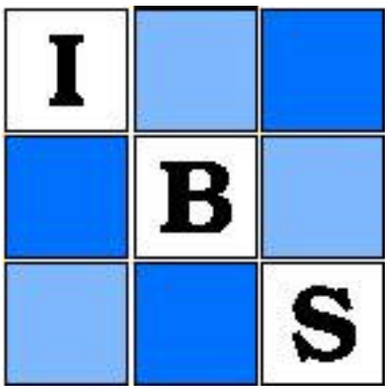

BIOMETRICS 46, 673–687
September 1990

Nonlinear Mixed Effects Models for Repeated Measures Data

Mary J. Lindstrom

Biostatistics Center, University of Wisconsin–Madison,
Madison, Wisconsin 53706, U.S.A.

and

Douglas M. Bates

Department of Statistics, University of Wisconsin–Madison,
Madison, Wisconsin 53706, U.S.A.

SUMMARY

We propose a general, nonlinear mixed effects model for repeated measures data and define estimators for its parameters. The proposed estimators are a natural combination of least squares estimators for nonlinear fixed effects models and maximum likelihood (or restricted maximum likelihood) estimators for linear mixed effects models. We implement Newton–Raphson estimation using previously developed computational methods for nonlinear fixed effects models and for linear mixed effects models. Two examples are presented and the connections between this work and recent work on generalized linear mixed effects models are discussed.

1. Introduction and Summary

1.1 *Introduction*

There has been a great deal of recent interest in mixed effects models for repeated measures data. By "repeated measures data" we mean data generated by observing a number of *individuals* repeatedly under differing experimental conditions where the individuals are assumed to constitute a random sample from a population of interest. For the reader familiar with split-plot designs, each individual can be thought of as a whole plot and each observation as a subplot. However, unlike split-plot models, repeated measures models usually include an underlying "functional" relationship between at least one of the predictor variables and the observations within individuals.

A common type of repeated measures data is *longitudinal data.* This is repeated measures data where the observations are ordered by time or position in space. More generally, longitudinal data can be defined as repeated measures data where the observations within individuals were not (or could not have been) randomly assigned to the levels of a "treatment" of interest (usually time or position in space). Serial correlation often results.

Mixed effects models for repeated measures data have become popular in part because their flexible covariance structure allows for nonconstant correlation among the observations and/or unbalanced data (designs that vary among individuals). Mixed effects models are also intuitively appealing. The notion that individuals' responses all follow a similar functional form with parameters that vary among individuals seems to be appropriate in many situations.

Key words: Longitudinal data; Newton–Raphson; Nonlinear least squares; Nonlinear models; Random effects.

The majority of work on methods for repeated measures data has focused on data that can be modeled by an expectation function that is linear in its parameters (e.g., Laird and Ware, 1982). More recent work has included the extension of generalized linear models to handle repeated measures data (e.g., Stiratelli, Laird, and Ware, 1984; Liang and Zeger, 1986; and Zeger, Liang, and Albert, 1988). Much of the recent work in mixed models (including GLIM models) has emphasized methods for data from epidemiological studies where a large number of individuals are observed at a small number of time points.

In this paper we are concerned with repeated measures data for which the assumption of normal errors is tenable (possibly after suitable transformation) but the proposed expectation function is nonlinear. One common example is nonlinear growth curve data. Data sets of this type typically include a moderate to large number of observations on each of a relatively small number of individuals. The proposed methods work equally well when many individuals are observed at a few time points each.

1.2 *Summary*

In Section 2 we present a general, nonlinear mixed effects model that is a generalization of both the linear mixed effects model and the standard fixed effects nonlinear model. We also discuss a number of nonlinear random effects models that have been proposed previously. The estimates we propose in Sections 3 and 4 for the variance components and the fixed effects are maximum likelihood (ML) or restricted maximum likelihood (RML) (Harville, 1974) for the linear mixed effects problem that in some sense most closely approximates the nonlinear problem. The estimates proposed for the random effects are posterior modes. These estimators reduce to the standard estimators for the linear mixed effects model when the expectation function is linear and to standard nonlinear least squares estimators when the variance of the random effects is zero.

The estimation method described in Section 5 is a two-step iterative procedure that draws on existing methods for linear mixed effects models and standard nonlinear least squares estimation. The calculations for both steps are implemented in a computationally efficient and stable manner.

In Section 6 we give details about the implementation and performance of the algorithm. We present two examples in Section 7 and in Section 8 we discuss relationships between this work and recent work on GLIM models for repeated measures data.

2. The Model

We define a general, nonlinear mixed effects model for the jth observation on the ith individual as

$$y_{ij} = f(\boldsymbol{\phi}_i, \mathbf{x}_{ij}) + e_{ij}, \tag{2.1}$$

where y_{ij} is the jth response on the ith individual, $\mathbf{x}_{ij}$ is the predictor vector for the jth response on the ith individual, f is a nonlinear function of the predictor vector and a parameter vector $\boldsymbol{\phi}_i$ of length r, and e_{ij} is a normally distributed noise term. There are no restrictions on the predictor vectors $\mathbf{x}_{ij}$. The parameter vector can vary from individual to individual. This is incorporated into the model by writing $\boldsymbol{\phi}_i$ as

$$\boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \quad \mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{D}), \tag{2.2}$$

where $\boldsymbol{\beta}$ is a p-vector of fixed population parameters, $\mathbf{b}_i$ is a q-vector of random effects associated with individual i, the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are design matrices of size $r \times p$ and $r \times q$ for the fixed and random effects, respectively, and $\sigma^2\mathbf{D}$ is a covariance matrix.

This is a general form of the nonlinear mixed effects model since any nonlinear function of fixed and random effects can be written as $f(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_{ij})$. The design matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are used to simplify model specification. For instance, if the data are in two groups then $\mathbf{A}_i$ can be chosen to allow different fixed effects for the different groups (e.g., $\mathbf{A}_i = [\mathbf{I}\ \mathbf{0}]$ for individuals in group 1 and $\mathbf{A}_i = [\mathbf{0}\ \mathbf{I}]$ for individuals in group 2). If the $\mathbf{B}_i$ are set equal for all i then both groups would still have random effects from the same distribution. Another common situation is where some but not all of the parameters have a random component. In this case $\mathbf{A}_i = \mathbf{I}$ but $\mathbf{B}_i$ would contain only some of the columns from $\mathbf{A}_i$ (see Section 7 for examples). Many other combinations are possible. An additional important reason for the use of these design matrices in the model is that they simplify model specification for computer implementation. Once the user supplies the derivatives of f with respect to $\boldsymbol{\phi}$, changes in the model that involve only $\mathbf{A}_i$ and $\mathbf{B}_i$ can be handled automatically.

