José C. Pinheiro
Douglas M. Bates

Mixed-Effects Models in S and S-PLUS

With 172 Illustrations

Springer

José C. Pinheiro
Department of Biostatistics
Novartis Pharmaceuticals
One Health Plaza
East Hanover, NJ 07936-1080
USA
jose.pinheiro@pharma.novartis.com

Douglas M. Bates
Department of Statistics
University of Wisconsin
Madison, WI 53706-1685
USA
bates@stat.wisc.edu

Library of Congress Cataloging-in-Publication Data
Pinheiro, José C.
Mixed-effects models in S and S-PLUS / José C. Pinheiro, Douglas M. Bates
p. cm. — (Statistics and computing)
Includes bibliographical references and index.
ISBN 0-387-98957-9 (alk. paper)
I. Bates, Douglas M. II. Title. III. Series.
QA76.73.S15P56 2000
005.13'3—dc21 99-053566

Printed on acid-free paper.

Printed in the United States of America. (HAM)

9 8 7 6 5 SPIN 10995662

Springer Verlag is a part of *Springer Science+Business Media*

springeronline.com

2

Theory and Computational Methods for Linear Mixed-Effects Models

In this chapter we present the theory for the linear mixed-effects model introduced in Chapter 1. A general formulation of LME models is presented and illustrated with examples. Estimation methods for LME models, based on the likelihood or the restricted likelihood of the parameters, are described, together with the computational methods used to implement them in the `lme` function. Asymptotic results on the distribution of the maximum likelihood estimators and the restricted maximum likelihood estimators are used to derive confidence intervals and hypotheses tests for the model's parameters.

The purpose of this chapter is to present an overview of the theoretical and computational aspects of LME models that allows the evaluation of the strengths and limitations of such models for practical applications. It is not the purpose of this chapter to present a thorough theoretical description of LME models. Such a comprehensive treatment of the theory of linear mixed-effects models can be found, for example, in Searle, Casella and McCulloch (1992) or in Vonesh and Chinchilli (1997).

Readers who are more interested in the applications of LME models and the use of the functions and methods in the nlme library to fit such models can, without loss of continuity, skip this chapter and go straight to Chapter 3. If you decide to skip this chapter at a first reading, it is recommended that you return to it (especially §2.1) at a later time to get a good understanding of the LME model formulation and its assumptions and limitations.

2.1 The LME Model Formulation

Linear mixed-effects models are mixed-effects models in which both the fixed and the random effects occur linearly in the model function. They extend linear models by incorporating random effects, which can be regarded as additional error terms, to account for correlation among observations within the same group.

In this section we present a general formulation for LME models proposed by Laird and Ware (1982). The original single-level formulation is described in §2.1.1 and its multilevel extension is described in §2.1.2.

2.1.1 Single Level of Grouping

For a single level of grouping, the linear mixed-effects model described by Laird and Ware (1982) expresses the n_i-dimensional response vector $\boldsymbol{y}_i$ for the ith group as

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, M, \\ \boldsymbol{b}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}), \end{aligned} \tag{2.1}$$

where $\boldsymbol{\beta}$ is the p-dimensional vector of *fixed effects*, $\boldsymbol{b}_i$ is the q-dimensional vector of *random effects*, $\boldsymbol{X}_i$ (of size $n_i \times p$) and $\boldsymbol{Z}_i$ (of size $n_i \times q$) are known fixed-effects and random-effects regressor matrices, and $\boldsymbol{\epsilon}_i$ is the n_i-dimensional *within-group error* vector with a spherical Gaussian distribution. The assumption $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \sigma^2\boldsymbol{I}$ can be relaxed as shown in Chapter 5, where we describe extensions that allow us to model nonconstant variances or special within-group correlation structures. The random effects $\boldsymbol{b}_i$ and the within-group errors $\boldsymbol{\epsilon}_i$ are assumed to be independent for different groups and to be independent of each other for the same group.

Because the distribution of the random effects vectors $\boldsymbol{b}_i$ is assumed to be normal (or Gaussian) with a mean of $\mathbf{0}$, it is completely characterized by its variance–covariance matrix $\boldsymbol{\Psi}$. This matrix must be symmetric and positive semi-definite; that is, all its eigenvalues must be non-negative. We will make the stronger assumption that it is *positive-definite* which is to say that all its eigenvalues must be strictly positive. We can make this restriction because an indefinite model can always be re-expressed as a positive-definite model of lower dimension.

The random effects $\boldsymbol{b}_i$ are defined to have a mean of $\mathbf{0}$ and therefore any nonzero mean for a term in the random effects must be expressed as part of the fixed-effects terms. Thus, the columns of $\boldsymbol{Z}_i$ are usually a subset of the columns of $\boldsymbol{X}_i$.

When computing with the model it is more convenient to express the variance–covariance matrix in the form of a *relative precision factor*, $\boldsymbol{\Delta}$,

which is any matrix that satisfies

$$\frac{\boldsymbol{\Psi}^{-1}}{1/\sigma^2} = \boldsymbol{\Delta}^T\boldsymbol{\Delta}.$$

If $\boldsymbol{\Psi}$ is positive-definite then such a $\boldsymbol{\Delta}$ will exist, but it need not be unique. The Cholesky factor (Thisted, 1988, §3.3) of $\sigma^2\boldsymbol{\Psi}^{-1}$ is one possible $\boldsymbol{\Delta}$. The matrix $\boldsymbol{\Delta}$ is called a *relative precision factor* because it factors the *precision matrix*, $\boldsymbol{\Psi}^{-1}$, of the random effects, expressed relative to the precision, $1/\sigma^2$, of the ϵ_i.

We use some of the examples in Chapter 1 to illustrate the general LME model formulation.

Railway Rails Experiment

In the case of the rails data introduced in §1.1, $M = 6$, $n_i = 3$, $i = 1, \ldots, 6$, $p = q = 1$, and the regressor matrices for the fixed and random effects are particularly simple:

$$\boldsymbol{X}_i = \boldsymbol{Z}_i = \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad i = 1, \ldots, 6.$$

The random effects $b_i, i = 1, \ldots, 6$ are scalars; hence their variance σ_b^2 is also a scalar, as is the relative precision factor, Δ. There is only one choice for Δ (up to changes in sign) and that is

$$\Delta = \sqrt{\sigma^2/\sigma_b^2}.$$

Ergometric Experiment of Types of Stools

The data for the stools ergometric experiment of §1.2 are balanced, with $M = 6$, $n_i = 4$, $i = 1, \ldots, 6$, $p = 4$, and $q = 1$. The fixed-effects regressor matrices $\boldsymbol{X}_i$ are determined by the contrasts chosen to represent the types of stool. For the Helmert contrasts parameterization used in the fit of the `fm1Stool` object in §1.2.1, we have

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 0 & 2 & -1 \\ 1 & 0 & 0 & 3 \end{bmatrix}, \quad i = 1, \ldots, 6.$$

The random-effects regression matrices $\boldsymbol{Z}_i$ and the relative precision factor Δ are the same as in the rails example.

Orthodontic Growth Curve in Girls

The orthodontic growth curve data for females presented in §1.4.1 are also balanced, with $M = 11$, $n_i = 4$, $i = 1, \ldots, 11$. For the LME model with

random effects for both the intercept and the slope, used to fit the `fm2OrthF` object in §1.4.1, we have $p = q = 2$ and the fixed- and random-effects regressor matrices are identical and given by

$$\boldsymbol{X}_i = \boldsymbol{Z}_i = \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix}, \quad i = 1, \ldots, 11.$$

Any square-root of the 2×2 matrix $\sigma^2 \boldsymbol{\Psi}^{-1}$ can be used as a relative precision factor in this case.

2.1.2 A Multilevel LME Model

The Laird–Ware formulation for single-level LME models presented in §2.1.1 can be extended to multiple, nested levels of random effects. In the case of two nested levels of random effects the response vectors at the innermost level of grouping are written $\boldsymbol{y}_{ij}, i = 1, \ldots, M, j = 1, \ldots, M_i$ where M is the number of first-level groups and M_i is the number of second-level groups within first-level group i. The length of $\boldsymbol{y}_{ij}$ is n_{ij}.

The fixed-effects model matrices are $\boldsymbol{X}_{ij}, i = 1, \ldots, M, j = 1, \ldots, M_i$ of size $n_{ij} \times p$. Using first-level random effects $\boldsymbol{b}_i$ of length q_1 and second-level random effects $\boldsymbol{b}_{ij}$ of length q_2 with corresponding model matrices $\boldsymbol{Z}_{i,j}$ of size $n_i \times q_1$ and $\boldsymbol{Z}_{ij}$ of size $n_i \times q_2$, we write the model as

$$\begin{gathered} \boldsymbol{y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{i,j}\boldsymbol{b}_i + \boldsymbol{Z}_{ij}\boldsymbol{b}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, M_i, \\ \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_1), \quad \boldsymbol{b}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_2), \quad \boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}). \end{gathered} \tag{2.2}$$

The level-1 random effects $\boldsymbol{b}_i$ are assumed to be independent for different i, the level-2 random effects $\boldsymbol{b}_{ij}$ are assumed to be independent for different i or j and to be independent of the level-1 random effects, and the within-group errors $\boldsymbol{\epsilon}_{ij}$ are assumed to be independent for different i or j and to be independent of the random effects.

Extensions to an arbitrary number Q of levels of random effects follow the same general pattern. For example, with $Q = 3$ the response for the kth level-3 unit within the jth level-2 unit within the ith level-1 unit is written

$$\begin{gathered} \boldsymbol{y}_{ijk} = \boldsymbol{X}_{ijk}\boldsymbol{\beta} + \boldsymbol{Z}_{i,jk}\boldsymbol{b}_i + \boldsymbol{Z}_{ij,k}\boldsymbol{b}_{ij} + \boldsymbol{Z}_{ijk}\boldsymbol{b}_{ijk} + \boldsymbol{\epsilon}_{ijk}, \\ i = 1, \ldots, M, \quad j = 1, \ldots, M_i, \quad k = 1, \ldots, M_{ij}, \\ \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_1), \quad \boldsymbol{b}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_2), \quad \boldsymbol{b}_{ijk} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_3), \quad \boldsymbol{\epsilon}_{ijk} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}). \end{gathered}$$

Note that the distinction between, say, the kth horizontal section of the regressor matrix for the level-2 random effect $\boldsymbol{b}_{ij}$, written $\boldsymbol{Z}_{ij,k}$, and the

jkth horizontal section of the regressor matrix for the level-1 random effect $\boldsymbol{b}_i$, written $\boldsymbol{Z}_{i,jk}$, is the position of the comma in the subscripts.

As with a single level of random effects, we will express the variance–covariance matrices, $\boldsymbol{\Psi}_q$, $q = 1, \ldots, Q$, in terms of relative precision factors $\boldsymbol{\Delta}_q$.

In this book, we only consider mixed-effects models with a multivariate normal (or Gaussian) distribution for the random effects and the within-group errors. Generally we assume that the variance–covariance matrix $\boldsymbol{\Psi}_q$ for the level-q random effects can be any positive-definite, symmetric matrix. In some models we will further restrict the form of $\boldsymbol{\Psi}_q$, say by requiring that it be diagonal or that it be a multiple of the identity.

Those familiar with the *multilevel modeling* literature (Bryk and Raudenbush, 1992; Goldstein, 1995) may notice that we count "levels" differently. In that literature the model (2.1) is called a two-level model because there are two levels of random variation. Similarly, the model (2.2) is called a three-level model. We prefer the terminology from the experimental design literature and count the number of "levels" as the number of nested levels of random effects.

Split-Plot Experiment on Varieties of Oats and Nitrogen Levels

We use the split-plot experiment on the yield of three different varieties of oats measured at four different concentrations of nitrogen, described in §1.6, to illustrate the multilevel LME model formulation. The final model used in that section, corresponding to the fitted object `fm4Oats`, represents the yield y_{ijk} for the jth variety of oat at the kth nitrogen concentration N_k in the ith block as

$$y_{ijk} = \beta_0 + \beta_1 N_k + b_i + b_{ij} + \epsilon_{ijk},\ i = 1, \ldots, 6,\ j = 1, \ldots, 3,\ k = 1, \ldots, 4,$$
$$b_i \sim \mathcal{N}\left(0, \sigma_1^2\right), \quad b_{ij} \sim \mathcal{N}\left(0, \sigma_2^2\right), \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2).$$

The fixed effects are the intercept β_0 and the nitrogen slope β_1. The b_i denote the Block random effects, the b_{ij} denote the Variety within Block random effects, and the ϵ_{ijk} denote the within-group errors. This is an example of a two-level mixed-effects model, with the b_{ij} random effects *nested* within the b_i random effects.

In this example, $M = 6$, $M_i = 3$, $n_{ij} = 4$, $i = 1, \ldots, 6$ $j = 1, \ldots, 3$, $p = 2$, and $q_1 = q_2 = 1$. The regressor matrices are

$$\boldsymbol{X}_{ij} = \begin{bmatrix} 1 & 0.0 \\ 1 & 0.2 \\ 1 & 0.4 \\ 1 & 0.6 \end{bmatrix}, \quad \boldsymbol{Z}_{i,j} = \boldsymbol{Z}_{ij} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad i = 1, \ldots, 6, \quad j = 1, \ldots, 3.$$

Because all the random effects are scalars, the precision factors are uniquely defined (up to changes in sign) as

$$\Delta_1 = \sqrt{\sigma^2/\sigma_1^2} \quad \text{and} \quad \Delta_2 = \sqrt{\sigma^2/\sigma_2^2}.$$

2.2 Likelihood Estimation for LME Models

Several methods of parameter estimation have been used for linear mixed-effects models. We will concentrate on two general methods: maximum likelihood (ML) and restricted maximum likelihood (REML). Descriptions and comparisons of the various estimation methods used for LME models can be found, for example, in Searle et al. (1992) and Vonesh and Chinchilli (1997).

