y_{ij}x_{ij}+y_{i\cdot}u_i} \prod_j (1+e^{\beta x_{ij}+u_i})^{-1} \\
&\qquad \times \exp\{-u_i^2/2\sigma^2\}/(2\pi\sigma^2)^{1/2}du_i.
\end{aligned}
\tag{7.2}
$$

There are several noteworthy features of the above calculations. First, the product appears in the fourth line because of the assumed conditional independence of the Y_{ij} given the random effects. Second, the product over the index i moves to the outside of the integration because the data form independent clusters (data within a cluster are dependent, but between clusters are independent). Finally, this integral cannot be evaluated in closed form, even though it about the simplest logit-normal model possible. On the other hand, it would not be too hard to evaluate this likelihood numerically since it is the product of single-dimensional integrals.

When the model has a single random effect or two nested random effects, it is relatively easy to evaluate the integrals in the likelihood. One can then maximize the likelihood numerically to find ML estimates and to perform likelihood ratio tests. This is the approach adopted in software such as SAS Proc NLMIXED (SAS Institute, 2001).

a. Numerical evaluation of the likelihood

When there is a single, normally distributed random effect, the likelihood can be written as a product of integrals of the form

$$
\int_{-\infty}^{+\infty} g(x)\exp\{-x^2\}dx. \tag{7.3}
$$

These can often be accurately evaluated using Gauss-Hermite quadrature, which addresses the usually troublesome appearance of an infinite range of integration:

$$
\int_{-\infty}^{+\infty} g(x)\exp\{-x^2\}dx \approx \sum_i w_i g(x_i), \tag{7.4}
$$

where the weights w_i and the evaluation points x_i can be found in books with details on numerical integration (e.g., Abramowitz and Stegun, 1964). This is a numerically simple approximation and is quite fast to compute, making numerical likelihood methods feasible.

As with any "automatic" numerical integration method, there are situations in which Gauss-Hermite quadrature for models like the logit-normal will give inaccurate results, generally having to do with the placement of the evaluation points. An improvement on simple Gauss-Hermite quadrature is adaptive quadrature, as exemplified in SAS Proc NLMIXED (SAS Institute, 2001) and Rabe-Hesketh et al. (2002), in which the point of evaluation of the integral is "centered" in order to improve accuracy.

While this approach works in simple problems it is not satisfactory in more difficult problems. The main complication is the design with regards to the random effects, since this affects which data are modeled as correlated. A particularly troublesome situation is when there are crossed random effects; in that case the data do not break into independent clusters, as opposed to the simple situation in (7.2). The leaf blight example has crossed random factors and evaluation of the likelihood would require the numerical evaluation of integrals of dimension greater than 200. Handling such situations is the topic of the remainder of the chapter.

7.3 Simulation approaches to ML

Since direct numerical evaluation of the likelihood is infeasible for many GLMMs, alternate approaches must be explored for approximating or calculating the likelihood (and then maximizing it). Many of the techniques have a genesis in Bayesian computational methods.

a. Model and notation

First I recall the notation for our GLMM from Chapter 4. Let $\mathbf{Y}$ denote the observed data vector and we will hypothesize the existence of a vector of random effects $\mathbf{u}$. We assume that the conditional distribution of $\mathbf{Y}$ given $\mathbf{u}$ follows a generalized linear model, with linear predictor, η_i, of the form $\eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}$, where $\mathbf{x}_i'$ denotes the ith row of $\mathbf{X}$, the model matrix for the fixed effects, and likewise with $\mathbf{z}_i'$ being the ith row of the model matrix for the random effects:

$$\begin{aligned} Y_i|\mathbf{u} &\sim \text{indep. } f_{Y_i|\mathbf{u}}(y_i|\mathbf{u},\boldsymbol{\beta},\phi), \\ \mathrm{E}[Y_i|\mathbf{u}] &= \mu_i, \\ g(\mu_i) &= \eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}, \\ \mathbf{u} &\sim f_{\mathbf{U}}(\mathbf{u}|\mathbf{D}). \end{aligned} \tag{7.5}$$

Note that we are assuming that the parameters of the conditional distribution of $\mathbf{Y}$ given $\mathbf{u}$ and those of $\mathbf{u}$ are distinct.