We will often wish to write the model for the ith individual's entire response vector. This is accomplished by letting

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) = \begin{bmatrix} f(\boldsymbol{\phi}_i, \mathbf{x}_{i1}) \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{i2}) \\ \vdots \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{in_i}) \end{bmatrix}.$$

Then

$$\mathbf{y}_i = \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) + \mathbf{e}_i, \tag{2.3}$$

where $\mathbf{e}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i)$ and $\boldsymbol{\Lambda}_i$ is a matrix that depends on i only through its dimension. For convenience we suppress the dependence of $\boldsymbol{\eta}_i$ on the predictor vectors $\mathbf{x}_{ij}$ $(j = 1, \ldots, n_i)$. In many situations $\boldsymbol{\Lambda}_i = \mathbf{I}$ but its inclusion in the model allows for the specification of a nonindependent marginal covariance structure—for example, an AR(1) correlation. Usually the number of parameters in $\boldsymbol{\Lambda}_i$ would be quite small.

We can write the M individual models as one by letting

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}, \quad \boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_M \end{bmatrix}, \quad \boldsymbol{\eta}(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{\eta}_1(\boldsymbol{\phi}_1) \\ \boldsymbol{\eta}_2(\boldsymbol{\phi}_2) \\ \vdots \\ \boldsymbol{\eta}_M(\boldsymbol{\phi}_M) \end{bmatrix},$$

$\tilde{\mathbf{D}} = \mathrm{diag}(\mathbf{D}, \mathbf{D}, \ldots, \mathbf{D})$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_M)$. Now, the overall model is

$$\mathbf{y} \mid \mathbf{b} \sim \mathrm{N}(\boldsymbol{\eta}(\boldsymbol{\phi}), \sigma^2\boldsymbol{\Lambda}), \quad \boldsymbol{\phi} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}, \tag{2.4}$$

$$\mathbf{b} \sim \mathrm{N}(\mathbf{0}, \sigma^2\tilde{\mathbf{D}}), \tag{2.5}$$

where

$$\mathbf{B} = \mathrm{diag}(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M), \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_M \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_M \end{bmatrix}.$$

Berkey (1982), Johansen (1984), and Racine-Poon (1985) all propose estimators for a random effects model that is a special case of the one given above where $\mathbf{A}_i = \mathbf{B}_i = \mathbf{I}$. That is, they model the expected value of the ith individual's response vector as $\boldsymbol{\eta}(\boldsymbol{\phi}_i)$ where

$\phi_i \sim N(\phi, \mathbf{D})$. This restriction ties the fixed and random effects together and is less flexible for modeling grouped data. These authors cope with the difficulties of a nonlinear expectation function by first calculating the nonlinear least squares estimates for the ϕ_i assuming an independent marginal error structure and then making a one-time linear approximation to the expectation function at these estimates. This method is appealingly simple but will produce poor estimates if this initial approximation is inappropriate at the final estimates of the ϕ_i. An additional drawback is that individuals who do not have sufficient data for an initial nonlinear fit must be excluded from the analysis. Sheiner and Beal (1980) describe a similar nonlinear random effects model and give an estimation algorithm that updates the linear approximation to the nonlinear expectation function appropriately. However, they use a suboptimal approximation when estimating the marginal covariance structure (see Section 4). This is corrected in the EM algorithm proposed by Wolf (unpublished Ph.D. thesis, Department of Statistics, University of Wisconsin–Madison, 1986) for finding maximum likelihood estimates (see Section 4).

3. Estimation of β and b

When the variance components $\mathbf{\Lambda}$ and $\mathbf{D}$ are known and η is a linear function of β and $\mathbf{b}$ ($\eta(\mathbf{A}_i\beta + \mathbf{B}_i\mathbf{b}_i) = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i$), then, if we define $\mathbf{X} = [\mathbf{X}_1^T, \mathbf{X}_2^T, \ldots, \mathbf{X}_M^T]^T$ and $\mathbf{Z} = \text{diag}(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_M)$, the standard estimators for β and $\mathbf{b}$ are the generalized least squares estimator

$$\hat{\beta}_{\text{lin}} = \hat{\beta}_{\text{lin}}(\theta) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{3.1}$$

and the posterior mean $\hat{\mathbf{b}}_{\text{lin}} = \hat{\mathbf{b}}_{\text{lin}}(\theta) = \tilde{\mathbf{D}}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}_{\text{lin}}(\theta))$, where $\mathbf{V} = \mathbf{\Lambda} + \mathbf{Z}\tilde{\mathbf{D}}\mathbf{Z}^T$ and θ contains the unique elements of $\mathbf{D}$ and the parameters in $\mathbf{\Lambda}$. The parameter vector θ is a convenience that allows us to write all of the variance parameters (except σ) in one vector without specifying exactly how $\mathbf{\Lambda}$ is parameterized. The "hat" notation is used consistently throughout to denote an estimator that depends on θ. The argument will often be suppressed to simplify notation. Note that $\mathbf{D}$ and $\mathbf{\Lambda}$ depend on θ by definition.

The estimates $\hat{\beta}_{\text{lin}}$ and $\hat{\mathbf{b}}_{\text{lin}}$ jointly maximize the function

$$g_{\text{lin}}(\beta, \mathbf{b} \mid \mathbf{y}) = -\tfrac{1}{2}\sigma^{-2}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})^T\mathbf{\Lambda}^{-1}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}) - \tfrac{1}{2}\sigma^{-2}\mathbf{b}^T\tilde{\mathbf{D}}^{-1}\mathbf{b}, \tag{3.2}$$

which for fixed β is the logarithm of the posterior density of $\mathbf{b}$ (up to a constant) and for fixed $\mathbf{b}$ is the log-likelihood for β (up to a constant). The two terms in (3.2) are a sum of squares and a quadratic term in $\mathbf{b}$. By transforming the quadratic term in $\mathbf{b}$ to an equivalent sum of squares term, we can treat the optimization purely as a least squares problem. This is then easy to translate into the nonlinear setting.

We create this least squares problem by augmenting the data vector with "pseudo-data" as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta + \tilde{\mathbf{Z}}\mathbf{b} + \tilde{\mathbf{e}},$$

where

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{y} \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{X} \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{Z} \\ \tilde{\mathbf{D}}^{-1/2} \end{bmatrix}, \quad \tilde{\mathbf{e}} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

and where $\tilde{\mathbf{D}}^{-1/2} = \text{diag}(\mathbf{L}^{-T}, \mathbf{L}^{-T}, \ldots, \mathbf{L}^{-T})$, and $\mathbf{L}$ is the Cholesky factor of $\mathbf{D}$ ($\mathbf{D} = \mathbf{L}^T\mathbf{L}$ and $\mathbf{L}$ is upper-triangular). Similarly, $\mathbf{\Lambda}^{-1/2}$ contains the Cholesky factors of the $\mathbf{\Lambda}_i$.

