UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

SCHOOL OF MATHEMATICS

STATISTICS PROJECT

REVIEW OF METHODS OF ESTIMATING PARAMETERS IN

NONLINEAR MIXED-EFFECTS (NLME) MODELS
A PROJECT SUBMITTED TO THE SCHOOL OF MATHEMATICS IN PARTIAL FULFILLMENT
FOR A DEGREE OF MASTER OF SCIENCE IN STATISTICS.

JULY 2007
by
Stephen Muteti Mbunzi
Supervised by
Dr. Thomas Achia
(University of Nairobi)

and
Sonal Nagda
(International Livestock Research Institute - Nairobi, Kenya)

Declaration

I the undersigned declare that this project is my original work
and to the best of my knowledge has not been presented for the
a ward of a degree in any other University.

Stephen Muteti Mbunzi
Reg. No I/56/7822/2005

Signature.................................... Date.....................................

Declaration By Supervisor

This project has been submitted for examination with my approval as supervisor

Dr. Thomas Achia
School of Mathematics,
University of Nairobi,
P.O. Box 30197 Nairobi,
KENYA.

Signature.................................... Date.....................................

Acknowledgments.

Gratitude is expressed to all my lecturers in the School of Mathematics , University of Nairobi, especially Prof. John Okoth Owino, Prof. M.M. Manene, Dr. F.N. Njui, Dr. Gachii, Prof. J.A.M. Otieno and Dr. R.W. Nguti, who contributed much in my search for knowledge.

My deepest gratitude goes to Dr. Achia, for allowing me do this work under his supervision, his endless advice, guidance and patience throughout my research work and even reading through the work.

Acknowledgment is also made of important assistance from the members of Research Method Groups of International Livestock Research Institute (ILRI) especially Jane Poole and Sonal Nagda, who helped me during my data analysis. I sincerely thank my classmates for helping me out at one time or another.

Dedicated:

To my wife who really offered me a lot of support and encouragement all throughout my study.

To my children, Faith and Francis for their continued love even when I was too busy during my studies.

Contents

Abstract.

This study is a critical review of theoretical issues that underline the linear mixed effects (LME) and nonlinear mixed effects (NLME) models. These two areas are revisited under maximum likelihood and restricted maximum likelihood estimation frameworks. We also review methods of estimating parameters in both linear and nonlinear mixed effects models. In the case of LME, we consider different ways of developing the likelihood estimators, key among these methods are the "pseudo-data" approach, orthogonal triangular decomposition method and the use of penalized least squares problem.

For NLME, we intended to investigate the computational efficiency and accuracy of computational methods, like the b-splines, that could be used to approximate the log-likelihood function in non-linear mixed effects models. This was not achieved in this study but can be an interesting area for further research work. We critically review the four methods of estimating parameters by Pinheiro and Bates (1995) through proving a number of lemmas. Our proves led us to same stated results by different researchers in different papers. This is a key issue in the investigation of other expansion methods and comparing their computational efficiency and accuracy with these existing ones.

We conclude by giving an insight into linear mixed effects models by analyzing a data set from livestock where we examine incorporation of random effects to study variations among rams (sires) and ewes (dams) and their influences on lamb weaning weight. Factors like year of birth of the lamb, sex of lamb, age at weaning, age of dam, ewe breed and ram breed are found to influence the weaning weight differently. With the random terms (ewes and rams) specified in the model the estimate of the residual among lamb variance is found to reduce due to taking into account the variations among rams and ewes within breeds. It was our intention to obtain heritability estimates which determine the proportion of the variation among offspring that have been handed down from parents out of these random estimates.

Keywords: repeated-measures data, multilevel data, longitudinal data, LME, NLME, "pseudo-data" and b-splines.

Chapter 1

Introduction

1.1 Introduction

A mixed-effects model may simply be defined as a model with both fixed effects and random effects. Mixed-effects models are primarily used to describe relationship between a response variable and some covariates in data that are grouped according to one or more classification factors. Examples of such grouped data include longitudinal data, repeated-measures data, multilevel data and block designs.

The increasing popularity of mixed-effects models is explained by the flexibility they offer in modeling the within-group correlation often present in grouped data, by the handling of balanced and unbalanced data in a unified framework, and by the availability of reliable and efficient softwares (for example R) for fitting them.