The likelihood for (7.5) is given by

$$L(\boldsymbol{\beta},\phi,\mathbf{D}) = \int \prod_{i=1}^{n} f_{Y_i|\mathbf{u}}(y_i|\mathbf{u},\boldsymbol{\beta},\phi) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D}) d\mathbf{u}, \tag{7.6}$$

which cannot usually be evaluated in closed form.

b. Monte Carlo EM

A possible approach to dealing with the high-dimensional integration is to set up an EM algorithm to compute the maximum likelihood estimates. To do so we need to define what will be the "missing data." A typical assumption in linear mixed models (see, e.g., Searle et al., 1992) is to consider the random effects, $\mathbf{u}$, to be the missing data. The complete data, $\mathbf{W}$, is then $\mathbf{W}=(\mathbf{Y},\mathbf{u})$ and the complete data loglikelihood is given by

$$\ln L_{\mathbf{W}} = \sum_i \ln f_{Y_i|\mathbf{u}}(y_i|\mathbf{u},\boldsymbol{\beta},\phi) + \ln f_{\mathbf{U}}(\mathbf{u}|\mathbf{D}). \tag{7.7}$$

This choice of missing data has two advantages. First, upon knowing $\mathbf{u}$, the Y_i are independent. Second, the M step of the EM algorithm maximizes with respect to $\boldsymbol{\beta}$, ϕ and $\mathbf{D}$. Since $\boldsymbol{\beta}$ and ϕ only enter the first term, the M step with respect to $\boldsymbol{\beta}$ and ϕ uses only the generalized linear model portion of the likelihood and so it is similar to a standard generalized linear model computation with the values of $\mathbf{u}$ treated as known. Maximizing with respect to $\mathbf{D}$ is just ML using the distribution of $\mathbf{u}$ after replacing sufficient statistics (in the case where $f_{\mathbf{U}}$ is in the exponential family) with their conditional expected values. The EM algorithm then takes the following form (where a superscript indicates the round of iteration):

1. Choose starting values $\boldsymbol{\beta}^{(0)}, \phi^{(0)}$, and $\mathbf{D}^{(0)}$. Set $m = 0$.
2. Calculate (with expectations evaluated under the current values).
   (a) $\boldsymbol{\beta}^{(m+1)}$ and $\phi^{(m+1)}$ to maximize $\mathrm{E}[\ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u},\boldsymbol{\beta},\phi)|\mathbf{y}]$.
   (b) $\mathbf{D}^{(m+1)}$ to maximize $\mathrm{E}[\ln f_{\mathbf{u}}(\mathbf{u}|\mathbf{D})|\mathbf{y}]$.
3. If convergence is achieved, declare the current values to be the MLEs; otherwise increment $m = m + 1$ and return to step 2.

In general, neither the expectation in 2(a) nor that in 2(b) can be computed in closed form. This is because the conditional distribution of $\mathbf{u}|\mathbf{y}$ involves $f_{\mathbf{Y}}$, that is, the likelihood, which is the quantity we are trying to avoid calculating directly.

It *is* possible, however, to produce random draws from the conditional distribution of $\mathbf{u}|\mathbf{y}$, without specifying or calculating $f_{\mathbf{Y}}$. One can then form Monte Carlo approximations to the required expectations.

There are a number of ways to produce the samples, including a Gibbs sampler (McCulloch, 1994), the Metropolis–Hastings algorithm (McCulloch, 1997) or the independence sampler (Booth and Hobert, 1999). For example, to specify a Metropolis algorithm, we must specify a candidate distribution, $h_{\mathbf{U}}(\mathbf{u})$, from which potential new values are drawn and also an acceptance function which gives the probability of accepting the new value (as opposed to keeping the previous value). This acceptance function is given by

$$A_k(\mathbf{u},\mathbf{u}^*) = \min\left\{1, \frac{f_{\mathbf{u}|\mathbf{Y}}(\mathbf{u}^*|\mathbf{y},\boldsymbol{\beta},\phi)h_{\mathbf{u}}(\mathbf{u})}{f_{\mathbf{u}|\mathbf{Y}}(\mathbf{u}|\mathbf{y},\boldsymbol{\beta},\phi)h_{\mathbf{u}}(\mathbf{u}^*)}\right\}, \tag{7.8}$$

where $\mathbf{u}^* = (u_1, u_2, \ldots, u_{k-1}, u_k^*, u_{k+1}, \ldots, u_q)'$ and A_k defines the probability of accepting the new coordinate u_k^* and replacing u_k with it.