In a nonlinear mixed effects model the maximum likelihood estimator $\hat{\beta}(\theta)$ and the posterior mode $\hat{\mathbf{b}}(\theta)$ maximize the function

$$g(\beta, \mathbf{b} \mid \mathbf{y}) = -\tfrac{1}{2}\sigma^{-2}(\mathbf{y} - \eta(\mathbf{A}\beta + \mathbf{B}\mathbf{b}))^T\mathbf{\Lambda}^{-1}(\mathbf{y} - \eta(\mathbf{A}\beta + \mathbf{B}\mathbf{b})) - \tfrac{1}{2}\sigma^{-2}\mathbf{b}^T\tilde{\mathbf{D}}^{-1}\mathbf{b}. \tag{3.3}$$

For fixed $\boldsymbol{\beta}$, g is a constant plus the log of the posterior density of $\mathbf{b}$. Thus it is clear that the $\mathbf{b}$ that maximizes g for a given value of $\boldsymbol{\beta}$ is the posterior mode. In Section 4 we show that $\hat{\boldsymbol{\beta}}$ is a maximum likelihood estimate relative to an approximate marginal distribution of $\mathbf{y}$. As in the linear case these estimates can be calculated as the solution to a (nonlinear) least squares problem formed by augmenting the data vector with "pseudo-data" as

$$\tilde{\mathbf{y}} = \tilde{\boldsymbol{\eta}}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}) + \tilde{\mathbf{e}},$$

where

$$\tilde{\mathbf{y}} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1/2}\mathbf{y} \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{e}} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}), \quad \text{and} \quad \tilde{\boldsymbol{\eta}}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}) = \begin{bmatrix} \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}) \\ \tilde{\mathbf{D}}^{-1/2}\mathbf{b} \end{bmatrix}.$$

4. Estimation of $\boldsymbol{\theta}$

4.1 *Maximum Likelihood*

We wish to define maximum likelihood estimators for $\boldsymbol{\theta}$ with respect to the marginal density of $\mathbf{y}$,

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \mathbf{b})p(\mathbf{b})\, d\mathbf{b},$$

but because the expectation function $\boldsymbol{\eta}$ is nonlinear in $\mathbf{b}$, there is no closed-form expression for this density and the calculation of such estimates would be very difficult. Instead, we approximate the conditional distribution of $\mathbf{y}$ [equation (2.4)] for $\mathbf{b}$ near $\hat{\mathbf{b}}(\boldsymbol{\theta})$ by a multivariate normal with expectation that is linear in $\mathbf{b}$. To accomplish this we approximate the residual $\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b})$ near $\hat{\mathbf{b}}$ as

$$\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}) \approx \mathbf{y} - [\boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\hat{\mathbf{b}}) + \hat{\mathbf{Z}}\mathbf{b} - \hat{\mathbf{Z}}\hat{\mathbf{b}}],$$

where

$$\hat{\mathbf{Z}}_i = \hat{\mathbf{Z}}_i(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\eta}_i}{\partial \mathbf{b}_i^{\mathrm{T}}}\bigg|_{\hat{\boldsymbol{\beta}},\hat{\mathbf{b}}} = \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\phi}^{\mathrm{T}}}\bigg|_{\hat{\boldsymbol{\beta}},\hat{\mathbf{b}}}\right)\mathbf{B}_i,$$

and

$$\hat{\mathbf{Z}} = \hat{\mathbf{Z}}(\boldsymbol{\theta}) = \mathrm{diag}(\hat{\mathbf{Z}}_1, \hat{\mathbf{Z}}_2, \ldots, \hat{\mathbf{Z}}_M) = \frac{\partial \boldsymbol{\eta}}{\partial \mathbf{b}^{\mathrm{T}}}\bigg|_{\hat{\boldsymbol{\beta}},\hat{\mathbf{b}}}. \tag{4.1}$$

Note that $\hat{\mathbf{Z}}$ is a function of $\boldsymbol{\theta}$ because $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ are (see §3). Then

$$\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\hat{\mathbf{b}}) + \hat{\mathbf{Z}}\hat{\mathbf{b}} - \hat{\mathbf{Z}}\mathbf{b} \mid \mathbf{b} \dot{\sim} \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda})$$

and the approximate conditional distribution of $\mathbf{y}$ is

$$\mathbf{y} \mid \mathbf{b} \dot{\sim} \mathrm{N}(\boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\hat{\mathbf{b}}) - \hat{\mathbf{Z}}\hat{\mathbf{b}} + \hat{\mathbf{Z}}\mathbf{b}, \sigma^2\boldsymbol{\Lambda}).$$

This expression, along with the distribution of $\mathbf{b}$ [equation (2.5)], allows us to approximate the marginal distribution of $\mathbf{y}$ as

$$\mathbf{y} \dot{\sim} \mathrm{N}(\boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\hat{\mathbf{b}}) - \hat{\mathbf{Z}}\hat{\mathbf{b}}, \sigma^2\hat{\mathbf{V}}), \tag{4.2}$$

where $\hat{\mathbf{V}} = \hat{\mathbf{V}}(\boldsymbol{\theta}) = \boldsymbol{\Lambda} + \hat{\mathbf{Z}}\tilde{\mathbf{D}}\hat{\mathbf{Z}}^{\mathrm{T}}$.

This method of approximation has been used in a similar setting by Stiratelli et al. (1984) and authors referenced therein. The algorithm of Sheiner and Beal (1980) uses a similar approximation but evaluated at the expectation of the random effects (at $\mathbf{b} = \mathbf{0}$ in our

model) rather than at the current estimates. This simplification reduces the computational burden if a Newton–Raphson-type algorithm is used but may result in poor estimates. The more accurate method of approximating the marginal distribution of $\mathbf{y}$ at the current estimates of the random effects is discussed by Wolf (unpublished thesis previously cited) in a proposed application of the EM algorithm to ML estimation for a nonlinear random effects model.

The log-likelihood corresponding to the approximate marginal distribution in equation (4.2) is

$$l_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \mathbf{y}) = -\tfrac{1}{2} \log \mid \sigma^2 \hat{\mathbf{V}} \mid - \tfrac{1}{2}\sigma^{-2}[\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\hat{\mathbf{b}}) + \hat{\mathbf{Z}}\hat{\mathbf{b}}]^T \hat{\mathbf{V}}^{-1}[\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\hat{\mathbf{b}}) + \hat{\mathbf{Z}}\hat{\mathbf{b}}], \quad (4.3)$$

where $\hat{\mathbf{b}}$ and $\hat{\mathbf{Z}}$ depend on $\boldsymbol{\theta}$. We define $\boldsymbol{\beta}^{(ML)}$, $\sigma^{(ML)}$, and $\boldsymbol{\theta}^{(ML)}$ to be the maximum likelihood estimators for $\boldsymbol{\beta}$, σ, and $\boldsymbol{\theta}$ with respect to l_F.

Note that, as in the linear case, the two estimators for $\boldsymbol{\beta}$ are equivalent—that is, $\boldsymbol{\beta}^{(ML)} = \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(ML)})$. This follows from the fact that they both maximize $l_F(\boldsymbol{\beta}, \sigma^{(ML)}, \boldsymbol{\theta}^{(ML)} \mid \mathbf{y})$.

The inverse second derivative matrix of l_F provides an estimate for an approximate variance–covariance matrix for $\boldsymbol{\beta}^{(ML)}$, $\sigma^{(ML)}$, and $\boldsymbol{\theta}^{(ML)}$.