2.2.1 The Single-Level LME Likelihood Function

Consider first the model (2.1) that has a single level of random effects. The parameters of the model are $\boldsymbol{\beta}$, σ^2, and whatever parameters determine $\boldsymbol{\Delta}$. We use $\boldsymbol{\theta}$ to represent an unconstrained set of parameters that determine $\boldsymbol{\Delta}$. We will discuss parameterizations of $\boldsymbol{\Delta}$ in §2.2.7—for now we will simply assume that a suitable parameterization has been chosen.

The *likelihood function* for the model (2.1) is the probability density for the data given the parameters, but regarded as a function of the parameters with the data fixed, instead of as a function of the data with the parameters fixed. That is,

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \boldsymbol{y}\right) = p(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2),$$

where L is the likelihood, p is a probability density, and $\boldsymbol{y}$ is the entire N-dimensional response vector, $N = \sum_{i=1}^{M} n_i$.

Because the nonobservable random effects $\boldsymbol{b}_i, i = 1, \ldots, M$ are part of the model, we must integrate the conditional density of the data given the random effects with respect to the marginal density of the random effects to obtain the marginal density for the data. We can use the independence of the $\boldsymbol{b}_i$ and the $\boldsymbol{\epsilon}_i$ to express this as

$$\begin{aligned} L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \boldsymbol{y}\right) &= \prod_{i=1}^{M} p\left(\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) \\ &= \prod_{i=1}^{M} \int p\left(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{b}_i | \boldsymbol{\theta}, \sigma^2\right) d\boldsymbol{b}_i, \end{aligned} \tag{2.3}$$

where the conditional density of $\boldsymbol{y}_i$ is multivariate normal

$$p\left(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\beta},\sigma^2\right)=\frac{\exp\left(-\left\|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{b}_i\right\|^2/2\sigma^2\right)}{\left(2\pi\sigma^2\right)^{n_i/2}} \tag{2.4}$$

and the marginal density of $\boldsymbol{b}_i$ is also multivariate normal

$$\begin{aligned} p\left(\boldsymbol{b}_i|\boldsymbol{\theta},\sigma^2\right) &= \frac{\exp\left(-\boldsymbol{b}_i^T\boldsymbol{\Psi}^{-1}\boldsymbol{b}_i/2\right)}{(2\pi)^{q/2}\sqrt{|\boldsymbol{\Psi}|}} \\ &= \frac{\exp\left(-\left\|\boldsymbol{\Delta}\boldsymbol{b}_i\right\|^2/2\sigma^2\right)}{\left(2\pi\sigma^2\right)^{q/2}\operatorname{abs}|\boldsymbol{\Delta}|^{-1}}, \end{aligned} \tag{2.5}$$

where $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$. Substituting (2.4) and (2.5) into (2.3) provides the likelihood as

$$\begin{aligned} &L\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y}\right)= \\ &\prod_{i=1}^{M}\frac{\operatorname{abs}|\boldsymbol{\Delta}|}{\left(2\pi\sigma^2\right)^{n_i/2}}\int\frac{\exp\left[-\left(\left\|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{b}_i\right\|^2+\left\|\boldsymbol{\Delta}\boldsymbol{b}_i\right\|^2\right)/2\sigma^2\right]}{\left(2\pi\sigma^2\right)^{q/2}}d\boldsymbol{b}_i \\ &=\prod_{i=1}^{M}\frac{\operatorname{abs}|\boldsymbol{\Delta}|}{\left(2\pi\sigma^2\right)^{n_i/2}}\int\frac{\exp\left(-\left\|\tilde{\boldsymbol{y}}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\beta}-\tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\right\|^2/2\sigma^2\right)}{\left(2\pi\sigma^2\right)^{q/2}}d\boldsymbol{b}_i, \end{aligned} \tag{2.6}$$

where

$$\tilde{\boldsymbol{y}}_i=\begin{bmatrix}\boldsymbol{y}_i\\ \boldsymbol{0}\end{bmatrix},\quad \tilde{\boldsymbol{X}}_i=\begin{bmatrix}\boldsymbol{X}_i\\ \boldsymbol{0}\end{bmatrix},\quad \tilde{\boldsymbol{Z}}_i=\begin{bmatrix}\boldsymbol{Z}_i\\ \boldsymbol{\Delta}\end{bmatrix}, \tag{2.7}$$

are augmented data vectors and model matrices. This approach of changing the contribution of the marginal distribution of the random effects into extra rows for the response and the design matrices is called a *pseudo-data* approach because it creates the effect of the marginal distribution by adding "pseudo" observations.

The exponent in the integral of (2.6) is in the form of a squared norm or, more specifically, a residual sum-of-squares. We can determine the conditional modes of the random effects given the data, written $\widehat{\boldsymbol{b}}_i$, by minimizing this residual sum-of-squares. This is a standard least squares problem for which we could write the solution as

$$\widehat{\boldsymbol{b}}_i=\left(\tilde{\boldsymbol{Z}}_i{}^T\tilde{\boldsymbol{Z}}_i\right)^{-1}\tilde{\boldsymbol{Z}}_i{}^T\left(\tilde{\boldsymbol{y}}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\beta}\right).$$

The squared norm can then be expressed as

$$\begin{aligned} \left\|\tilde{\boldsymbol{y}}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\beta}-\tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\right\|^2 &= \left\|\tilde{\boldsymbol{y}}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\beta}-\tilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2+\left\|\tilde{\boldsymbol{Z}}_i\left(\boldsymbol{b}_i-\widehat{\boldsymbol{b}}_i\right)\right\|^2 \\ &= \left\|\tilde{\boldsymbol{y}}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\beta}-\tilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2+\left(\boldsymbol{b}_i-\widehat{\boldsymbol{b}}_i\right)^T\tilde{\boldsymbol{Z}}_i{}^T\tilde{\boldsymbol{Z}}_i\left(\boldsymbol{b}_i-\widehat{\boldsymbol{b}}_i\right). \end{aligned} \tag{2.8}$$

The first term in (2.8) does not depend on $\boldsymbol{b}_i$ so its exponential can be factored out of the integral in (2.6). Integrating the exponential of the second term in (2.8) is equivalent, up to a constant, to integrating a multivariate normal density function. Note that

$$\begin{aligned}
&\frac{\sqrt{|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i|}}{\sqrt{|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i|}} \int \frac{\exp\left[-\left(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right)^T \tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i \left(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right)/2\sigma^2\right]}{(2\pi\sigma^2)^{q/2}} d\boldsymbol{b}_i \\
&\quad = \frac{1}{\sqrt{|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i|}} \int \frac{\exp\left[-\left(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right)^T \tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i \left(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i\right)/2\sigma^2\right]}{(2\pi\sigma^2)^{q/2}/\sqrt{|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i|}} d\boldsymbol{b}_i \\
&\quad = \frac{1}{\sqrt{|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i|}} = \frac{1}{\sqrt{|\boldsymbol{Z}_i^T \boldsymbol{Z}_i + \boldsymbol{\Delta}^T \boldsymbol{\Delta}|}}. \qquad (2.9)
\end{aligned}$$

By combining (2.8) and (2.9) we can express the integral in (2.6) as

$$\begin{aligned}
&\int \frac{\exp\left[-\left\|\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\right\|^2/2\sigma^2\right]}{(2\pi\sigma^2)^{q/2}} d\boldsymbol{b}_i \\
&\qquad = \frac{\exp\left(-\left\|\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2/2\sigma^2\right)}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i\right|}}
\end{aligned}$$

to give

$$\begin{aligned}
&L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \boldsymbol{y}\right) \\
&\quad = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\frac{-\sum_{i=1}^M \left\|\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2}{2\sigma^2}\right) \prod_{i=1}^M \frac{\operatorname{abs}|\boldsymbol{\Delta}|}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i\right|}}. \qquad (2.10)
\end{aligned}$$

The expression (2.10) could be used directly in an optimization routine to calculate the maximum likelihood estimates for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\sigma}^2$. However, the optimization is much simpler if we first *concentrate* or *profile* the likelihood so it is a function of $\boldsymbol{\theta}$ only. That is, we calculate the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}^2(\boldsymbol{\theta})$ as the values that maximize $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$ for a given $\boldsymbol{\theta}$. Notice that the parts of (2.10) involving $\boldsymbol{\beta}$ and σ^2 are identical in form to the likelihood for a linear regression model so $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}^2(\boldsymbol{\theta})$ can be determined from standard linear regression theory.

We do need to be careful because the least squares estimates for $\boldsymbol{\beta}$ will depend on the conditional modes $\widehat{\boldsymbol{b}}_i$ and these, in turn, depend on $\boldsymbol{\beta}$. Thus, we must determine these least squares values jointly as the least squares solution to

$$\left(\widehat{\boldsymbol{b}}_1^T, \ldots, \widehat{\boldsymbol{b}}_M^T, \widehat{\boldsymbol{\beta}}^T\right)^T = \arg \min_{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M, \boldsymbol{\beta}} \|\boldsymbol{y}_e - \boldsymbol{X}_e(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M, \boldsymbol{\beta})^T\|^2,$$

where

$$\boldsymbol{X}_e = \begin{bmatrix} \boldsymbol{Z}_1 & \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{X}_1 \\ \boldsymbol{\Delta} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_2 & \ldots & \mathbf{0} & \boldsymbol{X}_2 \\ \mathbf{0} & \boldsymbol{\Delta} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{Z}_M & \boldsymbol{X}_M \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Delta} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \boldsymbol{y}_e = \begin{bmatrix} \boldsymbol{y}_1 \\ \mathbf{0} \\ \boldsymbol{y}_2 \\ \mathbf{0} \\ \vdots \\ \boldsymbol{y}_M \\ \mathbf{0} \end{bmatrix}. \tag{2.11}$$

Conceptually we could write

$$\left(\widehat{\boldsymbol{b}}_1^T, \ldots, \widehat{\boldsymbol{b}}_M^T, \widehat{\boldsymbol{\beta}}^T\right)^T = (\boldsymbol{X}_e^T \boldsymbol{X}_e)^{-1} \boldsymbol{X}_e^T \boldsymbol{y}_e,$$

but we definitely would *not* want to calculate these values this way. The matrix $\boldsymbol{X}_e$ is sparse and can be very large. If possible we want to take advantage of the sparsity and avoid working directly with $\boldsymbol{X}_e$.

Linear regression theory also gives us the conditional maximum likelihood estimate for σ^2

$$\widehat{\sigma}^2(\boldsymbol{\theta}) = \frac{\|\boldsymbol{y}_e - \boldsymbol{X}_e(\widehat{\boldsymbol{b}}_1^T, \ldots, \widehat{\boldsymbol{b}}_M^T, \widehat{\boldsymbol{\beta}}^T)^T\|^2}{N}. \tag{2.12}$$

Notice that the maximum likelihood estimate of σ^2 is the residual sum-of-squares divided by N, not by $N - p$.

Substituting these conditional estimates back into (2.10) provides the profiled likelihood

$$L(\boldsymbol{\theta}) = L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \widehat{\sigma}^2(\boldsymbol{\theta})) = \frac{\exp(-N/2)}{[2\pi\widehat{\sigma}^2(\boldsymbol{\theta})]^{N/2}} \prod_{i=1}^{M} \frac{\text{abs}\,|\boldsymbol{\Delta}|}{\sqrt{\left|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i\right|}}. \tag{2.13}$$

We do not actually need to calculate the values of $\widehat{\boldsymbol{b}}_1, \ldots, \widehat{\boldsymbol{b}}_M$ or $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ to evaluate the profiled likelihood. We only need to know the norm of the residual from the augmented least squares problem. The decomposition methods described in §2.2.2 provide us with fast, convenient methods of calculating this.

The pseudo-data representation of the marginal density $p(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$ used in (2.6) is just one way of expressing this density and deriving the likelihood. It is also possible to describe this density as a normal distribution with mean $\mathbf{0}$ and a patterned variance–covariance matrix $\boldsymbol{\Sigma}_i$—a representation that is often used to derive the likelihood for the parameters in a linear mixed-effects model. Although we will not use this representation extensively in this chapter, we will use it in Chapter 5, so we present some of this derivation of the likelihood here.

The model (2.1) can be re-expressed as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i^*, \quad i = 1, \ldots, M, \tag{2.14}$$

where $\boldsymbol{\epsilon}_i^* = \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$. Because the $\boldsymbol{\epsilon}_i^*$ are the sum of two independent multivariate normal random vectors, they are independently distributed as multivariate normal vectors with mean $\mathbf{0}$ and variance–covariance matrix $\sigma^2\boldsymbol{\Sigma}_i$, where $\boldsymbol{\Sigma}_i = \boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T/\sigma^2$. It then follows from (2.14) that the $\boldsymbol{y}_i$ are independent multivariate normal random vectors with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and variance–covariance matrix $\sigma^2\boldsymbol{\Sigma}_i$. That is,

$$p\left(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \exp\left(\frac{\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)}{-2\sigma^2}\right) |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}.$$

For a given value of $\boldsymbol{\theta}$, the values of $\boldsymbol{\beta}$ and σ^2 that maximize the likelihood could be written as

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\sum_{i=1}^{M} \boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{M} \boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i,$$

$$\widehat{\sigma}^2(\boldsymbol{\theta}) = \frac{\sum_{i=1}^{M}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right)}{N}.$$

Computationally these expressions are much more difficult than (2.11) and (2.12). Using these expressions for $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}^2(\boldsymbol{\theta})$ we could derive the profiled likelihood or log-likelihood.

We present these expressions for completeness only. We prefer to use the expressions from the pseudo-data representation for computation, especially when the pseudo-data representation is combined with orthogonal-triangular decompositions described in the next section.