What should be used for the candidate distibution, $h_{\mathbf{u}}(\mathbf{u}^*)$? One choice is to use $h_{\mathbf{u}} = f_{\mathbf{u}}$, in which case we get a simplification:

$$\begin{aligned}
&\frac{f_{\mathbf{u}|\mathbf{Y}}(\mathbf{u}^*|\mathbf{y},\boldsymbol{\beta},\phi)h_{\mathbf{U}}(\mathbf{u})}{f_{\mathbf{u}|\mathbf{Y}}(\mathbf{u}|\mathbf{y},\boldsymbol{\beta},\phi)h_{\mathbf{U}}(\mathbf{u}^*)} \\
&\quad = \frac{\prod_{i=1}^n f_{y_i|\mathbf{u}}(y_i|\mathbf{u},\boldsymbol{\beta},\phi)f_{\mathbf{u}}(\mathbf{u}^*|\mathbf{D})f_{\mathbf{u}}(\mathbf{u}|\mathbf{D})}{\prod_{i=1}^n f_{y_i|\mathbf{u}}(y_i|\mathbf{u}^*,\boldsymbol{\beta},\phi)f_{\mathbf{u}}(\mathbf{u}|\mathbf{D})f_{\mathbf{u}}(\mathbf{u}^*|\mathbf{D})} \\
&\quad = \frac{\prod_{i=1}^n f_{y_i|\mathbf{u}}(y_i|\mathbf{u},\boldsymbol{\beta},\phi)}{\prod_{i=1}^n f_{y_i|\mathbf{u}}(y_i|\mathbf{u}^*,\boldsymbol{\beta},\phi)}.
\end{aligned} \tag{7.9}$$

This calculation only involves the specification of the generalized linear model portion of the model, namely the conditional distribution of $\mathbf{Y}|\mathbf{u}$.

Incorporating the Metropolis step into the EM algorithm gives an algorithm as follows:

1. Choose starting values $\boldsymbol{\beta}^{(0)}, \phi^{(0)}$, and $\mathbf{D}^{(0)}$. Set $m = 0$.
2. Generate N values, $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(N)}$, from the conditional distribution of $\mathbf{u}|\mathbf{Y}$ using a Metropolis algorithm like the one described above and using the current parameter values.
3. Choose
   (a) $\boldsymbol{\beta}^{(m+1)}$ and $\phi^{(m+1)}$ to maximize a Monte Carlo estimate of $\mathrm{E}[\ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u},\boldsymbol{\beta},\phi)|\mathbf{y}]$, namely $\frac{1}{N}\sum_{k=1}^N \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}^{(k)},\boldsymbol{\beta},\phi)$.
   (b) $\mathbf{D}^{(m+1)}$ to maximize $\frac{1}{N}\sum_{i=1}^N \ln f_{\mathbf{u}}(\mathbf{u}^{(k)}|\mathbf{D})$.
4. If convergence is achieved, declare the current values to be the MLEs; otherwise increment $m = m + 1$ and return to step 2.

While computationally intensive, this approach remains feasible for a variety of data configurations.

c. Monte Carlo Newton–Raphson

Although the EM algorithm is stable, in the sense that it is often able to converge from a wide variety of starting values, it is also well known to require many iterations to converge in a variety of problems. Algorithms that are quadratically convergent are often preferred. I next consider a version of Monte Carlo ML using a algorithm more akin to Newton–Raphson.