4.2 *Restricted Maximum Likelihood*

The method for defining the RML estimators is the same as that for the maximum likelihood estimators except that l_F in equation (4.3) becomes

$$l_R(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \mathbf{y}) = -\tfrac{1}{2} \log \mid \sigma^{-2} \hat{\mathbf{X}}^T \hat{\mathbf{V}}^{-1} \hat{\mathbf{X}} \mid + l_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \mathbf{y}), \quad (4.4)$$

where

$$\hat{\mathbf{X}}_i = \hat{\mathbf{X}}_i(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}^T} \Bigg|_{\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}} = \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\phi}^T} \Bigg|_{\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}} \right) \mathbf{A}_i$$

and

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}(\boldsymbol{\theta}) = \begin{bmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{X}}_2 \\ \vdots \\ \hat{\mathbf{X}}_M \end{bmatrix} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}^T} \Bigg|_{\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}}. \quad (4.5)$$

We define the estimators $\boldsymbol{\beta}^{(RML)}$, $\sigma^{(RML)}$, and $\boldsymbol{\theta}^{(RML)}$ to be those that maximize l_R. An argument parallel to the one given above will show that $\boldsymbol{\beta}^{(RML)} = \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(RML)})$. The restricted likelihood is based on $N - p$ linearly independent error contrasts. In the nonlinear model, the derivative matrix $\hat{\mathbf{X}}$, which defines these errors contrasts, depends on $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$. However, since the subspace spanned by the columns of this matrix depends only on intrinsic nonlinearity and not on parameter effects nonlinearity (Bates and Watts, 1980), it will be nearly constant near the estimates.

5. Two-Step Algorithm

We propose a two-step algorithm for finding $\boldsymbol{\theta}^{(ML)}$:

1. *Pseudo-data* (PD) *step.* Given the current estimate $\boldsymbol{\theta}^{(\omega)}$ for $\boldsymbol{\theta}$, calculate $\boldsymbol{\beta}^{(\omega)} = \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(\omega)})$, $\mathbf{b}^{(\omega)} = \hat{\mathbf{b}}(\boldsymbol{\theta}^{(\omega)})$, $\mathbf{X}^{(\omega)} = \hat{\mathbf{X}}(\boldsymbol{\theta}^{(\omega)})$, and $\mathbf{Z}^{(\omega)} = \hat{\mathbf{Z}}(\boldsymbol{\theta}^{(\omega)})$.

2. *Linear mixed effects* (LME) *step.* Given $\mathbf{b}^{(\omega)}$ and $\mathbf{Z}^{(\omega)}$, let $\boldsymbol{\beta}^{(\omega+1)}$, $\sigma^{(\omega+1)}$, and $\boldsymbol{\theta}^{(\omega+1)}$ be the values that maximize

$$l_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(\omega)}, \mathbf{y})$$
$$= -\tfrac{1}{2} \log \mid \sigma^2(\boldsymbol{\Lambda} + \mathbf{Z}^{(\omega)}\tilde{\mathbf{D}}\mathbf{Z}^{(\omega)\mathrm{T}}) \mid$$
$$- \tfrac{1}{2}\sigma^{-2}[\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}^{(\omega)}) + \mathbf{Z}^{(\omega)}\mathbf{b}^{(\omega)}]^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{Z}^{(\omega)}\tilde{\mathbf{D}}\mathbf{Z}^{(\omega)\mathrm{T}})^{-1}[\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}^{(\omega)}) + \mathbf{Z}^{(\omega)}\mathbf{b}^{(\omega)}].$$

We write the dependence of l_F on $\boldsymbol{\theta}^{(\omega)}$ explicitly to emphasize the dependence on the value of $\boldsymbol{\theta}$ at which $\hat{\mathbf{b}}$ and $\hat{\mathbf{Z}}$ are evaluated. The algorithm consists of iterating between these two steps until convergence. (We discuss the computational details in Section 6.)

What we have called the LME step does not quite correspond to a linear mixed effects estimation problem because the residual $\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}^{(\omega)}) + \mathbf{Z}^{(\omega)}\mathbf{b}^{(\omega)}$ is nonlinear in $\boldsymbol{\beta}$. We could still use the Newton–Raphson method to optimize l_F but this would require second derivatives of the model function, $\boldsymbol{\eta}$, with respect to the fixed effects, $\boldsymbol{\beta}$. To avoid this, we *approximate the residual near* $\boldsymbol{\beta}^{(\omega)}$ *by*

$$\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}^{(\omega)}) + \mathbf{Z}^{(\omega)}\mathbf{b}^{(\omega)} \approx \mathbf{y} - [\boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta}^{(\omega)} + \mathbf{B}\mathbf{b}^{(\omega)}) + \mathbf{X}^{(\omega)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(\omega)}) - \mathbf{Z}^{(\omega)}\mathbf{b}^{(\omega)}]$$
$$= \mathbf{w}^{(\omega)} - \mathbf{X}^{(\omega)}\boldsymbol{\beta},$$

where $\mathbf{X}^{(\omega)} = \hat{\mathbf{X}}(\boldsymbol{\theta}^{(\omega)})$ [equation (4.5)] and $\mathbf{w}^{(\omega)} = \mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta}^{(\omega)} + \mathbf{B}\mathbf{b}^{(\omega)}) + \mathbf{X}^{(\omega)}\boldsymbol{\beta}^{(\omega)} + \mathbf{Z}^{(\omega)}\mathbf{b}^{(\omega)}$. If we define

$$\tilde{l}_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(\omega)}, \mathbf{y}) = -\tfrac{1}{2} \log \mid \sigma^2(\boldsymbol{\Lambda} + \mathbf{Z}^{(\omega)}\tilde{\mathbf{D}}\mathbf{Z}^{(\omega)\mathrm{T}}) \mid$$
$$- \tfrac{1}{2}\sigma^{-2}[\mathbf{w}^{(\omega)} - \mathbf{X}^{(\omega)}\boldsymbol{\beta}]^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{Z}^{(\omega)}\tilde{\mathbf{D}}\mathbf{Z}^{(\omega)\mathrm{T}})^{-1}[\mathbf{w}^{(\omega)} - \mathbf{X}^{(\omega)}\boldsymbol{\beta}], \quad (5.1)$$

then the LME step with $\tilde{l}_F$ substituted for l_F is a linear mixed effects estimation problem of the type discussed in Laird and Ware (1982). This new LME step will result in the desired estimates since $\boldsymbol{\beta}^{(\mathrm{ML})}$, $\sigma^{(\mathrm{ML})}$, and $\boldsymbol{\theta}^{(\mathrm{ML})}$ maximize $\tilde{l}_F$.