2.2.2 *Orthogonal-Triangular Decompositions*

Orthogonal-triangular decompositions of rectangular matrices are a preferred numerical method for solving least squares problems (Chambers, 1977; Kennedy and Gentle, 1980; Thisted, 1988). They are also called *QR decompositions* as the decomposition is often written

$$\boldsymbol{X} = \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}\\ \mathbf{0}\end{bmatrix} = \boldsymbol{Q}_t\boldsymbol{R},$$

where $\boldsymbol{X}$ is an $n \times p$ matrix ($n \geq p$) of rank p, $\boldsymbol{Q}$ is $n \times n$ and orthogonal, $\boldsymbol{R}$ is $p \times p$ and upper triangular, and $\boldsymbol{Q}_t$ ($\boldsymbol{Q}$-truncated) consists of the first p columns of $\boldsymbol{Q}$. To say that $\boldsymbol{Q}$ is *orthogonal* means that $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$. This implies that $\boldsymbol{Q}_t^T\boldsymbol{Q}_t = \boldsymbol{I}$.

The S function `qr` is used to create a QR decomposition from a matrix. For example, in §1.4.1 we present a model where the fixed-effects model matrices for each subject are

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix}, \quad i = 1, \ldots, 11.$$

We can generate such a matrix in S and create its decomposition by

```
> Xmat <- matrix( c(1, 1, 1, 1, 8, 10, 12, 14), ncol = 2 )
> Xmat
     [,1] [,2]
[1,]    1    8
[2,]    1   10
[3,]    1   12
[4,]    1   14
> Xqr <- qr( Xmat )                    # creates a QR structure
> qr.R( Xqr )                          # returns R
     [,1]     [,2]
[1,]   -2 -22.0000
[2,]    0  -4.4721
> qr.Q( Xqr )                          # returns Q-truncated
     [,1]     [,2]
[1,] -0.5  0.67082
[2,] -0.5  0.22361
[3,] -0.5 -0.22361
[4,] -0.5 -0.67082
> qr.Q( Xqr, complete = TRUE )         # returns the full Q
     [,1]     [,2]      [,3]     [,4]
[1,] -0.5  0.67082  0.023607  0.54721
[2,] -0.5  0.22361 -0.439345 -0.71202
[3,] -0.5 -0.22361  0.807869 -0.21760
[4,] -0.5 -0.67082 -0.392131  0.38240
```

Although we will write expressions that involve $\boldsymbol{Q}$, this matrix is not usually evaluated explicitly. Products such as $\boldsymbol{Q}^T\boldsymbol{y}$ or $\boldsymbol{Q}\boldsymbol{y}$ can be calculated directly from information about the decomposition without having to generate this $n \times n$ matrix. See Dongarra, Bunch, Moler and Stewart (1979, Chapter 9) for details. The S functions `qr.qty` and `qr.qy` evaluate these products directly.

An important property of orthogonal matrices is that they preserve norms of vectors under multiplication either by $\boldsymbol{Q}$ or by $\boldsymbol{Q}^T$. That is,

the transformation represented by $\boldsymbol{Q}$ is a generalization of a rotation or a reflection in the plane. In particular,

$$\|\boldsymbol{Q}^T\boldsymbol{y}\|^2 = (\boldsymbol{Q}^T\boldsymbol{y})^T\boldsymbol{Q}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y} = \|\boldsymbol{y}\|^2.$$

If we apply this to the residual vector in a least squares problem we get

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 &= \left\|\boldsymbol{Q}^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right\|^2 \\
&= \|\boldsymbol{Q}^T\boldsymbol{y} - \boldsymbol{Q}^T\boldsymbol{X}\boldsymbol{\beta}\|^2 \\
&= \left\|\boldsymbol{c} - \boldsymbol{Q}^T\boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{bmatrix}\boldsymbol{\beta}\right\|^2 \\
&= \left\|\boldsymbol{c} - \begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{bmatrix}\boldsymbol{\beta}\right\|^2 \\
&= \|\boldsymbol{c}_1 - \boldsymbol{R}\boldsymbol{\beta}\|^2 + \|\boldsymbol{c}_2\|^2,
\end{aligned}$$

where $\boldsymbol{c} = \left(\boldsymbol{c}_1^T\boldsymbol{c}_2^T\right)^T = \boldsymbol{Q}^T\boldsymbol{y}$ is the rotated response vector. The components $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are of lengths p and $n-p$, respectively.

Because $\boldsymbol{X}$ has rank p, the $p \times p$ matrix $\boldsymbol{R}$ is nonsingular and upper-triangular. The least-squares solution $\widehat{\boldsymbol{\beta}}$ is easily evaluated as the solution to

$$\boldsymbol{R}\widehat{\boldsymbol{\beta}} = \boldsymbol{c}_1$$

and the residual sum-of-squares is $\|\boldsymbol{c}_2\|^2$. Notice that the residual sum-of-squares can be evaluated without having to calculate $\widehat{\boldsymbol{\beta}}$.

2.2.3 Evaluating the Likelihood Through Decompositions

Returning to the linear mixed-effects model, we take an orthogonal-triangular decomposition of the augmented model matrix $\tilde{\boldsymbol{Z}}_i$ from (2.7) as

$$\tilde{\boldsymbol{Z}}_i = \boldsymbol{Q}_{(i)}\begin{bmatrix}\boldsymbol{R}_{11(i)}\\ \boldsymbol{0}\end{bmatrix},$$

where $\boldsymbol{Q}_{(i)}$ is $(n_i+q)\times(n_i+q)$ and $\boldsymbol{R}_{11(i)}$ is $q \times q$. Then

$$\begin{aligned}
\left\|\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\right\|^2 &= \left\|\boldsymbol{Q}_{(i)}^T\left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} - \tilde{\boldsymbol{Z}}_i\boldsymbol{b}_i\right)\right\|^2 \\
&= \left\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\right\|^2 + \left\|\boldsymbol{c}_{0(i)} - \boldsymbol{R}_{00(i)}\boldsymbol{\beta}\right\|^2,
\end{aligned}$$

where the $q \times p$ matrix $\boldsymbol{R}_{10(i)}$, the $n_i \times p$ matrix $\boldsymbol{R}_{00(i)}$, the q-vector $\boldsymbol{c}_{1(i)}$ and the n_i-vector $\boldsymbol{c}_{0(i)}$ are defined by

$$\begin{bmatrix}\boldsymbol{R}_{10(i)}\\ \boldsymbol{R}_{00(i)}\end{bmatrix} = \boldsymbol{Q}_{(i)}^T\tilde{\boldsymbol{X}}_i \quad \text{and} \quad \begin{bmatrix}\boldsymbol{c}_{1(i)}\\ \boldsymbol{c}_{0(i)}\end{bmatrix} = \boldsymbol{Q}_{(i)}^T\tilde{\boldsymbol{y}}_i.$$

Another way of thinking of these matrices is as components in an orthogonal-triangular decomposition of an augmented matrix

$$\begin{bmatrix} \boldsymbol{Z}_i & \boldsymbol{X}_i & \boldsymbol{y}_i \\ \boldsymbol{\Delta} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_{(i)} \begin{bmatrix} \boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)} \\ \boldsymbol{0} & \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)} \end{bmatrix},$$

where the reduction to triangular form is halted after the first q columns. (The peculiar numbering scheme for the submatrices and subvectors is designed to allow easy extension to more than one level of random effects.) Returning to the integral in (2.6) we can now remove a constant factor and reduce it to

$$\begin{aligned} &\int \frac{\exp\left[-\left(\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i\|^2 + \|\boldsymbol{\Delta}\boldsymbol{b}_i\|^2\right)/2\sigma^2\right]}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{b}_i \\ &= \exp\left[\frac{\|\boldsymbol{c}_{0(i)} - \boldsymbol{R}_{00(i)}\boldsymbol{\beta}\|^2}{-2\sigma^2}\right] \int \frac{\exp\left[\frac{\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\|^2}{-2\sigma^2}\right]}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{b}_i. \end{aligned} \tag{2.15}$$

Because $\boldsymbol{R}_{11(i)}$ is nonsingular, we can perform a change of variable to $\boldsymbol{\phi}_i = (\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i)/\sigma$ with differential $d\boldsymbol{\phi}_i = \sigma^{-q} \operatorname{abs}|\boldsymbol{R}_{11(i)}|\, d\boldsymbol{b}_i$ and write the integral as

$$\begin{aligned} &\int \frac{\exp\left(-\|\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\boldsymbol{\beta} - \boldsymbol{R}_{11(i)}\boldsymbol{b}_i\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{b}_i \\ &\qquad = \frac{1}{\operatorname{abs}|\boldsymbol{R}_{11(i)}|} \int \frac{\exp\left(-\|\boldsymbol{\phi}_i\|^2/2\right)}{(2\pi)^{q/2}}\, d\boldsymbol{\phi}_i \\ &\qquad = 1/\operatorname{abs}|\boldsymbol{R}_{11(i)}|. \end{aligned} \tag{2.16}$$

This is the same result as (2.10) because

$$\begin{aligned} \sqrt{\left|\tilde{\boldsymbol{Z}}_i^T \tilde{\boldsymbol{Z}}_i\right|} &= \sqrt{\left|\left[\boldsymbol{R}_{11(i)}^T \boldsymbol{0}\right] \boldsymbol{Q}_{(i)}^T \boldsymbol{Q}_{(i)} \begin{bmatrix} \boldsymbol{R}_{11(i)} \\ \boldsymbol{0} \end{bmatrix}\right|} \\ &= \sqrt{\left|\boldsymbol{R}_{11(i)}^T \boldsymbol{R}_{11(i)}\right|} \\ &= \sqrt{\left|\boldsymbol{R}_{11(i)}^T\right| \left|\boldsymbol{R}_{11(i)}\right|} \\ &= \sqrt{\left(\left|\boldsymbol{R}_{11(i)}^T\right|\right)^2} \\ &= \operatorname{abs}|\boldsymbol{R}_{11(i)}|. \end{aligned}$$

Because $\boldsymbol{R}_{11(i)}$ is triangular, its determinant is simply the product of its diagonal elements.

Substituting (2.16) into (2.15) into (2.6) provides the likelihood as

$$L\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y}\right)=\prod_{i=1}^{M}\frac{\exp\left[-\left\|\boldsymbol{c}_{0(i)}-\boldsymbol{R}_{00(i)}\boldsymbol{\beta}\right\|^2/2\sigma^2\right]}{\left(2\pi\sigma^2\right)^{n_i/2}}\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right)$$
$$=\frac{\exp\left(-\sum_{i=1}^{M}\left\|\boldsymbol{c}_{0(i)}-\boldsymbol{R}_{00(i)}\boldsymbol{\beta}\right\|^2/2\sigma^2\right)}{\left(2\pi\sigma^2\right)^{N/2}}\prod_{i=1}^{M}\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right).$$

The term in the exponent has the form of a residual sum-of-squares for $\boldsymbol{\beta}$ pooled over all the groups. Forming another orthogonal-triangular decomposition

$$\begin{bmatrix}\boldsymbol{R}_{00(1)} & \boldsymbol{c}_{0(1)}\\ \vdots & \vdots\\ \boldsymbol{R}_{00(M)} & \boldsymbol{c}_{0(M)}\end{bmatrix}=\boldsymbol{Q}_0\begin{bmatrix}\boldsymbol{R}_{00} & \boldsymbol{c}_0\\ \boldsymbol{0} & \boldsymbol{c}_{-1}\end{bmatrix} \tag{2.17}$$

produces the reduced form

$$L\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y}\right)$$
$$=\left(2\pi\sigma^2\right)^{-N/2}\exp\left(\frac{\|\boldsymbol{c}_{-1}\|^2+\|\boldsymbol{c}_0-\boldsymbol{R}_{00}\boldsymbol{\beta}\|^2}{-2\sigma^2}\right)\prod_{i=1}^{M}\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right). \tag{2.18}$$

For a given $\boldsymbol{\theta}$, the values of $\boldsymbol{\beta}$ and σ^2 that maximize (2.18) are

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})=\boldsymbol{R}_{00}^{-1}\boldsymbol{c}_0 \quad\text{and}\quad \widehat{\sigma}^2(\boldsymbol{\theta})=\frac{\|\boldsymbol{c}_{-1}\|^2}{N}, \tag{2.19}$$

which give the profiled likelihood

$$L(\boldsymbol{\theta}|\boldsymbol{y})=L\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta},\widehat{\sigma}^2(\boldsymbol{\theta})|\boldsymbol{y}\right)$$
$$=\left(\frac{N}{2\pi\|\boldsymbol{c}_{-1}\|^2}\right)^{N/2}\exp\left(-\frac{N}{2}\right)\prod_{i=1}^{M}\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right), \tag{2.20}$$

or the profiled log-likelihood

$$\ell(\boldsymbol{\theta}|\boldsymbol{y})=\log L(\boldsymbol{\theta}|\boldsymbol{y})$$
$$=\frac{N}{2}\left[\log N-\log(2\pi)-1\right]-N\log\|\boldsymbol{c}_{-1}\|+\sum_{i=1}^{M}\log\operatorname{abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right). \tag{2.21}$$

The profiled log-likelihood (2.21) is maximized with respect to $\boldsymbol{\theta}$, producing the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$. The maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are then obtained by setting $\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}$ in (2.19).

Although technically the random effects $\boldsymbol{b}_i$ are not parameters for the statistical model, they do behave in some ways like parameters and often we want to "estimate" their values. The conditional modes of the random effects, evaluated at the conditional estimate of $\boldsymbol{\beta}$, are the *Best Linear Unbiased Predictors* or *BLUPs* of the $\boldsymbol{b}_i, i = 1, \ldots, M$. They can be evaluated, using the matrices from the orthogonal-triangular decompositions, as

$$\widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}) = \boldsymbol{R}_{11(i)}^{-1}\left(\boldsymbol{c}_{1(i)} - \boldsymbol{R}_{10(i)}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right). \tag{2.22}$$

In practice, the unknown vector $\boldsymbol{\theta}$ is replaced by its maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$, producing estimated BLUPs $\widehat{\boldsymbol{b}}_i(\widehat{\boldsymbol{\theta}})$.