Whenever the marginal density of $\mathbf{Y}$ is formed with distinct parameters for $f_{\mathbf{Y}|\mathbf{u}}$ and $f_{\mathbf{u}}$ then the ML equations for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi)'$ and $\mathbf{D}$ take the following form:

$$\mathrm{E}\left[\frac{\partial \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\middle|\mathbf{y}\right] = 0, \tag{7.10}$$

$$\mathrm{E}\left[\frac{\partial \ln f_{\mathbf{u}}(\mathbf{U}|\mathbf{D})}{\partial \mathbf{D}}\middle|\mathbf{y}\right] = 0, \tag{7.11}$$

where I have momentarily used the standard notation of $\mathbf{U}$ to denote a random variable, to more clearly indicate that the expectation is with respect to the distribution of the random effects, $\mathbf{U}$, conditional on $\mathbf{Y}$. Equality (7.11) only involves the distribution of $\mathbf{U}$ and is often fairly easy to solve, for example, when the distribution is normal. On the other hand, solving the first equality is similar to a standard generalized linear model and is amenable to a Newton–Raphson or scoring approach, which I now develop.

Expanding $\partial \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\theta})/\partial\boldsymbol{\beta}$ as a function of $\boldsymbol{\beta}$ around the value $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \phi_0)'$ gives

$$\begin{aligned}&\frac{\partial \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\theta})}{\partial\boldsymbol{\beta}}\\ &\doteq \frac{\partial \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\theta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{\partial^2 \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\boldsymbol{\beta}-\boldsymbol{\beta}_0).\end{aligned} \tag{7.12}$$

Specializing this to our model, and after utilizing the fact that one term has a conditional expected value of $\mathbf{0}$, the approximation becomes

$$\begin{aligned}&\frac{\partial \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\theta})}{\partial\boldsymbol{\beta}}\\ &\quad\doteq \mathbf{X}'\mathbf{W}(\boldsymbol{\theta}_0,\mathbf{U})/a(\phi_0)\,\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\bigg|_{\theta=\theta_0}[\mathbf{Y}-\boldsymbol{\mu}(\boldsymbol{\theta}_0,\mathbf{U})]\\ &\qquad-\mathbf{X}'\mathbf{W}(\boldsymbol{\theta}_0,\mathbf{U})/a(\phi_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0),\end{aligned} \tag{7.13}$$

where $\mathbf{W}(\boldsymbol{\theta},\mathbf{U}) = \operatorname{diag}\{(\partial\eta_i/\partial\mu_i)^2\operatorname{var}(Y_i|\boldsymbol{\mu})\}$, $\mu_i(\boldsymbol{\theta},\mathbf{u}) = \mathrm{E}[Y_i|\mathbf{u}]$, and $\partial\boldsymbol{\eta}/\partial\boldsymbol{\mu} = \operatorname{diag}\{\partial\eta_i/\partial\boldsymbol{\mu}_i\}$.

Using this approximation in (7.10) leads to an iteration equation of

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} &+ \mathrm{E}\left[\mathbf{X}'\mathbf{W}\left(\boldsymbol{\theta}^{(m)},\mathbf{U}\right)\mathbf{X}|\mathbf{y}\right]^{-1}\\ &\times \mathrm{E}\left[\mathbf{X}'\mathbf{W}\left(\boldsymbol{\theta}^{(m)},\mathbf{U}\right)\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\bigg|_{\theta=\theta^{(m)}}\right.\\ &\left.\times\left\{\mathbf{y}-\boldsymbol{\mu}\left(\boldsymbol{\beta}^{(m)},\mathbf{U}\right)\right\}\Big|\,y\right].\end{aligned} \tag{7.14}$$

This analog of scoring would proceed by iteratively solving (7.11), (7.14) and an equation for ϕ. An advantage of the scoring approach over MCEM is that it makes automatic the maximization step in 2(a).

Again, the expectations cannot typically be evaluated in closed form which leads to a Monte Carlo Newton–Raphson (MCNR) approach. The 3(b) step in MCEM would be replaced by

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} &+ \tilde{\mathrm{E}}\left[\mathbf{X}'\mathbf{W}\left(\boldsymbol{\theta}^{(m)},\mathbf{U}\right)\mathbf{X}|\mathbf{y}\right]^{-1}\\ &\times \tilde{\mathrm{E}}\left[\mathbf{X}'\mathbf{W}\left(\boldsymbol{\theta}^{(m)},\mathbf{U}\right)\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\bigg|_{\theta=\theta^{(m)}}\right.\\ &\left.\times\{\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(m)},\mathbf{U})\}\Big|\,y\right],\end{aligned} \tag{7.15}$$

where $\tilde{\mathrm{E}}$ represents a Monte Carlo approximant to the expectation.