These mixed-effects models may be divided into linear mixed-effects models and nonlinear mixed-effects models.

Linear mixed-effects models (MLE) 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.

Nonlinear mixed-effects models (NLME) are mixed-effects models in which some, or all, of the fixed and random effects occur nonlinearly in the model function.

Several approximation methods to the log-likelihood in the nonlinear mixed-effects model have been proposed and described by different researchers. Such methods include linear mixed-effects approximation method, a modified Laplacian approximation, importance sampling and Gaussian quadrature.

Linear mixed-effects (LME) approximation is an estimation algorithm which alternates between two steps, a penalized nonlinear least squares(PNLS) step, and a linear mixed-effects (LME) step.

Laplacian approximations are used frequently in Bayesian inference to estimate marginal posterior densities and predictive distributions can also be used for approximating the likelihood function in NLME models.

Importance sampling has been applied to the problem of belief inference in Bayesian networks (BNs) and action selection in influence diagrams (IDs). In its simpler form, the importance-

sampling distribution used is the "prior" distribution of the BN resulting from setting the value of the evidence. It provides estimates with larger variance than necessary hence far from optimal.
Adaptive Gaussian quadrature rules are used to approximate integrals of functions with respect to a given kernel by a weighted average of the integrand evaluated at predetermined abscissas.

1.2 Literature Review

Nonlinear mixed-effects models offer very high flexibility in handling the unbalanced repeated-measures data arising in different areas of investigation such as pharmacokinetics and economics. According to Pinheiro and Bates(1995), such repeated-measures data are generated by observing a number of subjects repeatedly under varying conditions. In longitudinal studies, observations on the same subject are made at different times.
Mixed-effects models assume that the form of the intra subject model that relates the response variable to time is common to all subjects, but some of the parameters that define the model may vary with subject. Nonlinear mixed-effects models are mixed-effects models in which the intra subject model relating the response variable to time is nonlinear in the parameters.
Pinheiro and Bates (1995) considered a nonlinear mixed-effects model which could be viewed as a hierarchical model that in some ways generalizes both the linear-mixed effects of Laird and Ware (1982) and the usual nonlinear model for independent data of Bates and Watts (1988).
Laird and Ware (1982) defined a family of models for serial measurements that included both growth models and repeated-measures models as special cases.
For measured, multivariate normal data, they proposed the following model in two stages:

- **Stage 1**
  For each individual unit, i,

$$\mathbf{y}_i = x_i\alpha_i + z_i\mathbf{b}_i + \mathbf{e}_i \tag{1.1}$$

  where $\mathbf{e}_i$ is distributed as $N(0, R_i)$ and assumed independent, R_i is an $n_i \times n_i$ positive definite covariance matrix and depends on i through its dimension n_i, but the set of unknown parameters in R_i will not depend upon i. α and $\mathbf{b}_i$ are considered fixed at this stage.

  Again α denote a $p \times 1$ vector of unknown population parameters, x_i a known $n_i \times p$ design matrix linking α to $\mathbf{y}_i$. b_i denote a $k \times 1$ vector of unknown individual effects, z_i a known $n_i \times k$ design matrix linking $\mathbf{b}_i$ to $\mathbf{y}_i$.

- **Stage 2**
  The $\mathbf{b}_i$ are distributed as $\mathrm{N}(0, D)$, independent of each other and of the $\mathbf{e}_i$.

  Here D is a $k \times k$ positive-definite covariance matrix. The population parameters, α, are treated as fixed effects.

Out of their formulation, marginally, the y_i are independent normals with mean $x_i\alpha$ and covariance matrix

$$R_i + Z_i D Z_i^T.$$

When $R_i = \sigma^2 I$, where I denotes an identity matrix, they got a simplified model called the "conditional-independence model". This implied that the n_i responses on individual i were independent, conditional on $\mathbf{b}_i$ and α.