The method for obtaining the RML estimates is exactly the same as that for the maximum likelihood estimates except that $\tilde{l}_F$ in equation (5.1) is replaced by

$$\tilde{l}_R(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(\omega)}, \mathbf{y}) = -\tfrac{1}{2} \log \mid \sigma^{-2}\mathbf{X}^{(\omega)\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{Z}^{(\omega)}\tilde{\mathbf{D}}\mathbf{Z}^{(\omega)\mathrm{T}})^{-1}\mathbf{X}^{(\omega)} \mid$$
$$+ \tilde{l}_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(\omega)}, \mathbf{y}).$$

6. Implementation and Convergence Behavior

6.1 *Pseudo-Data Step*

The pseudo-data calculations can be implemented by modifying a standard nonlinear least squares estimation routine. The unmodified routine will calculate the expected value ($\boldsymbol{\eta}$) vector of length N and derivative matrix of size $N \times r$ where r is the length of $\boldsymbol{\phi}$, the parameter vector for $\boldsymbol{\eta}$. For the pseudo-data problem the code must be modified to produce the $(N + Mq)$-vector

$$\tilde{\boldsymbol{\eta}} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\eta} \\ \tilde{\mathbf{D}}^{-1/2}\mathbf{b} \end{bmatrix}$$

and the $(N + Mq) \times (p + Mq)$ derivative matrix

$$\begin{bmatrix} \boldsymbol{\Lambda}^{-1/2}\mathbf{X}^{(\omega)} & \boldsymbol{\Lambda}^{-1/2}\mathbf{Z}^{(\omega)} \\ \mathbf{0} & \tilde{\mathbf{D}}^{-1/2} \end{bmatrix}.$$

If no advantage is taken of the special structure of the derivative matrix the computation of the parameter increment in the nonlinear least squares algorithm is of order $(N + Mq)(p + Mq)^2$. As the number of individuals M increases the time required could become prohibitive. However, by taking advantage of the "loosely coupled" structure of the problem, it is possible to reduce this computation to order $(N + Mq)(p + q)^2 + M(p + q)^3$ (Soo and Bates, in press).

6.2 *Linear Mixed Effects Step*

The LME step is a linear mixed effects model estimation problem where $\mathbf{w}^{(\omega)}$ is the response vector, $\mathbf{X}^{(\omega)}$ contains the predictor variables that correspond to the fixed effects, and $\mathbf{Z}^{(\omega)}$ contains the predictor variables that correspond to the random effects. Both the EM and Newton–Raphson (NR) algorithms have been proposed for this problem (Laird and Ware, 1982; Jennrich and Schluchter, 1986). We prefer NR to the EM algorithm for the reasons given in Lindstrom and Bates (1988). In particular, the low number of iterations usually required by the NR algorithm is especially important as the LME step may be executed many times. Also, the convergence of the NR algorithm can be checked with the orthogonality criterion (Bates and Watts, 1981). If the NR algorithm is implemented as described in Lindstrom and Bates (1988) then the iterations within the LME step will be of order Mq^4 for ML estimation and $M(q^4 + q^2p^4)$ for RML estimation. Because $\mathbf{X}^{(\omega)}$ and $\mathbf{Z}^{(\omega)}$ are not constant from one LME step to the next, initial calculations of order $(p + q)^2(N - M)$ must be done at the start of each LME step. The matrix decomposition methods proposed in Lindstrom and Bates (1988) must be modified if $\boldsymbol{\Lambda} \neq \mathbf{I}$. Note that in general $\hat{\mathbf{b}}$ need be calculated only at the final iteration of the LME step to provide starting values for the PD step.

6.3 *Convergence*

The augmented sum of squares

$$-\tfrac{1}{2}\sigma^{-2}[\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b})]^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}[\mathbf{y} - \boldsymbol{\eta}(\mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b})] - \tfrac{1}{2}\sigma^{-2}\mathbf{b}^{\mathrm{T}}\tilde{\mathbf{D}}^{-1}\mathbf{b}$$

is the objective function for the PD step. Each iteration in the PD step requires a decrease in this augmented sum of squares but it may increase from the end of one PD step to the beginning of the next since $\boldsymbol{\theta}$ may have changed in the intervening LME step. A similar pattern holds for $\tilde{l}_{\mathrm{F}}$ or $\tilde{l}_{\mathrm{R}}$ (the objective function for the LME step), which may decrease from the end of one LME step to the beginning of the next although it must increase within any LME step. The orthogonality convergence criterion measures convergence in the two steps independently and will also not necessarily decrease continuously from step to step. However, by requiring that three successive steps converge in one iteration we ensure that the whole algorithm has converged. In actual practice it seems to be most efficient to set a relatively low number (say, 5) for the maximum number of iterations within any one step.

The convergence behavior of the algorithm is in general quite good, especially considering that both of its steps are themselves iterative. As in standard nonlinear estimation, poor starting values can produce poor results and it is a good idea to take the time to find reasonable starting estimates. As in linear mixed effects models, the algorithm may perform poorly if unnecessary random effects are included in the model. Monitoring the covariance matrix $\mathbf{D}$ for singularities will indicate when this is a problem.

6.4 *Starting Values*

Starting values for $\boldsymbol{\beta}$, $\mathbf{b}$, $\mathbf{D}$, and $\boldsymbol{\Lambda}$ must be specified for the first PD step. We suggest using $\mathbf{b} = \mathbf{0}$ and $\boldsymbol{\Lambda} = \mathbf{I}$. As in all nonlinear regression problems, a starting value for $\boldsymbol{\beta}$ must be

inferred from the data and the form of the expectation function. One method for improving the initial guess for $\boldsymbol{\beta}$ is to fit the model without random effects to the pooled data using a standard nonlinear least squares package.

To obtain a starting value for $\mathbf{D}$ we suggest calculating the derivatives of the model function evaluated at the starting values for $\boldsymbol{\beta}$ and $\mathbf{b}$, then calculating $\mathbf{w}$ (see §5). These values can then be used in the starting value formula for $\mathbf{D}$ described in Laird, Lange, and Stram (1987).

6.5 *Uncertainty Estimates*

Approximate standard errors and correlations for the components of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ can be obtained after the last LME step using the Hessian of the likelihood function or the restricted likelihood function. It is also possible to get approximate standard errors for the components of $\boldsymbol{\beta}$ from the last PD step based on the linear approximation to the pseudo-data model but, since these standard errors are conditional on $\boldsymbol{\theta}$, they would tend to underestimate the actual uncertainty in $\boldsymbol{\beta}$. We would therefore recommend using the uncertainty estimates from the last LME step.

These uncertainty estimates are approximate in that they are based on a linear approximation to the model function at the parameter estimates. This type of approximation is commonly used to estimate the uncertainty in the parameters of nonlinear models. As pointed out in Bates and Watts (1988, Chap. 6), these uncertainty estimates can be quite inaccurate and a better appreciation of the uncertainty can be obtained by evaluating the profile likelihood and creating pairwise plots of the projected likelihood contours. The techniques applied there for the nonlinear regression model can also be applied here to the nonlinear mixed effects model.