The decomposition (2.17) is equivalent to calculating the QR decomposition of the potentially huge matrix $\boldsymbol{X}_e$ defined in (2.11). If we determined the least-squares solution to (2.11) using an orthogonal-triangular decomposition

$$\boldsymbol{X}_e = \boldsymbol{Q}_e \begin{bmatrix} \boldsymbol{R}_e \\ \boldsymbol{0} \end{bmatrix},$$

the triangular part of the decomposition and the leading part of the rotated, augmented response vector would be

$$\boldsymbol{R}_e = \begin{bmatrix} \boldsymbol{R}_{11(1)} & \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{R}_{10(1)} \\ \boldsymbol{0} & \boldsymbol{R}_{11(2)} & \ldots & \boldsymbol{0} & \boldsymbol{R}_{10(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{R}_{11(M)} & \boldsymbol{R}_{10(M)} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R}_{00} \end{bmatrix} \quad \text{and} \quad \boldsymbol{c}_1 = \begin{bmatrix} \boldsymbol{c}_{1(1)} \\ \boldsymbol{c}_{1(2)} \\ \vdots \\ \boldsymbol{c}_{1(M)} \\ \boldsymbol{c}_0 \end{bmatrix}.$$

Thus, the $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}^2(\boldsymbol{\theta})$ from (2.19) are the same as those from (2.17) and (2.12). The vector $\boldsymbol{c}_{-1}$ is the residual vector in the coordinate system determined by $\boldsymbol{Q}_e$. Because $\boldsymbol{Q}_e$ is orthogonal, $\|\boldsymbol{c}_{-1}\|^2$ is the residual sum-of-squares for the least squares problem defined by $\boldsymbol{X}_e$ and $\boldsymbol{y}_e$.

The profiled log-likelihood (2.20) has the same form as (2.13). It consists of three additive components; a constant, a scaled logarithm of the residual sum-of-squares, and a sum of ratios of the logarithms of determinants. In the next section we examine these terms in detail.

2.2.4 Components of the Profiled Log-Likelihood

Returning to the example of the rails data of §1.1, let us consider the different components of the profiled log-likelihood as expressed in (2.21). Recall that the relative precision factor $\boldsymbol{\Delta}$ will be a scalar in this case so let us write it as Δ. There are three additive terms in the profiled log-likelihood:

1. The constant $\frac{N}{2}\left[\log N - \log(2\pi) - 1\right]$, which can be neglected for the purposes of optimization.
2. $-N \log \|\boldsymbol{c}_{-1}\|$, a multiple of the logarithm of the norm of a residual vector from the penalized least-squares fit for Δ, $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, and $\boldsymbol{y}_i$.
3. $\sum_{i=1}^{M} \log\left(\Delta / \operatorname{abs}|\boldsymbol{R}_{11(i)}|\right) = \sum_{i=1}^{M} \log\left(\Delta / \sqrt{|\boldsymbol{Z}_i^T \boldsymbol{Z}_i + \Delta^2|}\right)$. In the general case this is the sum of the logarithms of the ratios of determinants.

In Figure 2.1 we show the two nonconstant terms and the resulting log-likelihood as a function of Δ for the rails example.

The shapes of the curves in Figure 2.1 indicate that it would be better to optimize the profiled log-likelihood with respect to $\theta = \log \Delta$ instead of Δ. This transformation will also help to ensure that Δ does not become negative during the course of the iterations of whatever optimization routine we use. In Figure 2.2 we show the components and the log-likelihood as a function of θ. We can see that the log-likelihood is closer to a quadratic with respect to θ than with respect to Δ.

There are patterns in Figure 2.2 that will hold in general for linear mixed-effects models. The log of the norm of the residual is an increasing sigmoidal, or "S-shaped," function with respect to θ. As $\theta \to -\infty$ (or $\Delta \to 0$), this log-norm approaches a horizontal asymptote at a value that corresponds to the log residual norm from an unpenalized regression of the form

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{X}_1 \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \dots & \boldsymbol{0} & \boldsymbol{X}_2 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{Z}_n & \boldsymbol{X}_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\epsilon}.$$

At the other extreme, large positive values of θ, and the correspondingly large values of Δ, put such a heavy penalty on the size of the b_i terms in the regression that these are forced to zero. Thus, as $\theta \to \infty$, the penalized residual norm approaches that from a regression of the entire response vector $\boldsymbol{y}$ on the $\boldsymbol{X}_i$ matrices alone.

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In the ratio of determinants term, very large values of Δ will dominate $\boldsymbol{Z}_i^T \boldsymbol{Z}_i$ in the denominator so the ratios approach Δ/Δ and the sum of the

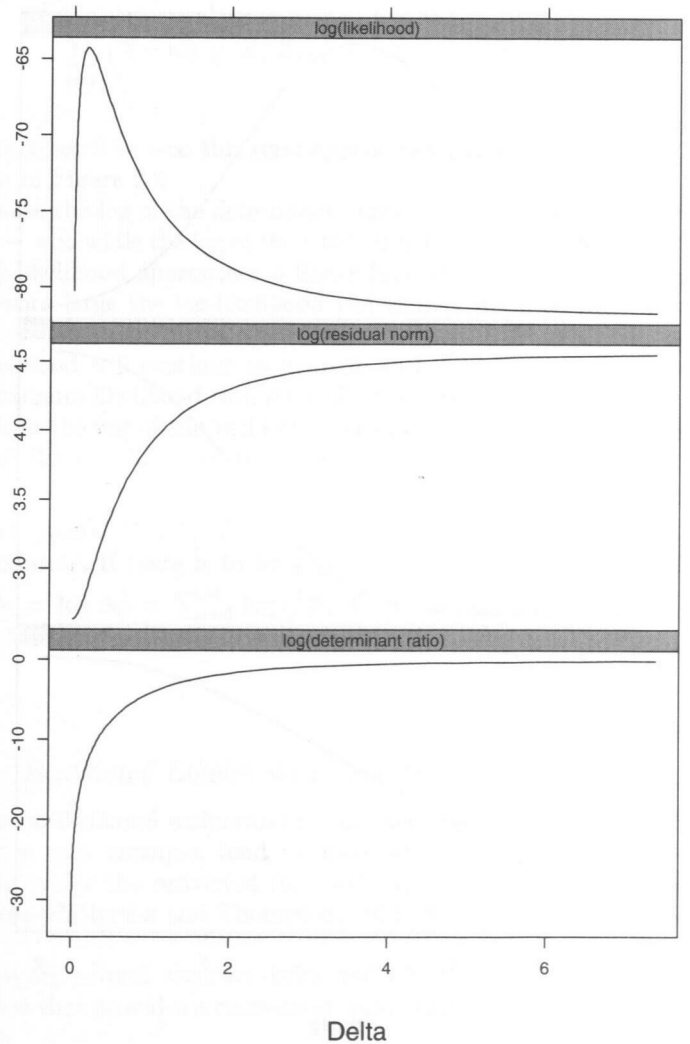

FIGURE 2.1. The profiled log-likelihood as a function of Δ for the rails example. Two of the components of the log-likelihood, $\log \|\mathbf{c}_{-1}\|$, the log of the length of the residual, and $\sum_{i=1}^{M} \log\left(\Delta / \operatorname{abs}|\boldsymbol{R}_{11(i)}|\right)$, the log of the determinant ratios, are shown on the same scale.

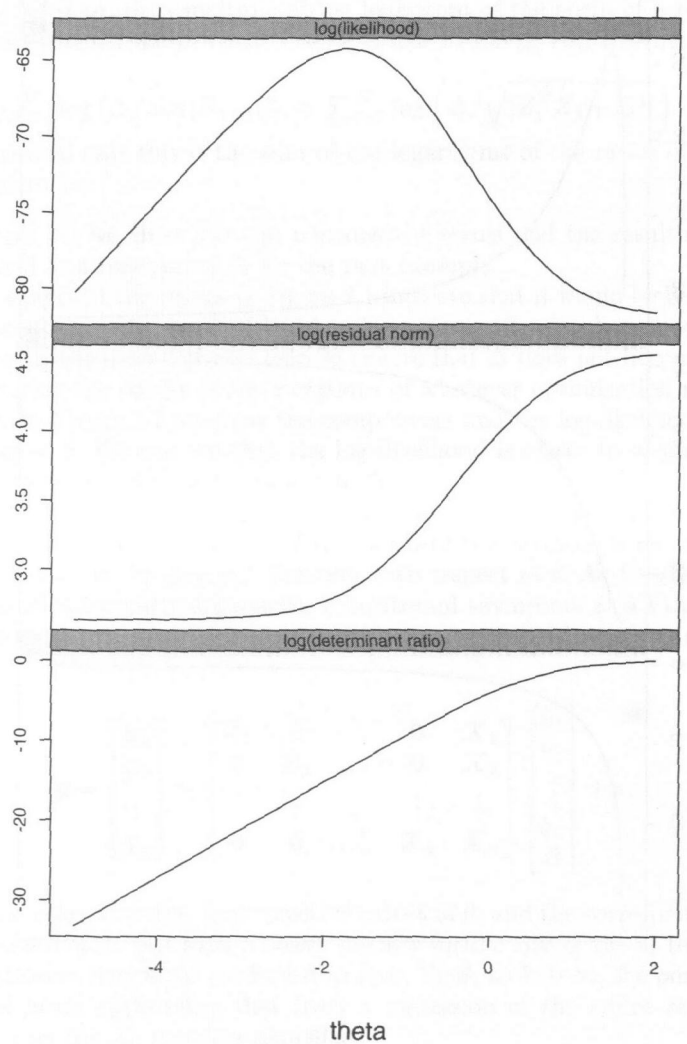

FIGURE 2.2. The profiled log-likelihood as a function of $\theta = \log(\Delta)$ for the rails example. Two of the components of the log-likelihood, $\log \|\boldsymbol{c}_{-1}\|$, the log of the length of the residual, and $\sum_{i=1}^{M} \log\left(\Delta / \operatorname{abs}|\boldsymbol{R}_{11(i)}|\right)$, the log of the determinant ratios, are shown on the same scale.

logarithms approaches zero. Very small values of Δ will have little effect on the denominator so the term has the form

$$\sum_{i=1}^{M}\left(\theta - \log\sqrt{|\boldsymbol{Z}_i^T\boldsymbol{Z}_i|}\right) = M\theta - \sum_{i=1}^{M}\log\sqrt{|\boldsymbol{Z}_i^T\boldsymbol{Z}_i|}.$$

That is, when $\theta \to -\infty$ this term approaches a linear function of θ, as can be seen in Figure 2.2.

Because the log of the determinant ratio approaches a linear function of θ as $\theta \to -\infty$ while the log of the residual norm tends to a finite asymptote, the log-likelihood approaches a linear function of θ. When Δ, and hence θ, becomes large the log-likelihood will usually decrease then approach a constant. This does not always occur, however. For some data sets, the log-likelihood will continue to increase with θ as $\theta \to \infty$. In these cases, the maximum likelihood estimator of σ_b^2 is zero.

Both in the log of the ratio of determinants term and in the log of the norm of the penalized residual term, the effect of Δ is determined by its size relative to the $\boldsymbol{Z}_i$ matrices. Values of Δ that are either much less than or much greater than $\sqrt{\boldsymbol{Z}_i^T\boldsymbol{Z}_i}$ will produce a log-likelihood that is near an asymptote. If there is to be a maximum for finite θ it will have to be near $\theta_0 = \log\Delta_0 = \sum_{i=1}^{M}\log\sqrt{\boldsymbol{Z}_i^T\boldsymbol{Z}_i}/M$. In the case of the rails data $\theta_0 = 0.549$.

2.2.5 Restricted Likelihood Estimation

Maximum likelihood estimates of "variance components," such as σ^2 and σ_b^2 in the rails example, tend to underestimate these parameters. Many analysts prefer the restricted (or residual) maximum likelihood (REML) estimates (Patterson and Thompson, 1971; Harville, 1977) for these quantities.

There are several ways to define the REML estimation criterion. One definition that provides a convenient computational form (Laird and Ware, 1982) is

$$L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \int L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{\beta},$$

which, within a Bayesian framework, corresponds to assuming a locally uniform prior distribution for the fixed effects $\boldsymbol{\beta}$ and integrating them out of the likelihood.

Using (2.18) and the same change-of-variable techniques as in (2.16) gives the log-restricted-likelihood

$$\begin{aligned}\ell_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) &= \log L_R\left(\boldsymbol{\theta}, \sigma^2|\,\boldsymbol{y}\right) \\ &= -\frac{N-p}{2}\log(2\pi\sigma^2) - \frac{\|\boldsymbol{c}_{-1}\|^2}{2\sigma^2} - \operatorname{log\,abs}|\boldsymbol{R}_{00}| + \sum_{i=1}^{M}\operatorname{log\,abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right).\end{aligned}$$

This produces the conditional estimate $\widehat{\sigma}^2_R(\theta) = \|\boldsymbol{c}_{-1}\|^2/(N-p)$ for σ^2, from which we obtain the profiled log-restricted-likelihood

$$\begin{aligned}\ell_R(\boldsymbol{\theta}|\boldsymbol{y}) &= \ell_R(\boldsymbol{\theta}, \widehat{\sigma}^2_R(\boldsymbol{\theta})|\boldsymbol{y}) \\ &= \text{const} - (N-p)\log\|\boldsymbol{c}_{-1}\| - \operatorname{log\,abs}|\boldsymbol{R}_{00}| + \sum_{i=1}^{M}\operatorname{log\,abs}\left(\frac{|\boldsymbol{\Delta}|}{|\boldsymbol{R}_{11(i)}|}\right).\end{aligned} \tag{2.23}$$

The components of the profiled log-restricted-likelihood in (2.23) are similar to those in the profiled log-likelihood (2.21) except that the log of the norm of the residual vector has a different multiplier and there is an extra determinant term of $\operatorname{log\,abs}|\boldsymbol{R}_{00}| = \log\left|\sum_{i=1}^{M}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right|/2$. A plot of the components of the profiled log-restricted-likelihood versus θ for the rails example would be similar in shape to Figure 2.2.