d. Simulated maximum likelihood and moments

While both MCEM and MCNR work on the log of the likelihood, Geyer and Thompson (1992), Gelfand and Carlin (1993) and Durbin and Koopman (1997) have suggested simulation to estimate the value of the likelihood directly. Starting from the likelihood we have

$$
\begin{aligned}
L(\boldsymbol{\beta}, \phi, \mathbf{D}|\mathbf{y}) &= \int f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \phi) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D}) d\mathbf{u} \\
&= \int \frac{f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \phi) f_{\mathbf{U}}(\mathbf{u}|\mathbf{D})}{h_{\mathbf{u}}(\mathbf{u})} h_{\mathbf{u}}(\mathbf{u}) d\mathbf{u} \qquad (7.16) \\
&\doteq \frac{1}{N} \sum_{k=1}^{N} \frac{f_{\mathbf{y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}^{(k)}, \boldsymbol{\beta}, \phi) f_{\mathbf{U}}(\mathbf{u}^{(k)}|\mathbf{D})}{h_{\mathbf{u}}(\mathbf{u}^{(k)})},
\end{aligned}
$$

where the $\mathbf{u}$'s are selected from the *importance sampling* distribution $h_{\mathbf{u}}(\mathbf{u})$ and N is the number of simulated values. This is an unbiased estimate no matter the choice of $h_{\mathbf{u}}(\mathbf{u})$. The simulated likelihood is then numerically maximized, either after a single simulation, or using multiple simulations in an iterative process where the importance sampling distribution is allowed to depend on the current parameter values. A distinction to this approach is that the simulation and the maximization take place in two separate steps, while the MCEM and MCNR methods directly approximate the terms needed to perform the maximization.

e. Stochastic approximation

Stochastic approximation is a well-researched method (e.g., Lai and Robbins, 1979; Wei, 1987; Ruppert, 1991) originally proposed for finding the root of a regression equation in cases where it was desired to avoid strong assumptions about the form of the regression. Starting from an initial guess of $x^{(0)}$, and under the assumption that $\mathrm{E}[Y|x]$ is increasing, new guesses are obtained via an equation of the form

$$
x^{(m+1)} = x^{(m)} - \alpha_m Y_m, \qquad (7.17)
$$

where Y_m is a value of Y sampled from the regression model with $x = x^{(m)}$, and with α_m decreasing slowly enough so that $\sum \alpha_m = \infty$, but quickly enough so that $\sum \alpha_m^2 < \infty$. The intuition behind (7.17) is that if the current guess as to the root gives a value of Y_m that is larger than zero, then the next value of x should be smaller than the current one. The values of α_m are chosen to be decreasing so that the sequence eventually converges to the root, without using step sizes so small that convergence takes too long.

How can this idea be applied to solving for the MLEs in GLMMs? The basic idea is to apply the root-finding to the ML equation, appropriately defined. Since

$$
f_{\mathbf{u}|\mathbf{Y}} = \frac{f_{\mathbf{Y},\mathbf{u}}}{f_{\mathbf{Y}}}
$$

we have

$$\ln f_{\mathbf{Y},\mathbf{u}} = \ln f_{\mathbf{Y}} + \ln f_{\mathbf{u}|\mathbf{Y}}$$

and

$$\frac{\partial \ln f_{\mathbf{Y},\mathbf{u}}}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f_{\mathbf{Y}}}{\partial \boldsymbol{\theta}} + \frac{\partial \ln f_{\mathbf{u}|\mathbf{Y}}}{\partial \boldsymbol{\theta}}. \tag{7.18}$$

Noting that the expectation (with respect to the conditional distribution of $\mathbf{u}$ given $\mathbf{Y}$) of the final term in (7.18) is zero (by the usual score identity) we can regard it as a regression equation with $\partial \ln f_{\mathbf{Y},\mathbf{u}}/\partial \boldsymbol{\theta}$ playing the role of the "response," $\partial \ln f_{\mathbf{Y}}/\partial \boldsymbol{\theta}$ playing the role of the regression equation (regarded as a function of the unknown parameter, $\boldsymbol{\theta}$), and $\partial \ln f_{\mathbf{u}|\mathbf{Y}}/\partial \boldsymbol{\theta}$ as a mean-zero "error" term. Now solving for the root of the "regression" equation is the same as solving for the value of $\boldsymbol{\theta}$ that makes $\partial \ln f_{\mathbf{Y}}/\partial \boldsymbol{\theta}$ equal to zero, that is, the ML estimate. See Delyon et al. (1999) for a recent example of such work.