Pinheiro and Bates (1995) considered a slightly modified version of the model proposed in Lindstrom and Bates (1990) which define the j^{th} observation on the i^{th} individual as

$$\mathbf{y}_{ij} = f(\phi_i, \mathbf{x}_{ij}) + \mathbf{e}_{ij}, i = 1, \cdots, M; j = 1, \cdots, n_i \tag{1.2}$$

where $\mathbf{y}_{ij}$ is the j^{th} response on the i^{th} individual, $\mathbf{x}_{ij}$ is the predictor vector for the j^{th} response on the i^{th} individual, f is a nonlinear function of the predictor vector and a parameter vector ϕ_i of length r and $\mathbf{e}_{ij}$ is a normally distributed noise term.
According to Lindstrom and Bates (1990), the predictor variables x_{ij} are not restricted.
The parameter vector could vary from individual to individual, thus the subject-specific parameter vector was modeled as

$$\phi_{ij} = A_{ij}\beta + B_{ij}\mathbf{b}_i, \ \mathbf{b}_i \sim N(0, \sigma^2 D), \tag{1.3}$$

where β is a p-dimensional vector of fixed population parameters, $\mathbf{b}_i$ is a q-dimensional random effects vector associated with individual i. A_{ij} and B_{ij} are design matrices of size $r \times p$ and $r \times q$ for the fixed and random effects respectively. $\sigma^2 D$ is a (general) variance-covariance matrix.
They also assumed that observations made on different subjects are independent and that e_{ij} are iid $\mathrm{N}(0, \sigma^2)$ and independent of b_i.
Though different methods can be used to estimate the parameters in model (1.2), Davidian et al (1991), Ramos et al (in press)), Pinheiro and Bates (1995) restricted themselves to maximum likelihood (ML) and restricted maximum likelihood (REML) estimation methods.
Pinheiro and Bates (1995) based their maximum likelihood estimate in (1.2) on the marginal density of y

$$f(\mathbf{y}|\beta, D, \sigma^2) = \int f(\mathbf{y}|\mathbf{b}, \beta, D, \sigma^2) \ f(\mathbf{b}) \ d\mathbf{b} \tag{1.4}$$

where $f(\mathbf{y}|\beta, D, \sigma^2)$ is the marginal density of $\mathbf{y}$, $f(\mathbf{y}|\mathbf{b}, \beta, D, \sigma^2)$ is the conditional density of $\mathbf{y}$ given the random effects $\mathbf{b}$ and $f(\mathbf{b})$ is the marginal distribution of $\mathbf{b}$.
Generally, the integral above does not have a closed-form expression when the model function f is nonlinear in b_i, so different approximations have been proposed for estimating it.
Some of these methods consist of:

- taking a first-order Taylor expansion of the model function f around the expected value of the random effects. Such an approach was used by Sheiner and Beal (1980) and Vonesh and Carter (1992).

- taking a first-order Taylor expansion of the model function f around the conditional (or D) modes of the random effects, as given in Lindstrom and Bates (1990).

- the use of Gaussian quadrature rules, as applied in Davidian and Gallant (1992).

Pinheiro and Bates (1995) considered four different approximations to the log-likelihood in the nonlinear mixed-effects model (1.2), namely:

(i) Lindstrom and Bate's (1990) linear mixed-effects (LME) method;

(ii) a modified Laplacian approximation (Tierney and Kadane (1986));

(iii) important sampling (Geweke (1989)) and

(iv) Gaussian quadrature (Davidian and Gallant (1992)).

They compared them based on their computational and statistical properties, using both real data examples and simulation results.
In the description of the four different approximations to the log-likelihood in the nonlinear mixed-effects model (1.2),they showed that there exists a close relationship between the Laplacian approximation, importance sampling and a Gaussian quadrature rule centered around the conditional modes of the random effects **b**.
Their first approximation considered (LME approximation) was from an algorithm proposed by Lindstrom and Bates (1990) for estimating the parameters in model (1.2). Their algorithm proceeded in two alternating steps, a penalized nonlinear least squares (PNLS) step and a linear mixed-effects (LME) step until some convergence criterion is met. Such alternating algorithms tend to be more efficient when the estimates of the variance-covariance components (D and σ^2) are not highly correlated with the estimates of the fixed effects β.
In the linear mixed-effects model, the maximum likelihood estimates of D and σ^2 are asymptotically independent of the maximum likelihood estimates of β as was demonstrated by Pinheiro (1994). Same results have not yet been extended to the nonlinear mixed-effects model (1.2).
To carry out Laplacian approximation, the integral they wanted to estimate for the marginal distribution of y_i in model (1.2) was

$$f(\mathbf{y}_i|\beta, D, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{(n_i+q)/2} \mid D \mid^{-1/2} \exp\left[-g(\beta, D, y_i, b_i)/2\sigma^2\right] db_i \qquad (1.5)$$

where

$$g(\beta, D, y_i, b_i) = \|y_i - f_i(\beta, b_i)\|^2 + b_i^T D^{-1} b_i \qquad (1.6)$$