The evaluation of the restricted maximum likelihood estimates is done by optimizing the profiled log-restricted-likelihood (2.23) with respect to $\boldsymbol{\theta}$ only, and using the resulting REML estimate $\widehat{\boldsymbol{\theta}}_R$ to obtain the REML estimate of σ^2, $\widehat{\sigma}^2_R(\widehat{\boldsymbol{\theta}}_R)$. Similarly, the REML estimated BLUPs of the random effects are obtained by replacing $\boldsymbol{\theta}$ with $\widehat{\boldsymbol{\theta}}_R$ in (2.22).

In some ways, it is blurring the definition of the REML criterion to speak of the REML estimate of $\boldsymbol{\beta}$. The REML criterion only depends on $\boldsymbol{\theta}$ and σ. However, it is still useful, and perhaps even sensible, to evaluate the "best guess" at $\boldsymbol{\beta}$ from (2.19) once $\widehat{\boldsymbol{\theta}}_R$ has been determined using the REML criterion.

An important difference between the likelihood function and the restricted likelihood function is that the former is invariant to one-to-one reparameterizations of the fixed effects (e.g., a change in the contrasts representing a categorical variable), while the latter is not. Changing the $\boldsymbol{X}_i$ matrices results in a change in $\operatorname{log\,abs}|\boldsymbol{R}_{00}|$ and a corresponding change in $\ell_R(\boldsymbol{\theta}|\boldsymbol{y})$. As a consequence, LME models with different fixed-effects structures fit using REML cannot be compared on the basis of their restricted likelihoods. In particular, likelihood ratio tests are not valid under these circumstances.

2.2.6 *Multiple Levels of Random Effects*

The likelihood function and the restricted likelihood function for multilevel LME models can be calculated using the same techniques described for the single-level model in §2.2.1, §2.2.3, and §2.2.5. We use the two-level LME model to illustrate the basic steps in the derivation of the multilevel likelihood function.

The likelihood for a model with two levels of random effects is defined as in (2.3), but integrating over both levels of random effects

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2 | \boldsymbol{y}) = \prod_{i=1}^{M} \int \prod_{j=1}^{M_i} \left[\int p(\boldsymbol{y}_{ij} | \boldsymbol{b}_{ij}, \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{b}_{ij} | \boldsymbol{\theta}_2, \sigma^2)\, d\boldsymbol{b}_{ij} \right] p(\boldsymbol{b}_i | \boldsymbol{\theta}_1, \sigma^2)\, d\boldsymbol{b}_i. \tag{2.24}$$

As with a single level of random effects, we can simplify the integrals in (2.24) if we augment the $\boldsymbol{Z}_{ij}$ matrices with $\boldsymbol{\Delta}_2$ and form orthogonal-triangular decompositions of these augmented arrays. This allows us to evaluate the inner integrals. To evaluate the outer integrals we iterate this process.

That is, we first form and decompose the arrays

$$\begin{bmatrix} \boldsymbol{Z}_{ij} & \boldsymbol{Z}_{i,j} & \boldsymbol{X}_{ij} & \boldsymbol{y}_{ij} \\ \boldsymbol{\Delta}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} = \boldsymbol{Q}_{(ij)} \begin{bmatrix} \boldsymbol{R}_{22(ij)} & \boldsymbol{R}_{21(ij)} & \boldsymbol{R}_{20(ij)} & \boldsymbol{c}_{2(ij)} \\ \mathbf{0} & \boldsymbol{R}_{11(ij)} & \boldsymbol{R}_{10(ij)} & \boldsymbol{c}_{1(ij)} \end{bmatrix}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, M_i. \tag{2.25}$$

The matrix $\boldsymbol{R}_{22(ij)}$ will be an upper-triangular matrix of dimension $q_2 \times q_2$. The other arrays in the first row of the decomposition in (2.25) are used only if the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ or the conditional modes $\widehat{\boldsymbol{b}}_{ij}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{b}}_i(\boldsymbol{\theta})$ are required. The arrays in the second row of the decomposition: $\boldsymbol{R}_{11(ij)}$, $\boldsymbol{R}_{10ij}$, and $\boldsymbol{c}_{1(ij)}$ each have n_{ij} rows.

To evaluate the outer integral in (2.24) we again form and decompose an augmented array

$$\begin{bmatrix} \boldsymbol{R}_{11(i1)} & \boldsymbol{R}_{10(i1)} & \boldsymbol{c}_{1(i1)} \\ \vdots & \vdots & \vdots \\ \boldsymbol{R}_{11(iM_i)} & \boldsymbol{R}_{10(iM_i)} & \boldsymbol{c}_{1(iM_i)} \\ \boldsymbol{\Delta}_1 & \mathbf{0} & \mathbf{0} \end{bmatrix} = \boldsymbol{Q}_{(i)} \begin{bmatrix} \boldsymbol{R}_{11(i)} & \boldsymbol{R}_{10(i)} & \boldsymbol{c}_{1(i)} \\ \mathbf{0} & \boldsymbol{R}_{00(i)} & \boldsymbol{c}_{0(i)} \end{bmatrix}, \quad i = 1, \ldots, M. \tag{2.26}$$

The final decomposition to produce $\boldsymbol{R}_{00}$, $\boldsymbol{c}_0$, and $\boldsymbol{c}_{-1}$ is the same as that in (2.17).

Using the matrices and vectors produced in (2.25), (2.26), and (2.17) and following the same steps as for the single level of nesting we can express

the profiled log-likelihood for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ as

$$\begin{aligned}\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{y}) &= \log L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \widehat{\sigma}^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|\boldsymbol{y}) \\ &= \text{const} - N \log \|\boldsymbol{c}_{-1}\| + \sum_{i=1}^{M} \log \text{abs}\left(\frac{|\boldsymbol{\Delta}_1|}{|\boldsymbol{R}_{11(i)}|}\right) \\ &\quad + \sum_{i=1}^{M}\sum_{j=1}^{M_i} \log \text{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(ij)}|}\right).\end{aligned}$$

Similarly, the profiled log-restricted-likelihood is

$$\begin{aligned}\ell_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{y}) &= \log L_R(\widehat{\boldsymbol{\beta}}_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \widehat{\sigma}^2_R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|\boldsymbol{y}) \\ &= \text{const} - (N-p) \log \|\boldsymbol{c}_{-1}\| - \log \text{abs}\,|\boldsymbol{R}_{00}| \\ &\quad + \sum_{i=1}^{M} \log \text{abs}\left(\frac{|\boldsymbol{\Delta}_1|}{|\boldsymbol{R}_{11(i)}|}\right) + \sum_{i=1}^{M}\sum_{j=1}^{M_i} \log \text{abs}\left(\frac{|\boldsymbol{\Delta}_2|}{|\boldsymbol{R}_{22(ij)}|}\right).\end{aligned}$$

The calculation methods extend in the obvious way to Q nested levels of random effects.

2.2.7 Parameterizing Relative Precision Factors

In a model with Q nested levels of random effects, there are Q symmetric, positive-definite, variance–covariance matrices $\boldsymbol{\Psi}_k$, $k = 1, \ldots, Q$. For computational purposes we express these in terms of relative precision factors $\boldsymbol{\Delta}_k, k = 1, \ldots, Q$ that satisfy

$$\boldsymbol{\Delta}_k^T \boldsymbol{\Delta}_k = \sigma^2 \boldsymbol{\Psi}_k^{-1}, \quad k = 1, \ldots, Q.$$

To optimize the log-likelihood or log-restricted-likelihood we express the scaled variance–covariance matrices $\boldsymbol{\Psi}_k/\sigma^2$, or equivalently the relative precision factors $\boldsymbol{\Delta}_k$, as a function of unconstrained parameter vectors $\boldsymbol{\theta}_k, k = 1, \ldots, Q$. For example, when Δ is a scalar, as in §2.2.4, we use the unconstrained parameter $\theta = \log \Delta$ when optimizing the log-likelihood.

For the general case where $\boldsymbol{\Psi}_k/\sigma^2$ is a positive-definite, symmetric matrix of size $q \times q$, we parameterize it through its *matrix logarithm*. To define this parameterization, we note that any positive-definite, symmetric matrix $\boldsymbol{A}$ can be expressed as the *matrix exponential* of another symmetric matrix $\boldsymbol{B}$. This means that

$$\boldsymbol{A} = e^{\boldsymbol{B}} = \boldsymbol{I} + \boldsymbol{B} + \frac{\boldsymbol{B}^2}{2!} + \frac{\boldsymbol{B}^3}{3!} + \ldots.$$

If $\boldsymbol{A}$ is the matrix exponential of $\boldsymbol{B}$, then $\boldsymbol{B}$ is the matrix logarithm of $\boldsymbol{A}$.

Suppose $\boldsymbol{A}$ is $q \times q$, symmetric and positive-definite. One way of evaluating its matrix logarithm $\boldsymbol{B}$ is to calculate an eigenvalue-eigenvector decomposition

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T,$$

where Λ is $q \times q$ and diagonal while U is $q \times q$ and orthogonal. If A is positive-definite, then all the diagonal elements of Λ must be positive. The matrix logarithm of Λ is the diagonal matrix whose diagonal elements are the logarithms of the corresponding elements of Λ. We will denote this by $\log \Lambda$. Finally

$$B = \log A = U \log \Lambda U^T.$$

We define θ_k to be the elements of the upper triangle of the matrix logarithm of Ψ_k/σ^2. This gives a nonredundant, unconstrained parameter vector for Ψ_k/σ^2.

Other unconstrained parameterizations for Ψ_k/σ^2 are used when the matrix is required to have a special structure beyond being symmetric and positive-definite. For example, if Ψ_k/σ^2 is to be diagonal and positive-definite then the diagonal elements must all be positive and an unconstrained parameterization uses the logarithms of these diagonal elements.

2.2.8 Optimization Algorithms

Optimization of the profiled log-likelihood or the profiled log-restricted-likelihood of an LME model is usually accomplished through EM iterations or through Newton–Raphson iterations (Laird and Ware, 1982; Lindstrom and Bates, 1988; Longford, 1993).

The EM algorithm (Dempster, Laird and Rubin, 1977) is a popular iterative algorithm for likelihood estimation in models with incomplete data. The EM iterations for the LME model are based on regarding the random effects, such as the $b_i, i = 1, \ldots, M$, as unobserved data. At iteration w we use the current variance–covariance parameter vector, $\theta^{(w)}$, to evaluate the distribution of $b|y$ and derive the expectation of the log-likelihood for a new value of θ given this conditional distribution. Because we are taking an expectation, this step is called the E step. The M step consists of maximizing this expectation with respect to θ to produce $\theta^{(w+1)}$. Each iteration of the EM algorithm results in an increase in the log-likelihood function, though a possibly small increase. Efficient implementations of the EM algorithm for LME models are described in Bates and Pinheiro (1998).

The Newton–Raphson algorithm (Thisted, 1988, §4.2.2) is one of the most widely used optimization procedures. It uses a first-order expansion of the score function (the gradient of the log-likelihood function) around the current estimate $\theta^{(w)}$ to produce the next estimate $\theta^{(w+1)}$. Each Newton–Raphson iteration requires the calculation of the score function and its derivative, the Hessian matrix of the log-likelihood. Under general conditions usually satisfied in practice, the Newton–Raphson algorithm converges quadratically. Because the calculation of the Hessian matrix at each iteration may be computationally expensive, simple, quicker to compute approximations are sometimes used, leading to the so-called Quasi–Newton algorithms (Thisted, 1988, §4.3.3.4) .

Any iterative optimization algorithm requires initial values for the parameters. Because we can express both the profiled log-likelihood and the profiled log-restricted-likelihood as a function of the $\boldsymbol{\theta}$ parameters, we only need to formulate starting values for $\boldsymbol{\theta}$ when performing iterative optimization for LME models. These may be obtained from a previous fit for similar data, or derived from the current data. A general procedure for deriving initial values for $\boldsymbol{\theta}$ from the data being fit is described in Bates and Pinheiro (1998) and is implemented in the `lme` function.

Individual iterations of the EM algorithm are quickly and easily computed. Although the EM iterations generally bring the parameters into the region of the optimum very quickly, progress toward the optimum tends to be slow when near the optimum. Newton–Raphson iterations, on the other hand, are individually more computationally intensive than the EM iterations, and they can be quite unstable when far from the optimum. However, close to the optimum they converge very quickly.

We therefore recommend a hybrid approach of starting with an initial $\boldsymbol{\theta}^{(0)}$, performing a moderate number of EM iterations, then switching to Newton–Raphson iterations. Essentially the EM iterations can be regarded as refining the starting estimates before beginning the more general optimization routine. The `lme` function implements such a hybrid optimization scheme. It begins by calculating initial estimates of the $\boldsymbol{\theta}$ parameters, then uses several EM iterations to get near the optimum, then switches to Newton–Raphson iterations to complete the convergence to the optimum. By default 25 EM iterations are performed before switching to Newton–Raphson iterations.

When fitting an LME model, it is often helpful to monitor the progress of the Newton–Raphson iterations to identify possible convergence problems. This is done by including an optional `control` argument in the call to `lme`. The value of `control` should be a list that can contain any of several flags or settings for the optimization algorithm. One of these flags is `msVerbose`. When it is set to `TRUE` or `T`, diagnostic output on the progress of the Newton–Raphson iterations in the indirect call of the `ms` function (Bates and Chambers, 1992, §10.2) is produced.