7. Examples

7.1 *Orange Trees*

The following examples illustrate the computational performance of the algorithm described in Section 6 for the special case where $\boldsymbol{\Lambda} = \mathbf{I}$. We have implemented this case in FORTRAN-77 on a Vax-11/750 using a modified version of a nonlinear least squares routine (Bates and Watts, 1988) and an implementation of the Newton–Raphson algorithm for linear mixed effects models estimation described in Lindstrom and Bates (1988). The extension to a parameterized $\boldsymbol{\Lambda}$ is straightforward and would require only the specification of the derivatives of $\boldsymbol{\Lambda}$ with respect to its parameters.

To illustrate the algorithm we give a detailed description of its behavior when used to fit the logistic model to data on the growth of orange trees over time given in Draper and Smith (1981, p. 524). The data are presented in Figure 1 and consist of seven measurements of the trunk circumference (in millimeters) on each of five orange trees. These data are balanced because the measurements are taken on the same days for each individual (days are counted consecutively from January 1, 1969). The logistic model $y = \phi_1/(1 + \phi_2 e^{\phi_3 x})$ seems well suited to these data. Inspection of the data and some preliminary fits indicated that the only parameter that varies appreciably from tree to tree is ϕ_1, the asymptotic tree circumference. In the notation of Section 2, $\boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i$, where $\mathbf{A}_i = \mathbf{I}$ and $\mathbf{B}_i = (0, 0, 1)^{\mathrm{T}}$ for all i. Thus the model is

$$y_{ij} = \frac{\beta_1 + b_{i1}}{1 + \beta_2 e^{\beta_3 x_{ij}}} + e_{ij},$$

where the e_{ij} are independent and normal with mean zero and variance σ^2, and $\mathbf{b}_i \sim \mathrm{N}(0, \sigma^2\mathbf{D})$. In this example $\boldsymbol{\Lambda} = \mathbf{I}$ and $\mathbf{D}$ is a 1×1 matrix. The starting values for this example are $\boldsymbol{\beta}^0 = [150, 10, -.001]^{\mathrm{T}}$, $\mathbf{D}^0 = 117.6$, and $\mathbf{b}_i^0 = \mathbf{0}$ for all i.

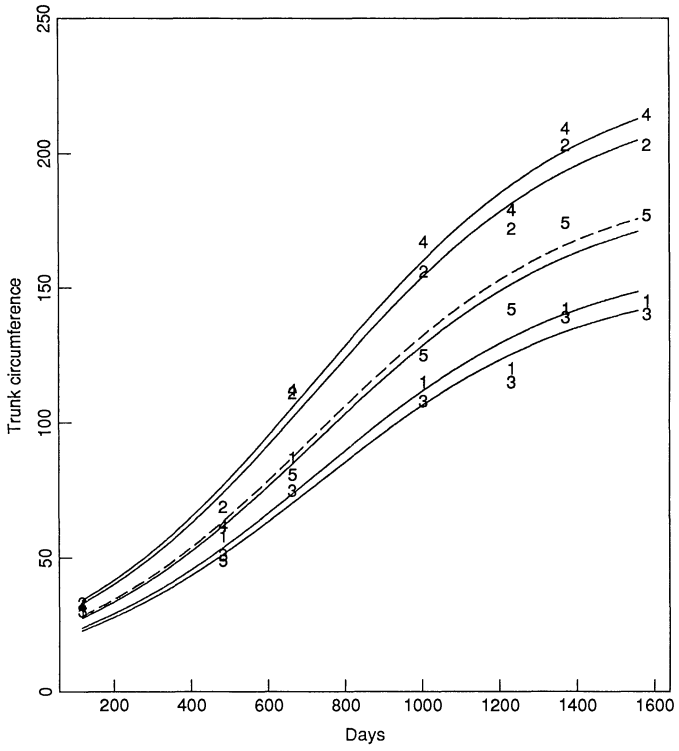

Figure 1. Trunk circumference (in millimeters) of five orange trees: Data and individual fitted curves from RML estimation. Dashed line represents the mean curve.

Table 1 gives the details of the iterations for RML estimation. The first iteration within each LME step is the calculation of $\hat{\beta}_{\text{lin}}$ [equation (3.1)] evaluated at the estimate of **D** and the appropriate derivative matrices returned from the preceding PD step. The converged value of $\mathbf{b} = [-29.51, 31.68, -37.13, 40.16, -5.20]^{\mathrm{T}}$ and the converged values of β, σ^2, and **D** can be found in Table 1. The predicted curves for the RML estimates are shown in Figure 1. The dashed line represents the mean curve, i.e., the curve with $\mathbf{b} = \mathbf{0}$.

These data were gathered over time so it is likely that there is some serial correlation among the measurements on each individual. This correlation is not modeled in the noise term $\mathbf{e}_{ij}$ but it is assumed that the random effects structure will be adequate to model its effect on the marginal covariance matrix. The correlation could be modeled more directly by defining Λ to be an AR(1) correlation matrix. However, difficulties with identifiability of parameters may result. Chi and Reinsel (1989) have investigated this question for linear mixed effects models.

The estimated marginal covariance–correlation matrix (correlations below, variances on, and covariances above the main diagonal) for each of the trees is

$$\begin{bmatrix} 92.94 & 61.08 & 83.81 & 127.08 & 149.16 & 159.13 & 169.34 \\ .44 & 204.07 & 189.55 & 287.39 & 337.35 & 359.88 & 382.97 \\ .48 & .73 & 326.06 & 394.39 & 462.94 & 493.86 & 525.55 \\ .51 & .78 & .85 & 663.80 & 701.89 & 748.77 & 796.82 \\ .52 & .79 & .86 & .91 & 889.83 & 878.92 & 935.31 \\ .52 & .79 & .86 & .92 & .93 & 1{,}003.57 & 997.79 \\ .52 & .80 & .87 & .92 & .93 & .94 & 1{,}127.75 \end{bmatrix}.$$

Table 1
Iterations for RML *estimation for the orange tree data*

Step	**D**	σ^2	β_1	β_2	β_3	$\tilde{l}_R$	Sum of sq.	Crit.	Sec.
Starting values	117.61		150.000	10.000	−.00100				
PD			150.000	10.000	−.00100		323,749	4.5×10^{2}	.33
			109.772	6.973	−.00110		319,175	5.0×10^{2}	.33
			83.223	4.950	−.00129		308,162	5.2×10^{2}	.15
			74.987	4.240	−.00161		277,050	5.0×10^{2}	.15
			83.794	4.649	−.00205		211,435	4.1×10^{2}	.15
LME	117.61	88.470	186.922	9.652	−.00384	−107.038			.32
	59.59	64.331	186.613	9.650	−.00386	−106.017		8.4×10^{-1}	.30
	36.90	65.963	186.231	9.649	−.00387	−105.468		1.9×10^{-1}	.45
	28.54	67.207	185.940	9.648	−.00388	−105.259		6.0×10^{-2}	.22
	23.16	68.472	185.644	9.647	−.00389	−105.147		6.4×10^{-2}	.27
	19.36	69.777	185.339	9.646	−.00390	−105.098		6.6×10^{-2}	.22
PD			185.339	9.646	−.00390		9,910	1.3×10	.33
			183.191	7.236	−.00286		2,650	1.0	.15
			191.292	8.104	−.00289		2,104	8.9×10^{-4}	.18
			191.206	8.151	−.00290		2,104	6.0×10^{-7}	.17
LME	19.36	69.777	191.205	8.152	−.00290	−105.753			.27
	18.88	65.943	191.185	8.153	−.00290	−105.752		1.1×10^{-2}	.27
	18.88	65.943	191.185	8.153	−.00290	−105.752		3.1×10^{-4}	.18
PD			191.185	8.153	−.00290		2,110	8.5×10^{-10}	.32
LME	18.88	65.943	191.184	8.153	−.00290	−105.752			.28
	18.88	65.943	191.184	8.153	−.00290	−105.752		6.3×10^{-4}	.15
PD			191.184	8.153	−.00290		2,110	6.5×10^{-13}	.33