7.4 Nonparametric maximum likelihood

Thus far, I have only considered parametric estimation of the random effects distribution. This is partially a personal bias, partly due to the fact that nonparametric estimation of random effects distributions is limited in the scope of problems to which it is applicable (in complicated problems with multiple random factors and covariates it will often give degenerate answers), and partially due to the feeling, backed up by some research, that the exact form assumed for a latent, unobserved variate is not terribly important. This is supported by, for example, Neuhaus et al. (1992) and Tan et al. (1999) and it is important to discount the results of Heckman and Singer (1984), which suggest great sensitivity to the assumed distribution but apply to situations in which there is a mixture distribution but *only one observation per level of the "random effect,"* a situation I consider unrealistic.

That said, there is considerable literature on nonparametric estimation of the random effects distribution. The nonparametric MLE takes the form of a discrete distribution on a finite number of support points (Lindsey, 1983). See Aitken (1999) for a nice recent review article. Various forms of the EM algorithm have been suggested for fitting these models; see Pilla and Lindsay (2001) for a recent example. Variations on complete nonparametric MLE fitting are the smooth nonparametric approaches of Magder and Zeger (1996) and Verbeke and Lesaffre (1996) and the "semi-nonparametric" approach of Zhang and Davidian (2001).

7.5 Bayesian methods

My approach throughout this monograph is unabashedly frequentist. In a sense, a Bayesian would never be concerned with a mixed model: she would consider all factors to be random. Of course, this makes the computations nearly as difficult as maximum likelihood. Modern computational methods for Bayesian analysis of generalized linear models are an active research area I won't attempt to summarize here. The state of the practice is exemplified in software packages like BUGS

(Spiegelhalter et al., 1999), which is available for free!

Conversely, many would not see much difference between a Bayesian approach and factors declared to be random. I would like to point out what I see as two major differences. First, I only incorporate random effects (and hence assign a distribution) for tangible entities, like subjects, sites and isolates. Conceptually, the effect for each of those exists (though we might not be able to gather sufficient data to know it very precisely) and the distribution across the various subjects or sites or isolates is also a tangible entity. This is in direct contrast to a Bayesian, who is required to specify a distribution for all parameters, for example, the overall intercept. Second, proper Bayesians assign distributions designed to capture belief. This is quite different than writing down a statistical model that describes the fact that, for exmaple, subjects have different baseline values in a model and the subjects in the study come from a larger collection of possible subjects.

Finally, I would like to comment on the feeling that Bayesian procedures with flat or diffuse priors will always mimic a maximum likelihood approach. This is not always true, since ML is possible in cases where a flat prior gives rise to an improper posterior (Natarajan and McCulloch, 1995). Of course, a proper prior will always give rise to a proper posterior. So what about using a proper, but diffuse prior? Natarajan and McCulloch (1998) show that, in some situations, there may be no compromise. That is, before the priors get diffuse enough to be essentially non-informative (in the sense that they mimic ML) the Bayesian computational machinery may break down.

7.6 Further notes

Other techniques for maximizing a simulated likelihood are explored in Borsch-Supan and Hajivassiliou (1993); Durbin and Koopman (1997) and Casella and Berger (1994); adaptive importance sampling is considered in Kong et al. (1994); Gelman and Meng (1998). Similar methods can be used to approximate either the likelihood ratio statistics directly or score tests (Kent, 1982; Boos, 1992; Geyer and Thompson, 1992; Geyer, 1994). Jank and Booth (to appear) investigate and compare some of the methods described in this chapter. Chan and Ledolter (1995) use Monte Carlo EM for a time-series model for count data.