This integrand was expanded only around $\hat{b}$.
Wolfinger (1993) had expanded the same integrand around both $\hat{\beta}$ and $\hat{b}$, by assuming a flat prior for β.
Pinheiro and Bates (1995), noted that, there did not seem to be a straightforward generalization of the concept of REML to nonlinear mixed-effects models. The difficulty is that REML depends heavily upon the linearity of the fixed effects in the model function, which does not occur in nonlinear models.
Lindstrom and Bates (1990) circumvented that problem by using an approximation to the model function f in which the fixed effects β occur linearly.

This could not be done for the Laplacian approximation, unless they considered yet another Taylor expansion of the model function that would lead then back to something very similar to Lindstrom and Bate's approach.
Though they got exact results with importance sampling approximation, they could not in general obtain a closed-form expression for the maximum likelihood estimate 0f σ^2 for fixed β and D. Thus profiling on σ^2 was no longer reasonable. For such exact results, the model function was linear in b.
The adaptive Gaussian quadrature approximation very closely resembled that obtained for importance sampling. It gave the exact log-likelihood when the model function was linear in b, but not true in general for the Gaussian quadrature approximation.
Like the importance sampling approximation, the Gaussian quadrature approximation could not be profiled on σ^2 to reduce the dimensionality of the optimization problem.
It is worth noting that according to Pinheiro and Bates (1995) LME approximation to the log-likelihood function in NLME models gave accurate and reliable estimation results. Its main advantage are its computational efficiency and the availability of a REML version of it.
With regard to REML estimation, Pinheiro and Bates (1995) found that the results from their data suggested that the bias correction ability of the method depended on the nonlinear model that was being considered.

1.2.1 Statement of the problem

Several numerical approaches to approximating the log-likelihood function in non-linear mixed effects models have been proposed in the literature. Amongst them are methods like the LME approximation method, that has great computational efficient and availability of a REML version yet having bias correction depended on the non-linear method being considered, the Gaussian quadrature approximation, that gave the exact log-likelihood when the model function was linear in b, but not true in general for the Gaussian quadrature approximation, and the importance sampling method, that gave exact results but could not in general give a closed-form expression for the maximum likelihood estimate 0f σ^2 for fixed β and D.
There is need to investigate the computational efficiency and accuracy of these and other computational methods, like the b-splines, that could be used to approximate the log-likelihood function in non-linear mixed effects models. There is need also to investigate the possibility of developing REML versions for approximation methods. Very importantly, one needs to understand these already existing computational and numeric methods to competently investigate other computational methods.
There are numerous case studies that have been developed using other softwares, like SAS and GENSTAT, to illustrate the use of these softwares in analysis linear and non-linear mixed effects models. There exists to the best of our knowledge no such works with the R software, a free-ware down loadable on the Internet.

1.3 Objectives

1.3.1 Broad objectives

In our research,we would like to

1. Review computational and numeric methods of approximating the likelihood in nonlinear mixed effects models.

2. Give a case study using R on a data set from livestock with linear mixed-effects.

1.3.2 Specific objectives

1. Review the theory on linear mixed-effects models.

2. Review the theory on nonlinear mixed-effects models.

3. Review the Laplacian approximations, importance sampling and Gaussian quadratures in approximating the log-likelihood and investigate possibilities of using other expansion methods to approximate the log-likelihood (e.g b-splines and cubic-splines) and possibilities of defining a REML estimation for these approximations/estimation methods.