If we set this flag in the first fit for the rails example of §1.1, the diagnostic output is not very interesting because the EM iterations leave the parameter estimates so close to the optimum that convergence of the Newton–Raphson iterations is declared almost immediately.

```
> fm1Rail.lme <- lme( travel ~ 1, data = Rail, random = ~ 1 | Rail,
+        control = list( msVerbose = TRUE ) )
Iteration:  0 ,  1  function calls, F=  61.049
Parameters:
[1] -1.8196
Iteration:  1 ,  2  function calls, F=  61.049
```

```
Parameters:
[1] -1.8196
```

Note that the parameter listed in the iteration output is

$$\widehat{\theta} = \log(\widehat{\Delta}) = \log(\widehat{\sigma}/\widehat{\sigma}_b) = \log(4.0208/24.805) = -1.8196$$

and this is the only parameter being directly controlled by the optimization algorithm. The function labelled `F` in the iteration output is the negative of the log-restricted-likelihood but without the constant term $\frac{N-p}{2}[\log(N-p) - \log(2\pi) - 1]$, which is -0.0396 when $N = 18$ and $p = 1$. Thus, the value of $F = 61.049$ at convergence corresponds to a log-likelihood of $-(61.049 + 0.0396) = -61.089$. Because most optimization algorithms are designed to minimize rather than maximize a function of the parameters, we minimize the negative of the log-likelihood instead of maximizing the log-likelihood.

If we eliminate the EM iterations altogether with another `control` argument, `niterEM`, we can observe the progress of the Newton–Raphson iterations for θ.

```
> fm1Rail.lme <- lme( travel ~ 1, data = Rail, random = ~ 1 | Rail,
+    control = list( msVerbose = TRUE, niterEM = 0 ))
Iteration:  0 ,  1  function calls, F=  67.894
Parameters:
[1] -0.43152
Iteration:  1 ,  3  function calls, F=  61.157
Parameters:
[1] -2.0007
Iteration:  2 ,  4  function calls, F=  61.05
Parameters:
[1] -1.8028
Iteration:  3 ,  5  function calls, F=  61.049
Parameters:
[1] -1.8195
```

The algorithm converged to a slightly different value of θ, but with essentially the same value of the log-likelihood.

2.3 Approximate Distributions

Inference on the parameters of a linear mixed-effects model usually relies on approximate distributions for the maximum likelihood estimates and the restricted maximum likelihood estimates derived from asymptotic results.

Pinheiro (1994) has shown that, under certain regularity conditions generally satisfied in practice, the maximum likelihood estimates in the general LME model described in §2.1 are consistent and asymptotically normal. The approximate variance–covariance matrix for the maximum likelihood estimates is given by the inverse of the information matrix (Cox

and Hinkley, 1974, §4.8) corresponding to the log-likelihood function $\ell = \ell(\boldsymbol{\beta}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q, \sigma^2)$. Because

$$\mathrm{E}\left[\partial^2 \ell / \partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_q^T\right] = \mathbf{0},\ q = 1, \ldots, Q \quad \text{and} \quad \mathrm{E}\left[\partial^2 \ell / \partial \boldsymbol{\beta} \partial \sigma^2\right] = \mathbf{0},$$

the information matrix corresponding to an LME model with Q levels of nesting is block diagonal and, therefore, the maximum likelihood estimates of the fixed effects $\boldsymbol{\beta}$ are asymptotically uncorrelated with the maximum likelihood estimates of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ and σ^2.

The approximate distributions of the maximum likelihood estimates in an LME model with Q levels of nesting are

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left[\boldsymbol{R}_{00}^{-1} \boldsymbol{R}_{00}^{-T}\right]\right),$$

$$\begin{bmatrix} \widehat{\boldsymbol{\theta}}_1 \\ \vdots \\ \widehat{\boldsymbol{\theta}}_Q \\ \log \widehat{\sigma} \end{bmatrix} \dot{\sim} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_Q \\ \log \sigma \end{bmatrix}, \ \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q, \sigma)\right), \tag{2.27}$$

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q, \sigma) = -\begin{bmatrix} \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T} & \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_1^T} & \cdots & \frac{\partial^2 \ell}{\partial \log \sigma \partial \boldsymbol{\theta}_1^T} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_1 \partial \log \sigma} & \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}_2 \partial \log \sigma} & \cdots & \frac{\partial^2 \ell}{\partial^2 \log \sigma} \end{bmatrix},$$

where $\ell = \ell(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q, \sigma^2)$ now denotes the log-likelihood function profiled on the fixed effects, $\boldsymbol{\mathcal{I}}$ denotes the *empirical* information matrix, and $\boldsymbol{R}_{00}$ is defined as in (2.17). We use $\log \sigma$ in place of σ^2 in (2.27) to give an unrestricted parameterization for which the normal approximation tends to be more accurate.

As shown by Pinheiro (1994), the REML estimates in an LME model also are consistent and asymptotically normal, with approximate distributions identical to (2.27) but with ℓ replaced by the log-restricted-likelihood ℓ_R defined in §2.2.5.

In practice, the unknown parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ and σ^2 are replaced by their respective ML or REML estimates in the expressions for the approximate variance–covariance matrices in (2.27). The approximate distributions for the maximum likelihood estimates and REML estimates are used to produce hypothesis tests and confidence intervals for the LME model parameters, as described in §2.4.

2.4 Hypothesis Tests and Confidence Intervals

After we have fit a statistical model to the data we usually want to assess the precision of the estimates and the "significance" of various terms in the

model or to compare how well one model fits the data relative to another model. This section presents approximate hypothesis tests and confidence intervals for the parameters in an LME model.

2.4.1 *Likelihood Ratio Tests*

A general method for comparing nested models fit by maximum likelihood is the *likelihood ratio test* (Lehmann, 1986, §1.7). Such a test can also be used with models fit by REML, but only if both models have been fit by REML and if the fixed-effects specification is the same for both models.

One statistical model is said to be *nested* within another model if it represents a special case of the other model. For example, in the analysis of the `Machines` data described in §1.3, we fit one model, `fm1Machine`, with a random effect for `Worker` only, then we fit a second model, `fm2Machine`, with random effects for `Worker` and for `Machine %in% Worker`. The model `fm1Machine` is nested within `fm2Machine` because it represents a special case of `fm2Machine` in which the variance of the `Machine %in% Worker` interaction term is zero.

If L_2 is the likelihood of the more general model (e.g., `fm2Machine`) and L_1 is the likelihood of the restricted model (e.g., `fm1Machine`) we must have $L_2 > L_1$ and, correspondingly, $\log L_2 > \log L_1$. The likelihood ratio test (LRT) statistic

$$2\log(L_2/L_1) = 2[\log(L_2) - \log(L_1)]$$

will be positive. If k_i is the number of parameters to be estimated in model i, then the asymptotic, or "large sample," distribution of the LRT statistic, under the null hypothesis that the restricted model is adequate, is a χ^2 distribution with $k_2 - k_1$ degrees of freedom.

In Chapter 1 we show several examples of likelihood ratio tests performed with the `anova` function. When given two arguments representing fits of nested models, this function displays the LRT statistic in the `L.Ratio` column and gives the p-value from the $\chi^2_{k_2-k_1}$ distribution. The column labelled `df` is the number of parameters in each model. For example, using model fits described in §1.3, we would have

```
> anova( fm1Machine, fm2Machine )
           Model df    AIC    BIC  logLik   Test L.Ratio p-value
fm1Machine     1  5 300.46 310.12 -145.23
fm2Machine     2  6 231.27 242.86 -109.64 1 vs 2  71.191  <.0001
```

The `anova` function also displays the values of the *Akaike Information Criterion* (*AIC*) (Sakamoto et al., 1986) and the *Bayesian Information Criterion* (*BIC*) (Schwarz, 1978). As mentioned in §1.1.1, these are model

comparison criteria evaluated as

$$\begin{aligned} AIC &= -2\ell\left(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}\right) + 2n_{par}, \\ BIC &= -2\ell\left(\widehat{\boldsymbol{\theta}}|\boldsymbol{y}\right) + n_{par}\log(N) \end{aligned} \tag{2.28}$$

for each model, where n_{par} denotes the number of parameters in the model. Under these definitions, "smaller is better." That is, if we are using AIC to compare two or more models for the same data, we prefer the model with the lowest AIC. Similarly, when using BIC we prefer the model with the lowest BIC. The REML versions of the AIC and the BIC simply replace $\ell(\widehat{\boldsymbol{\theta}}|\boldsymbol{y})$ by $\ell_R(\widehat{\boldsymbol{\theta}}|\boldsymbol{y})$ and $\log(N)$ by $\log(N-p)$ in (2.28).

We will generally use likelihood-ratio tests to evaluate the significance of terms in the random-effects structure. That is, we fit different nested models in which the random-effects structure changes and apply likelihood-ratio tests. Stram and Lee (1994), using the results of Self and Liang (1987), argued that tests on the random effects structure conducted in this way can be conservative. That is, the p-value calculated from the $\chi^2_{k_2-k_1}$ distribution is greater than it should be. As Stram and Lee (1994) explain, changing from the more general model to the more specific model involves setting the variance of certain components of the random effects to zero, which is on the boundary of the parameter region. The asymptotic results for likelihood ratio tests have to be adjusted for boundary conditions. In the next section we use simulations to demonstrate the effect of these adjustments.

Simulating Likelihood Ratio Test Statistics

One way to check on the distribution of the likelihood ratio test statistic under the null hypothesis is through simulation. The `simulate.lme` function takes two model specifications, the null model and the alternative model. These may be given as lme objects corresponding to each model, or as lists of arguments used to produce such fits. In the latter case, only those characteristics that change between the two models need to be specified in the argument list for the alternative model.

For example, in the analysis of the `OrthoFem` data presented in §1.4.1, the `fm1OrthF` fit to the `OrthoFem` data has the specification

```
> fm1OrthF <- lme( distance ~ age, data = OrthoFem,
+     random = ~ 1 | Subject )
```

while `fm2OrthF` is fit as

```
> fm2OrthF <- update( fm1OrthF, random = ~ age | Subject )
```

Both models correspond to $\boldsymbol{X}_i$ matrices of

$$\boldsymbol{X}_1 = \boldsymbol{X}_2 = \cdots = \boldsymbol{X}_{11} = \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix}$$

with a two-dimensional $\boldsymbol{\beta}$ vector. In `fm1OrthF` the $\boldsymbol{Z}_i$ matrices are

$$\boldsymbol{Z}_1 = \boldsymbol{Z}_2 = \cdots = \boldsymbol{Z}_{11} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

and the $\boldsymbol{b}_i$ are one-dimensional random vectors with variance $\boldsymbol{\Psi} = \sigma_1^2$. In `fm2OrthF` the $\boldsymbol{Z}_i$ matrices are

$$\boldsymbol{Z}_1 = \boldsymbol{Z}_2 = \cdots = \boldsymbol{Z}_{11} = \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix}$$

and the $\boldsymbol{b}_i$ are two-dimensional random vectors with variance–covariance matrix

$$\boldsymbol{\Psi} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

In the terminology of hypothesis tests, `fm1OrthF` is the null model and `fm2OrthF` is the alternative model. In this case, the null model is a special case of the alternative model, with one fewer random effect. The model being fit as `fm1OrthF` is obtained from the model for `fm2OrthF` by requiring the last row and column of the 2×2 $\boldsymbol{\Psi}$ to be zero. Although there are three distinct entries in this row and column, these entries are determined by only two parameters because $\boldsymbol{\Psi}$ must be symmetric. Notice that one of these entries, σ_2^2, must be non-negative so setting it to zero corresponds to a boundary condition.

To simulate the likelihood ratio test statistic comparing model `fm1OrthF` to model `fm2OrthF` we generate data according to the null model using the parameter values from `fm1OrthF`. We then fit both the null and the alternative model to each set of simulated data and calculate the likelihood ratio test statistic. This is repeated for `nsim` cases. By doing this we obtain an empirical distribution of the likelihood ratio test statistic under the null hypothesis. We can then compare the empirical distribution to different χ^2 distributions as in Figure 2.3, which is produced by

```
> orthLRTsim <- simulate.lme( fm1OrthF, fm2OrthF, nsim = 1000 )
> plot( orthLRTsim, df = c(1, 2) )              # produces Figure 2.3
```

Figure 2.3 is a probability–probability plot—similar to a quantile–quantile plot but on the p-value scale, rather than on the scale of the likelihood ratio test (LRT) statistic. The nominal p-values for the simulated LRT statistics, under χ^2 distributions with 1 and 2 degrees of freedom and an equal-weight mixture of those χ^2 distributions (denoted `Mix(1,2)` in Figure 2.3), for both

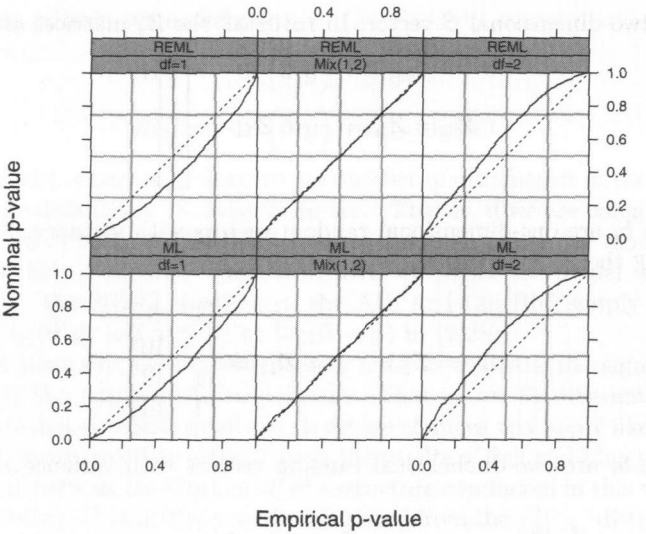

FIGURE 2.3. Plots of the nominal versus empirical p-values for the likelihood ratio test statistic comparing two models for the orthodontic data, female subjects only. The null model, `fm1OrthF`, has a random effect for the intercept only. The alternative model, `fm2OrthF`, has random effects for both the intercept and the slope. The null model was simulated 1000 times, both models were fit to the simulated data, and the likelihood ratio test statistic was calculated for both maximum likelihood and REML estimates. In each panel, the nominal p-values for the LRT statistics under the corresponding distribution are plotted versus the empirical p-values.