The marginal covariance matrices of the trees are all equal because balanced data coupled with a random effect that is conditionally linear results in derivative matrices $\hat{\mathbf{Z}}_i$ [equation (4.1)] that are equal for all i. (A conditionally linear parameter is one for which the derivative of the model with respect to the parameter does not depend on the value of the parameter.) The marginal variances increase as time increases, reflecting the increasing deviation of the individual curves from the mean curve. Also, the marginal correlation between observations on the same individual is quite high and is greater at the larger time points. This reflects the large variability between trees as compared to within trees and this is most pronounced at the higher time points.

7.2 *Guinea Pigs*

Another example of data that can be modeled using nonlinear mixed effects models is the guinea pig data discussed in Johansen (1984). In this experiment 50 tissue samples were taken from the intestine of each of eight guinea pigs. For each guinea pig, five tissue samples were assigned randomly to each of ten different concentrations of B-methyl-glucoside and the uptake volume was measured in micromoles per milligram of fresh tissue per 2 minutes. Only the means of the five tissue samples at each concentration for each animal are reported and these 80 data points are plotted in Figure 2. In the terminology defined in Section 1, these are repeated measures data. However, the data are not longitudinal because the subsampling units (tissue sample) were randomly assigned to the treatments (concentrations) and there is no reason to suspect the presence of serial correlation. As in the first example, the data are balanced because the same set of concentrations was used for the samples from all of the guinea pigs. The proposed mechanistic model for uptake volume as a function of concentration is (Johansen, 1984)

$$y = \frac{\phi_1 x}{\phi_2 + x} + \phi_3 x.$$

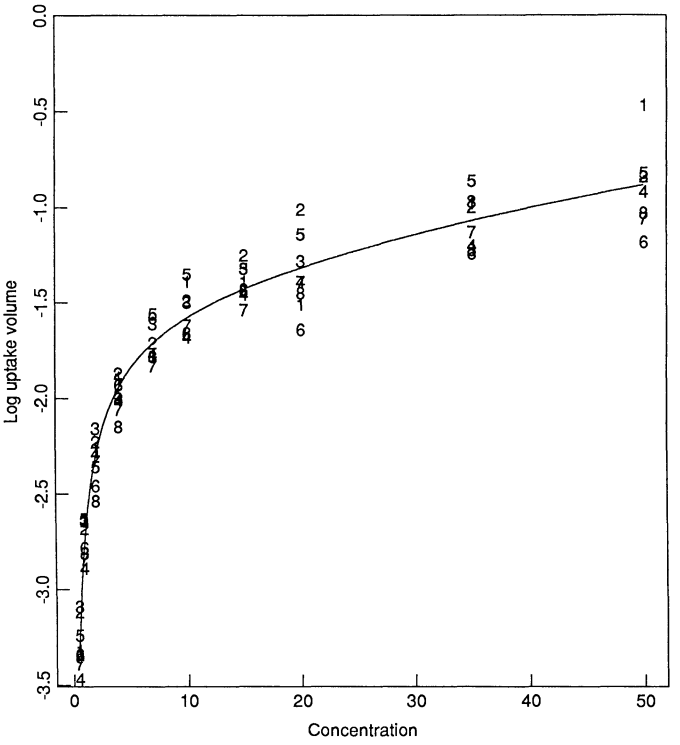

Figure 2. Uptake volume of B-methyl-glucoside in tissue samples from eight guinea pigs. Plot character indicates individual; solid line represents the mean curve.

We chose to model ϕ_2 and ϕ_3 as random effects. This implies that ϕ_1, the maximal rate of uptake, is a fixed effect, β_1; that ϕ_2, the Michaelis or affinity constant, is a random effect with mean β_2 and individual deviations b_{i1}; and that ϕ_3, a diffusion constant, is a random effect with mean β_3 and individual deviations b_{i2}. That is,

$$\mathbf{A}_i = \mathbf{I}, \quad \text{and} \quad \mathbf{B}_i = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{for all } i.$$

Thus the proposed mixed effects model for these data is

$$y_{ij} = \frac{\beta_1 x_{ij}}{\beta_2 + b_{i1} + x_{ij}} + (\beta_3 + b_{i2})x_{ij} + e_{ij},$$

where y_{ij} is the jth uptake volume for individual i, x_{ij} is the jth concentration level for individual i, $e_{ij} \sim \mathrm{N}(0, \sigma^2)$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]^{\mathrm{T}}$ is a vector of fixed population effects, and $\mathbf{b}_i = [b_{i1}, b_{i2}]^{\mathrm{T}}$ is a vector of individual random effects with $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{D})$. The model was fit in the natural log scale where both the data and the model were transformed to preserve the interpretability of the parameter estimates. The starting values used were

$$\boldsymbol{\beta}^0 = [.3, 5.0, .01]^{\mathrm{T}}, \quad \mathbf{b}_i^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{D}^0 = \begin{bmatrix} 18.04 & -.0698 \\ -.0698 & 5.22 \times 10^{-4} \end{bmatrix}.$$

The algorithm converged in five "first-level" iterations (where each first-level iteration consists of an LME and a PD step). The "second-level" iterations within the LME and PD