1.4 Methodology

We have already found that taking a first-order Taylor expansion of the model function f around the expected value of the random effects, and around the conditional modes of the random effects and use of Gaussian quadrature rules were the three methods used in the approximation to the log-likelihood in the nonlinear mixed-effects (NLME) model (1.2). The different approximations to the log-likelihood in the NLME models are linear mixed-effects (LME) method, a modified Laplacian approximation, importance sampling and Gaussian quadrature.
In our research we would attempt various expansion methods like Taylor expansion, b-splines and cubic-splines to approximate the function f. As much as possible we would carry out these expansion methods on linear mixed-effects (LME) approximation. Out of our data sets we would approximate both the first-order compartment model and the logistic model using Taylor expansion, b-splines and cubic-splines methods. We would also simulate results for the variance-covariance components and for the fixed effects in both logistic model and the first-order compartment model.

1.5 Expected output

At the end of our work we expect to review the theory on linear mixed-effects models and nonlinear mixed-effects models. It is our expectation that we would be able to review linear mixed effects approximation method, Laplacian approximations, importance sampling and

Gaussian quadratures in approximating the log-likelihood. Before the end of our study we would wish to compare Taylor expansion, b-splines and cubic-splines as expansion methods in linear mixed-effect (LME) approximation to the log-likelihood in nonlinear mixed-effects model . We would give a well-detailed example in linear mixed effects models using R on a data-set from livestock to illustrate these expansion methods.

Chapter 2

Critical Review of literature

Several different nonlinear mixed-effects (NLME) models have been proposed in recent years (Sheiner and Beal (1980); Mallet, Mentre, Steimer and Lokiek (1988); Lindstrom and Bates (1990); Davidian and Gallant (1992); Vonesh and Carter (1992); Wakefield, Smith, Racine-Poon and Gelfand in press; and Pinheiro and Bates (1995))
Our aim is to review nonlinear mixed-effects model by Pinheiro and Bates (1995) which is slightly a modified version of the model proposed in Lindstrom and Bates (1990). This model can be viewed as a hierarchical model that in some ways generalizes both the linear mixed-effects model of Laird and Ware (1982) and the usual nonlinear model for independent data (Bates and Watts (1988)).
Different methods can be used to estimate the parameters in NLME model. We will restrict ourselves to considering maximum likelihood and restricted maximum likelihood estimation In this chapter, we do a critical review on the theory on LME and NLME models.

2.1 Review of Linear Mixed Effects (LME) Models

2.1.1 Likelihood Estimation for LME Models

Consider the model

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i b_i + \epsilon_i,\; i = 1, \ldots, M \tag{2.1}$$

where β is a p-dimensional vector of fixed effects, $b_i \sim N(0, \psi)$ and is a q-dimensional vector of random effects, $\epsilon_i \sim N(0, \sigma^2 I)$ and is n_i-dimensional within-group error vector with a spherical Gaussian distribution, $\mathbf{X}_i$ is a known fixed effects regressor $(n_i \times p)$ matrix and $\mathbf{Z}_i$ is a known random effects regressor $(n_i \times q)$ matrix.
(2.1) is a Mixed Effect Model with a single level of random effects. The parameters of the model are β, σ^2, and Δ, the relative precision factor.
We use θ to represent an unconstrained set of parameters that determine Δ. The likelihood function of 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,

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = f(\mathbf{y}|\beta, \theta, \sigma^2), \tag{2.2}$$

where L is the likelihood, f is a probability density, and y is the entire N-dimensional response vector,$N = \sum_{i=1}^{M} n_i$.
In the subsequent sections we will make the following assumptions:

1. $b_i \sim N(0, \psi)$ and that ψ is a symmetric and positive definite matrix.

2. the matrix ψ can be expressed in the form of a relative precision factor, Δ, which satisfies

$$\left(\frac{\psi^{-1}}{1/\sigma^2}\right) = \Delta^T \Delta \tag{2.3}$$

3. Δ factors precision matrix, ψ^{-1}, of the random effects relative to the precision $1/\sigma^2$, of the ϵ_i.