ML and REML estimation, are plotted versus the empirical p-values, obtained from the empirical distribution of the simulated LRT statistics.

For both REML and ML estimates, the nominal p-values for the LRT statistics under a χ^2 distribution with 2 degrees of freedom are much greater than the empirical p-values. This is the sense in which the likelihood ratio test using χ^2_2 for the reference distribution will be conservative—the actual p-value is smaller than the p-value that is reported. Stram and Lee (1994) suggest a $0.5\chi^2_1 + 0.5\chi^2_2$ mixture as a reference distribution, which is confirmed in Figure 2.3, for both ML and REML estimation. A χ^2_1 appears to be "anti-conservative" in the sense that the nominal p-values are smaller than the empirical p-values.

The adjustment suggested by Stram and Lee (1994) is not always this successful. According to this adjustment, the null distribution of the likelihood ratio test statistic for comparing `fm1Machine` to `fm2Machine` should have approximately a $0.5\chi^2_0 + 0.5\chi^2_1$ mixture distribution, where χ^2_0 represents a distribution with a point mass at 0. When simulated

```
> machineLRTsim <- simulate.lme(fm1Machine, fm2Machine, nsim= 1000)
```

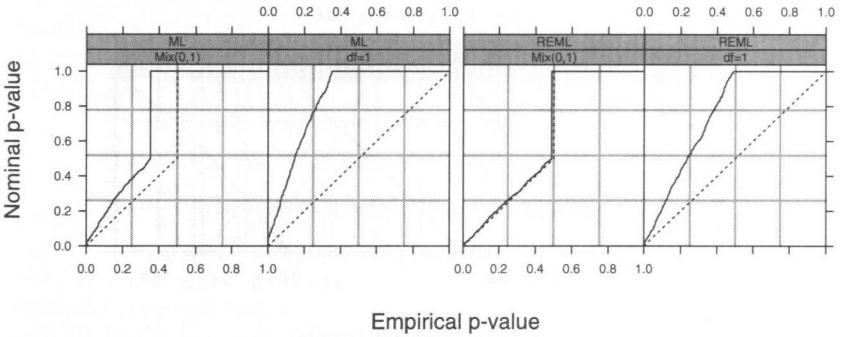

FIGURE 2.4. Plots of the nominal versus empirical p-values for the likelihood ratio test statistic comparing two models for the `Machines` data. The null model, `fm1Machine`, has a random effect for `Worker` only. The alternative model, `fm2Machine`, has random effects for the `Worker` and a random interaction for `Machine %in% Worker`. Both models were fit to 1000 sets of data simulated from the null model and the likelihood ratio test statistics were calculated.

```
> plot( machineLRTsim, df = c(0, 1),          # produces Figure 2.4
+   layout = c(4,1), between = list(x = c(0, 0.5)) )
```

it produces a distribution for the LRT statistics that closely agrees with the equal-weight mixture in the REML case, but which resembles a $0.65\chi^2_0 + 0.35\chi^2_1$ mixture in the ML case.

It is difficult to come up with general rules for approximating the distribution of the LRT statistic for such nested mixed-effects models. The naive approach of using a χ^2 distribution with the number of degrees of freedom determined by the difference in the number of nonredundant parameters in the models as the reference is easily implemented and tends to be conservative. This is the reference distribution we use to calculate the p-values quoted in the multiargument form of `anova`. One should be aware that these p-values may be conservative. That is, the reported p-value may be greater than the true p-value for the test and, in some cases, it may be much greater.

2.4.2 Hypothesis Tests for Fixed-Effects Terms

When two nested models differ in the specification of their fixed-effects terms, a likelihood ratio test can be defined for maximum likelihood fits only. As described in §2.2.5 a likelihood ratio test for REML fits is not feasible, because there is a term in the REML criterion that changes with the change in the fixed-effects specification.

Even though a likelihood ratio test for the ML fits of models with different fixed effects can be calculated, we do **not** recommend using such tests. Such

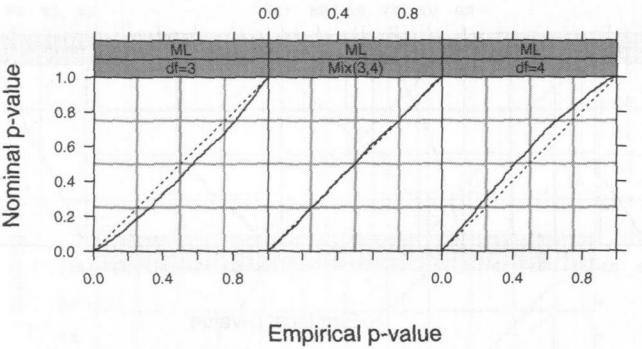

FIGURE 2.5. Plots of the nominal versus empirical p-values for the likelihood ratio test statistic comparing two models for the `ergoStool` data. The alternative model has a fixed effect for `Type` but the null model does not. The random effects specifications are the same. Both models were fit to 1000 sets of data simulated from the null model and the likelihood ratio test statistics from the maximum likelihood estimates were calculated.

likelihood ratio tests using the standard χ^2 reference distribution tend to be "anticonservative"—sometimes quite badly so.

As an example, consider the `ergoStool` example analyzed in §1.2.1. Suppose we compare `fm1Stool`, the model for the `ergoStool` data with a fixed effect for the `Type` factor, to a model without a fixed effect for the `Type` factor.

```
> stoolLRTsim <-
+   simulate.lme( m1 = list(fixed = effort ~ 1, data = ergoStool,
+                           random = ~ 1 | Subject),
+                 m2 = list(fixed = effort ~ Type),
+                 method = "ML", nsim = 1000 )
> plot( stoolLRTsim, df = c(3, 4) )                    # Figure 2.5
```

We can see from Figure 2.5 that, at 3 degrees of freedom, which is the difference in the number of parameters in the two models, the χ^2 distribution gives p-values that are "anticonservative." At 4 degrees of freedom the p-values will be conservative. The nominal p-values for the equal-weight mixture of χ^2_3 and χ^2_4 distributions, represented in the middle panel of Figure 2.5, are in close agreement with the empirical p-values.

In this case the slight anticonservative nature of the reported p-values may not be too alarming. However, as the number of parameters being removed from the fixed effects becomes large, compared to the total number of observations, this inaccuracy in the reported p-values can be substantial. For example, Littell, Milliken, Stroup and Wolfinger (1996, §1.5) provide analyses of data from a partially balanced incomplete block (PBIB) design. The design is similar to the randomized block design in the ergometric experiment described in §1.2 except that not every level of the treatment

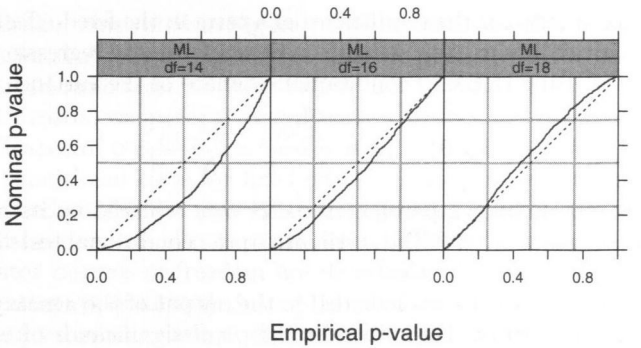

FIGURE 2.6. Plots of the nominal versus empirical p-values for the likelihood ratio test statistic comparing two models for the PBIB data. The alternative model has a fixed effect for Type, but the null model does not. The random effects specifications are the same. Both models were fit to 1000 sets of data simulated from the null model, and the likelihood ratio test statistics from the maximum likelihood estimates were calculated.

appears with every level of the blocking factor. This is the sense in which it is an "incomplete" block design. It is "partially balanced" because every pair of treatments occur together in a block the same number of times. These data are described in greater detail in Appendix A.22 and are given as an object called PBIB that is available with the nlme library.

The important point with regard to the likelihood ratio tests is that there are 15 levels of the Treatment factor and only 60 observations in total. The blocking factor also has 15 levels. If we simulate the likelihood ratio test and plot the p-values calculated from the χ^2_{14} distribution,

```
> pbibLRTsim <-
+     simulate.lme( m1 = list( fixed = response ~ 1, data = PBIB,
+                              random = ~ 1 | Block ),
+                   m2 = list( fixed = response ~ Treatment ),
+                   method = "ML", nsim = 1000 )
> plot( pbibLRTsim, df = c(14,16,18), weights = FALSE ) # Figure 2.6
```

we can see, from Figure 2.6, that the p-values calculated using χ^2_{14} as the reference distribution are seriously "anticonservative."

Another, perhaps more conventional, approach to performing hypothesis tests involving terms in the fixed-effects specification is to condition on the estimates of the random effects variance–covariance parameters, $\widehat{\boldsymbol{\theta}}$. As described in §2.2.1, for a fixed value of $\boldsymbol{\theta}$, the conditional estimates of the fixed effects, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, are determined as standard least-squares estimates. The approximate distribution of the maximum likelihood or the REML estimates of the fixed effects in (2.27) is exact for the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$.

Conditional tests for the significance of a term in the fixed-effects specification are given by the usual F-tests or t-tests for linear regression models, based on the usual (REML) conditional estimate of the variance

$$\widehat{\sigma}_R^2(\boldsymbol{\theta}) = s^2 = \frac{RSS}{N-p} = \frac{\|\boldsymbol{c}_{-1}\|^2}{N-p}.$$

In practice, the unknown parameter vector $\boldsymbol{\theta}$ is replaced by its maximum likelihood estimate or its REML estimate so the conditional tests hold only approximately.

The conditional t-tests are included in the output of the `summary` method applied to `lme` objects. They test the *marginal* significance of each fixed effect coefficient when all other fixed effects are present in the model. For example the t-tests for the `fm2Machine` model

```
> summary( fm2Machine )
. . .
Fixed effects: score ~ Machine
             Value Std.Error DF t-value p-value
(Intercept) 59.650    2.1447 36  27.813  <.0001
   Machine1  3.983    1.0885 10   3.660  0.0044
   Machine2  3.311    0.6284 10   5.269  0.0004
. . .
```

indicate that all the fixed-effects terms are significant.

The conditional F-tests are implemented in the single-argument form of the `anova` method for fitted models from `lme`. They test the significance of *terms* in the fixed effects model, which may include several coefficients. By default, the terms are tested *sequentially* in the order they enter the model, but the argument `type` to `anova` can be used to specify *marginal* F-tests. For example, to jointly test the significance of all 14 coefficients corresponding to `Treatment` in the `PBIB` example we use

```
> fm1PBIB <- lme( response ~ Treatment, data = PBIB, random = ~ 1 )
> anova( fm1PBIB )
            numDF denDF F-value p-value
(Intercept)     1    31  1654.2  <.0001
  Treatment    14    31     1.5  0.1576
```

We will compare this result, a p-value of 15.8%, with that from the likelihood ratio test. Because a likelihood ratio test for terms in the fixed-effects specification must be done on ML fits, we first re-fit `fm1PBIB` using maximum likelihood, then modify the model.

```
> fm2PBIB <- update( fm1PBIB, method = "ML" )
> fm3PBIB <- update( fm2PBIB, response ~ 1 )
> anova( fm2PBIB, fm3PBIB )
        Model df    AIC    BIC  logLik   Test L.Ratio p-value
fm2PBIB     1 17 56.571 92.174 -11.285
fm3PBIB     2  3 52.152 58.435 -23.076 1 vs 2  23.581  0.0514
```

The simulation illustrated in Figure 2.6 shows that the 15.8% p-value from the conditional F-test is much more realistic than the 5.1% p-value from the likelihood ratio test.

For this reason, we prefer the conditional F-tests and t-tests for assessing the significance of terms in the fixed effects.

These conditional tests for fixed-effects terms require denominator degrees of freedom. In the case of the conditional F-tests, the numerator degrees of freedom are also required, being defined by the term itself. The denominator degrees of freedom are determined by the grouping level at which the term is estimated. A term is called *inner* relative to a grouping factor if its value can change within a given level of the grouping factor. A term is *outer* to a grouping factor if its value does not change within levels of the grouping factor. A term is said to be estimated at level i, if it is inner to the $i-1$st grouping factor and outer to the ith grouping factor. For example, the term `Machine` in the `fm2Machine` model is outer to `Machine %in% Worker` and inner to `Worker`, so it is estimated at level 2 (`Machine %in% Worker`). If a term is inner to all Q grouping factors in a model, it is estimated at the level of the within-group errors, which we denote as the $Q+1$st level.

The intercept, which is the parameter corresponding to the column of 1's in the model matrices $\boldsymbol{X}_i$, is treated differently from all the other parameters, when it is present. As a parameter it is regarded as being estimated at level 0 because it is outer to all the grouping factors. However, its denominator degrees of freedom are calculated as if it were estimated at level $Q+1$. This is because the intercept is the one parameter that pools information from all the observations at a level even when the corresponding column in $\boldsymbol{X}_i$ doesn't change with the level.

Letting m_i denote the total number of groups in level i (with the convention that $m_0 = 1$ when the fixed effects model includes an intercept and 0 otherwise, and $m_{Q+1} = N$) and p_i denote the sum of the degrees of freedom corresponding to the terms estimated at level i, the ith level denominator degrees of freedom is defined as

$$denDF_i = m_i - (m_{i-1} + p_i), \quad i = 1, \ldots, Q+1.$$

This definition coincides with the classical decomposition of degrees of freedom in balanced, multilevel ANOVA designs and gives a reasonable approximation for more general mixed-effects models.