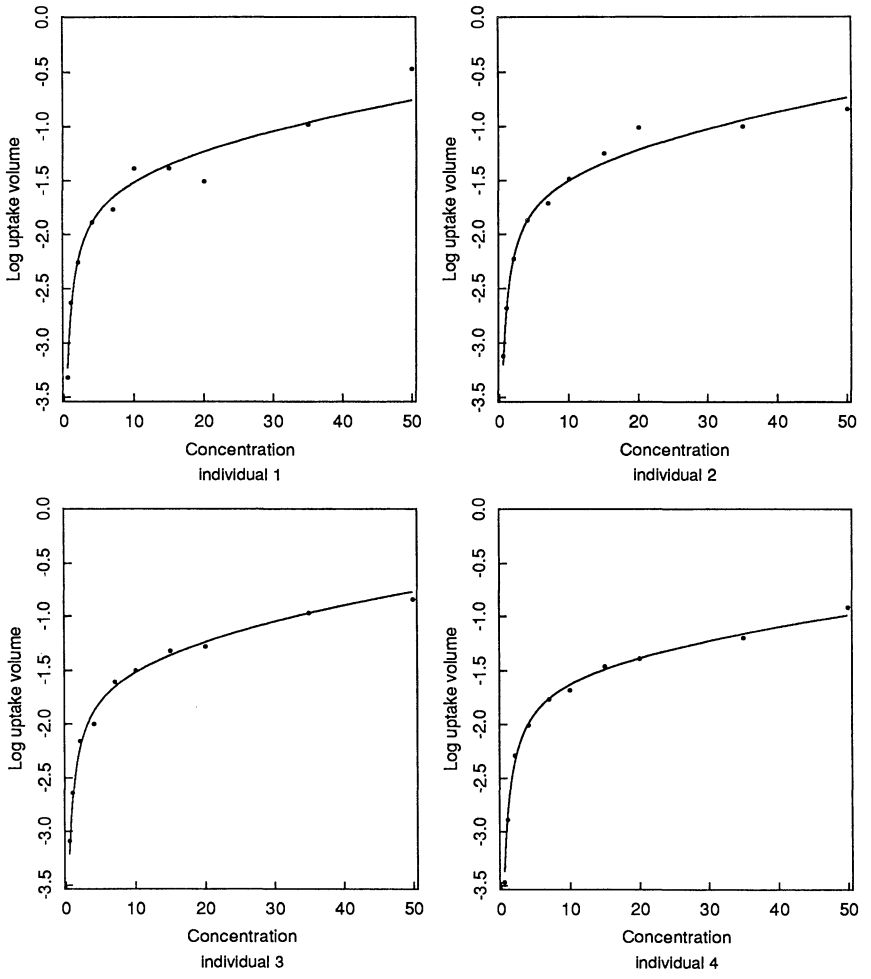

Figure 3. Guinea pig example: Four individual fitted curves and data.

steps used an average of .67 and .84 CPU second, respectively. The converged estimates are

$$\boldsymbol{\beta} = [.2006,\ 2.393,\ .004452]^{\mathrm{T}}, \quad \mathbf{D} = \begin{bmatrix} 6.69 & -.0329 \\ -.0329 & 1.88 \times 10^{-4} \end{bmatrix}, \quad \text{and} \quad \sigma^2 = .00908.$$

The predicted curves from the RML converged values are shown for the first four guinea pigs in Figure 3. The solid curve in Figure 2 is the mean curve ($\mathbf{b} = \mathbf{0}$).

8. Discussion

There remain a number of unanswered questions about nonlinear mixed effects models. For instance, procedures for selecting and criticizing models need to be developed and studied using real data. Also, little is known about the properties of the "naive" population estimators defined as simple weighted or unweighted means of nonlinear least squares individual estimates. The naive estimators will not be ML except in the case of completely balanced data with a linear expectation function. However, they are simple and intuitive and deserve further study.

There are similarities between our approach to nonlinear random effects models and recent work in the area of generalized linear models (GLIM) for longitudinal data (e.g., Stiratelli et al., 1984; Liang and Zeger, 1986). Zeger, Liang, and Albert (1988) discuss the current literature in this area and categorize the types of models that have been proposed into subject-specific (SS) models and population average (PA) or marginal models. The distinction between SS and PA models is that SS models concentrate on modeling individuals in order to understand the population, whereas PA models concentrate on the marginal distribution only.

For linear expectation functions (or identity link in GLIM models) the distinction between the two types of models is less important than for nonlinear expectations. In linear models the fixed effects will have the same interpretation in both PA and SS models, as discussed by Zeger et al. In fact, the SS linear mixed effects model can be viewed as just one way to generate a parameterization for the marginal covariance in the PA model. Mixed effects models are often used because of the need to parameterize the marginal covariance matrix either because of imbalance in the data or because the number of individuals is not large compared to the number of observations per individual.

When the expectation function is nonlinear in the parameters, the choice between SS and PA models has consequences for the interpretation of parameter estimates. As would be expected, the estimates in an SS model refer to individual behavior, whereas the parameters in a PA model are describing the group as a whole. For example, if the predictor designs for the fixed and random effects are the same, then in SS models the fixed effects are "typical" parameters, whereas in PA models the fixed effects are the parameters that produce a "typical" response vector.

The nonlinear mixed effect model we have proposed is an SS model and is similar in form to the GLIM SS model described for longitudinal data by Zeger et al. However, the GLIM model assumes a more restrictive conditional expectation function while allowing a more general conditional error structure. The GLIM SS model for the conditional expectation of y_{ij} is $\mathrm{E}(y_{ij} \mid \mathbf{b}_i) = h^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i)$. The link function h is more restrictive since it is a function of one variable rather than a general nonlinear function of the parameters and predictor variables. In practice, GLIM models are most widely studied for a small set of useful link functions including the logit, probit, and log links.

Acknowledgements

The authors wish to thank the referees for their helpful comments and suggestions. We also wish to acknowledge the support of the National Science Foundation through Grant Number DMS-8801996 and the National Institutes of Health through Grant Number CA18332-13.

Résumé

On propose un modèle général non linéaire à effets mixtes pour des données à mesures répétées en définissant les estimateurs de ses paramètres. Les estimateurs proposés sont une combinaison naturelle des estimateurs des moindres carrés pour la partie non linéaire et fixe des effets et des estimateurs du maximum de vraisemblance pour la partie linéaire et mixte des effets. En ce qui concerne l'estimation, on met en oeuvre la méthode de Newton–Raphson en se servant de techniques de calcul préalablement développées pour les modèles non linéaires à effets fixes et pour les modèles linéaires à effets mixtes. On présente deux exemples et l'accent est mis sur les liens existant entre ce travail et les travaux récents sur les modèles linéaires généralisés à effets mixtes.

References

Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of nonlinearity (with Discussion). *Journal of the Royal Statistical Society, Series B* **42,** 1–25.

Bates, D. M. and Watts, D. G. (1981). A relative offset orthogonality convergence criterion for nonlinear least squares. *Technometrics* **23,** 179–183.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications.* New York: Wiley.

Berkey, C. S. (1982). Bayesian approach for a nonlinear growth model. *Biometrics* **38,** 953–961.

Chi, E. M. and Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association* **84,** 452–459.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis,* 2nd edition. New York: Wiley.

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

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated measures models with structural covariance matrices. *Biometrics* **42,** 805–820.

Johansen, S. (1984). *Functional Relations, Random Coefficients, and Nonlinear Regression with Application to Kinetic Data.* New York: Springer-Verlag.

Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association* **82,** 97–105.

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

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

Lindstrom, M. J. and Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association* **83,** 1014–1022.

Racine-Poon, A. (1985). Bayesian approach to nonlinear random effects models. *Biometrics* **41,** 1015–1023.

Sheiner, L. B. and Beal, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis–Menten model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics* **8,** 553–571.

Soo, Y. and Bates, D. M. Loosely coupled nonlinear least squares. *Computational Statistics and Data Analysis* (in press).

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

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

Received September 1988; *revised August* 1989; *accepted August* 1989.