For example, in the `fm2Machine` model, $Q = 2$, $m_0 = 1$, $m_1 = 6$, $m_2 = 18$, $m_3 = 54$, $p_0 = 1$, $p_1 = 0$, $p_2 = 2$, and $p_3 = 0$, giving $denDF_1 = 5$, $denDF_2 = 10$, and $denDF_3 = 36$.

```
> anova( fm2Machine )
             numDF denDF F-value p-value
(Intercept)      1    36  773.57  <.0001
    Machine      2    10   20.58   3e-04
```

Because `Machine` is estimated at level 2, its denominator degrees of freedom is 10.

Another example is provided by the analysis of the `Oats` data, presented in §1.6, which shows an `anova` of the form

```
> anova( fm2Oats )
                numDF denDF F-value p-value
  (Intercept)       1    51  245.14  <.0001
ordered(nitro)      3    51   41.05  <.0001
      Variety       2    10    1.49  0.2724
```

In this example, $Q = 2$, $m_0 = 1$, $m_1 = 6$, $m_2 = 18$, $m_3 = 72$, $p_0 = 1$, $p_1 = 0$, $p_2 = 2$, and $p_3 = 3$, giving $denDF_1 = 5$, $denDF_2 = 10$, and $denDF_3 = 51$. The `nitro` factor changes within levels of the first-level grouping factor, `Block`, and within levels of the second-level grouping factor, `Variety %in% Block`. Thus, it is inner to each of these grouping factors and is estimated at level 3, with 51 denominator degrees of freedom. By definition, the `Variety` factor cannot change within levels of `Variety %in% Block`, but it changes within levels of `Block`. It is therefore outer to `Variety %in% Block` and inner to `Block`, being estimated at level 2 with 10 denominator degrees of freedom.

2.4.3 Confidence Intervals

Approximate confidence intervals on the variance–covariance components and the fixed effects are obtained using the approximate distributions for the maximum likelihood estimates and the REML estimates described in §2.3 and the conditional t-tests described in §2.4.2.

Letting df_j denote the denominator degrees-of-freedom for the conditional t-test corresponding to the the jth fixed effect based on the $\widehat{\boldsymbol{\beta}}$, an approximate confidence interval of level $1 - \alpha$ for β_j is

$$\widehat{\beta}_j \pm t_{df_j}(1 - \alpha/2)\widehat{\sigma}_R \sqrt{\left[\boldsymbol{R}_{00}^{-1}\boldsymbol{R}_{00}^{-T}\right]_{jj}},$$

where $t_{df_j}(1 - \alpha/2)$ denotes the $1 - \alpha/2$ quantile of the t-distribution with df_j degrees of freedom and $\widehat{\sigma}_R$ denotes the REML estimate of σ. The matrix $\boldsymbol{R}_{00}$ is evaluated at the estimated value of $\boldsymbol{\theta}$.

Confidence intervals on the within-group standard deviation σ are obtained from the approximate distribution in (2.27). Letting $\left[\boldsymbol{\mathcal{I}}^{-1}\right]_{\sigma\sigma}$ represent the last diagonal element of the inverse empirical information matrix defined in (2.27), an approximate confidence interval of level $1 - \alpha$ for σ is

$$\left[\widehat{\sigma}\exp\left(-z(1 - \alpha/2)\sqrt{\left[\boldsymbol{\mathcal{I}}^{-1}\right]_{\sigma\sigma}}\right), \widehat{\sigma}\exp\left(z(1 - \alpha/2)\sqrt{\left[\boldsymbol{\mathcal{I}}^{-1}\right]_{\sigma\sigma}}\right)\right],$$

where $z(1 - \alpha/2)$ denotes the $1 - \alpha/2$ quantile of the standard normal distribution. This confidence interval formulation works for both ML and REML estimates, with the obvious modifications.

Confidence intervals on the variance–covariance components for the random effects are a bit trickier to obtain. In practice, one is interested in getting confidence intervals on the original scale of the elements of $\boldsymbol{\Psi}$ and not in the scale of the unconstrained parameters $\boldsymbol{\theta}$ used in the optimization. For some simple forms of $\boldsymbol{\Psi}$, such as a diagonal structure or a multiple of the identity structure, it is easy to transform the confidence intervals obtained in the unconstrained scale (the logarithm of the standard deviations in the two examples mentioned) back to the original parameter scale (by exponentiating the confidence limits in the case of the diagonal and multiple of the identity structures).

In the case of a general positive-definite $\boldsymbol{\Psi}$, however, usually it is not possible to transform back to the original scale the confidence intervals obtained for the unconstrained parameter used in the optimization. This is true, for example, for the matrix logarithm parameterization described in §2.2.7, when the dimension of $\boldsymbol{\Psi}$ is greater than one.

The approach used in `lme` is to consider a different parameterization for general positive-definite $\boldsymbol{\Psi}$ when calculating confidence intervals. This parameterization, which we call the *natural* parameterization, uses the logarithm of the standard deviations and the generalized logits of the correlations. For a given correlation parameter $-1 < \rho < 1$, its generalized logit is $\log[(1+\rho)/(1-\rho)]$ which can take any value in the real line. We denote by $\boldsymbol{\eta}$ the parameter vector determining the natural parameterization. The elements of $\boldsymbol{\eta}$ are *individually* unconstrained, but not *jointly* so. Therefore, the natural parameterization cannot be used for optimization. However, the elements of $\boldsymbol{\eta}$ can be individually transformed into meaningful parameters in the original scale, so it is a useful parameterization for constructing confidence intervals.

If η_j corresponds to the logarithm of a standard deviation in $\boldsymbol{\Psi}$ and letting $[\boldsymbol{\mathcal{I}}^{-1}]_{jj}$ denote its associated diagonal element in the inverse empirical information matrix, an approximate level $1-\alpha$ confidence interval for the corresponding standard deviation is

$$\left[\exp\left(\widehat{\eta}_j - z(1-\alpha/2)\sqrt{[\boldsymbol{\mathcal{I}}^{-1}]_{jj}}\right), \exp\left(\widehat{\eta}_j + z(1-\alpha/2)\sqrt{[\boldsymbol{\mathcal{I}}^{-1}]_{jj}}\right)\right].$$

An approximate confidence interval for a correlation coefficient represented by η_j in the natural parameter vector is

$$\left[\frac{\exp\left(\widehat{\eta}_j - z(1-\frac{\alpha}{2})\sqrt{[\boldsymbol{\mathcal{I}}^{-1}]_{jj}}\right)-1}{\exp\left(\widehat{\eta}_j - z(1-\frac{\alpha}{2})\sqrt{[\boldsymbol{\mathcal{I}}^{-1}]_{jj}}\right)+1}, \frac{\exp\left(\widehat{\eta}_j + z(1-\frac{\alpha}{2})\sqrt{[\boldsymbol{\mathcal{I}}^{-1}]_{jj}}\right)-1}{\exp\left(\widehat{\eta}_j + z(1-\frac{\alpha}{2})\sqrt{[\boldsymbol{\mathcal{I}}^{-1}]_{jj}}\right)+1}\right].$$

2.5 Fitted Values and Predictions

Fitted values, which are the predicted values for the observed responses under the fitted model, are often of interest for model checking. Predicted values for new observations are one of the primary quantities of interest when making inferences from a fitted model.

In a mixed-effects model, fitted values and predictions may be obtained at different levels of nesting, or at the population level. Population level predictions estimate the marginal expected value of the response. For example, letting $\boldsymbol{x}_h$ represent a vector of fixed effects covariates, the marginal expected value of the corresponding response y_h is

$$\mathrm{E}\,[y_h] = \boldsymbol{x}_h^T \boldsymbol{\beta}. \tag{2.29}$$

Predicted values at the kth level of nesting estimate the conditional expectation of the response, given the random effects at levels $\leq k$. For example, letting $\boldsymbol{z}_h(i)$ denote a vector of covariates corresponding to random effects associated with the ith group at the first level of nesting, the *level-1* predictions estimate

$$\mathrm{E}\,[y_h(i)|\boldsymbol{b}_i] = \boldsymbol{x}_h^T \boldsymbol{\beta} + \boldsymbol{z}_h(i)^T \boldsymbol{b}_i. \tag{2.30}$$

Similarly, letting $\boldsymbol{z}_h(i,j)$ denote a covariate vector associated with the jth level-2 group within the ith level-1 group, the *level-2* predicted values estimate

$$\mathrm{E}\,[y_h(i)|\boldsymbol{b}_i, \boldsymbol{b}_{ij}] = \boldsymbol{x}_h^T \boldsymbol{\beta} + \boldsymbol{z}_h(i)^T \boldsymbol{b}_i + \boldsymbol{z}_h(i,j)^T \boldsymbol{b}_{ij}. \tag{2.31}$$

This extends naturally to an arbitrary level of nesting.

The *Best Linear Unbiased Predictors* (*BLUPs*) of the population expected values and the conditional expectations given the random effects are obtained by replacing, in the expressions defining the expectations, $\boldsymbol{\beta}$ with its conditional estimate $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and the random effects with their BLUPs. For example, the BLUPs corresponding to the expected values in (2.29), (2.30), and (2.31) are

$$\begin{aligned}
\widehat{y}_h &= \boldsymbol{x}_h^T \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) \\
\widehat{y}_h(i) &= \boldsymbol{x}_h^T \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \boldsymbol{z}_h(i)^T \widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}) \\
\widehat{y}_h(i,j) &= \boldsymbol{x}_h^T \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \boldsymbol{z}_h(i)^T \widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}) + \boldsymbol{z}_h(i,j)^T \widehat{\boldsymbol{b}}_{ij}(\boldsymbol{\theta}).
\end{aligned}$$

In practice, the unknown parameter vector $\boldsymbol{\theta}$ is replaced by its maximum likelihood estimate or its REML estimate, producing estimated BLUPs of the expected values.

2.6 Chapter Summary

This chapter presents the theory and computational methods for linear mixed-effects models. We express linear mixed-effects models in the Laird–

Ware formulation. For a single grouping factor that divides the observations into M groups of $n_i, i = 1, \ldots, M$ observations, the model is written

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, M, \\ \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}). \end{aligned}$$

The parameters in the model are the p-dimensional fixed effects, $\boldsymbol{\beta}$, the $q \times q$ variance–covariance matrix, $\boldsymbol{\Psi}$, for the random effects, and the variance σ^2 of the "noise" $\boldsymbol{\epsilon}_i$. We estimate these parameters by maximum likelihood (ML) or by restricted maximum likelihood (REML).

Although the random effects $\boldsymbol{b}_i, i = 1, \ldots, M$ are not formally parameters in the model, we will often want to formulate our "best guess" for these values given the data. We use the Best Linear Unbiased Predictors (BLUPs) for this.

For computational purposes the variance–covariance matrix $\boldsymbol{\Psi}$ is re-expressed in terms of the relative precision factor $\boldsymbol{\Delta}$ which satisfies

$$\boldsymbol{\Delta}^T\boldsymbol{\Delta} = \boldsymbol{\Psi}/\sigma^2$$

and the matrix $\boldsymbol{\Delta}$ is expressed as a function of an unconstrained parameter vector $\boldsymbol{\theta}$.

The *profiled* log-likelihood function with respect to $\boldsymbol{\theta}$ can be easily calculated using matrix decompositions. That is, the log-likelihood corresponding to the conditionally best estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}^2(\boldsymbol{\theta})$ can be evaluated as a function of $\boldsymbol{\theta}$ alone. This simplifies the problem of optimizing the likelihood to get maximum likelihood estimates because it reduces the dimension of the optimization. The same simplification applies to REML estimation.

We describe approximate distributions for the maximum likelihood estimates and the REML estimates using results from asymptotic theory for linear mixed-effects models.

We compare models that differ in the random effects specification by likelihood ratio tests or by simulation-based parametric bootstrap evaluations.

We assess the significance of terms in the fixed-effects specification by standard linear regression tests conditional on the value of $\widehat{\boldsymbol{\theta}}$. These tests include t-tests for individual coefficients or F-tests for more complicated terms or linear combinations of coefficients. The degrees of freedom for a t-test (or the denominator degrees of freedom for an F-test) depend on whether the factor being considered is inner to the grouping factor (changes within levels of the grouping factor) or outer to the grouping factor (is invariant within levels of the grouping factor).

Approximate confidence intervals for the fixed effects and the variance–covariance parameters are produced from the approximate distributions of the maximum likelihood estimates and REML estimates.

All these results extend to multiple nested levels of random effects. A model with two levels of nested random effects, for example, is written

$$\boldsymbol{y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{i,j}\boldsymbol{b}_i + \boldsymbol{Z}_{ij}\boldsymbol{b}_{ij} + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, M_i,$$
$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_1), \quad \boldsymbol{b}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_2), \quad \boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

The two variance–covariance matrices $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are written in terms of relative precision factors $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$, parameterized by unconstrained parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The profiled log-likelihood or the profiled log-restricted-likelihood, a function of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ only, is maximized to produce the estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$, $\widehat{\boldsymbol{\Psi}_1}$, and $\widehat{\boldsymbol{\Psi}_2}$.

Exercises

1. The simulation results presented in Figure 2.4 (p. 87) indicate that the null distribution of the REML likelihood ratio test statistic comparing a null model with a single level of scalar random effects to an alternative model with nested levels of scalar random effects is approximately an equally weighted mixture of a χ^2_0 and a χ^2_1.

   Confirm this result by simulating a LRT statistic on the `Oats` data, considered in §1.6. The preferred model for those data, `fm4Oats`, was defined with `random = ~1 | Block/Variety`. Re-fit this model with `random = ~1 | Block`. Using this fit as the null model and `fm4Oats` as the alternative model, obtain a set of simulated LRT statistics with `simulate.lme`. Plot these simulated LRT statistics setting `df = c(0,1)` to obtain a plot like Figure 2.4. Are the conclusions from this simulation similar to those from the simulation shown in Figure 2.4?

   Note that `simulate.lme` must fit both models to `nsim` simulated sets of data. By default `nsim = 1000`, which could tie up your computer for a long time. You may wish to set a lower value of `nsim` if the default number of simulations will take too long.