It follows therefore that

$$\left(\frac{\psi^{-1}}{1/\sigma^2}\right) = \Delta^T \Delta \Rightarrow \psi^{-1} = \left(\frac{\Delta^T \Delta}{\sigma^2}\right) \tag{2.4}$$

Lemma 1. *The likelihood function of the sample data is given as*

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} . \exp\{-\Sigma_{i=1}^{M} \left\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i\hat{b}_i\right\|^2 /2\sigma^2\} \prod_{i=1}^{M} \frac{abs|\Delta|}{\sqrt{|\bar{Z}_i^T \bar{Z}_i|}}$$

Proof. Because the non-observable random effects $b_i, i = 1, \ldots, M$ 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 b_i and the ϵ_i to express this as

$$f(\mathbf{y_i}|b_i\beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n_i}{2}} \exp\left\{-\frac{\|(y_i - X_i\beta - Z_i b_i)\|^2}{2\sigma^2}\right\} \tag{2.5}$$

$$\begin{aligned} L(\beta, \theta, \sigma^2|\mathbf{y}) &= f(\mathbf{y}|\beta, \theta, \sigma^2) \\ &= f((y_1, y_2, \ldots, y_M)'|\beta, \theta, \sigma^2) \end{aligned} \tag{2.6}$$

and by independence of the y_i's we have

$$\begin{aligned} L(\beta, \theta, \sigma^2|\mathbf{y}) = &\ f(y_1|\beta, \theta, \sigma^2) f(y_2|\beta, \theta, \sigma^2) \ldots f(y_M|\beta, \theta, \sigma^2) \\ &= \prod_{i=1}^{M} f(y_i|\beta, \theta, \sigma^2) \end{aligned} \tag{2.7}$$

Thus,

$$L(\beta,\theta,\sigma^2|\mathbf{y}) = \prod_{i=1}^{M} f(\mathbf{y_i}|\beta,\theta,\sigma^2), \tag{2.8}$$

By applying the Bayesian Theorem, we have

$$f(\mathbf{y_i}|\beta,\theta,\sigma^2) = \frac{f(y_i\beta,\theta,\sigma^2)}{f(\beta,\theta,\sigma^2)} \tag{2.9}$$

we get

$$\begin{aligned}\prod_{i=1}^{M} f(y_i|\beta,\theta,\sigma^2) &= \prod_{i=1}^{M} \frac{f(y_i,\beta,\theta,\sigma^2)}{f(\beta,\theta,\sigma^2)},\\ &= \prod_{i=1}^{M} \int \frac{f(\mathbf{y}_i,b_i,\beta,\theta,\sigma^2)}{f(\beta,\theta,\sigma^2)} dbi,\\ &= \prod_{i=1}^{M} \int \frac{f(\mathbf{y_i}|b_i,\beta,\theta,\sigma^2).f(b_i\beta,\theta,\sigma^2)}{f(\beta,\theta,\sigma^2)} dbi.\end{aligned} \tag{2.10}$$

The likelihood function then becomes

$$\prod_{i=1}^{M} f(\mathbf{y_i}|\beta,\theta,\sigma^2) = \prod_{i=1}^{M} \int \frac{f(\mathbf{y_i}|b_i\beta,\theta,\sigma^2).f(b_i\beta,\theta,\sigma^2)}{f(\beta,\theta,\sigma^2)} dbi, \tag{2.11}$$

which leads to

$$L(\beta,\theta,\sigma^2|\mathbf{y}) = \prod_{i=1}^{M} \int f(\mathbf{y_i}|b_i\beta,\theta,\sigma^2).f(b_i|\beta,\theta,\sigma^2) dbi, \tag{2.12}$$

giving

$$L(\beta,\theta,\sigma^2|\mathbf{y}) = \prod_{i=1}^{M} \int f(\mathbf{y_i}|b_i\beta,\sigma^2).f(b_i|\theta,\sigma^2) dbi, \tag{2.13}$$

since $E(\mathbf{y_i}|b_i\beta,\sigma^2) = X_i\beta + Z_ib_i$ and $var(\mathbf{y_i}|b_i\beta,\sigma^2) = \sigma^2\mathbf{I}$.

Then

$$f(\mathbf{y_i}|b_i\beta,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n_i/2} \exp -\frac{1}{2}\left[(y_i - X_i\beta - Z_ib_i)^T(\sigma^2\mathbf{I})^{-1}(y_i - X_i\beta - Z_ib_i)\right], \tag{2.14}$$

which simplifies to

$$\begin{aligned}f(\mathbf{y_i}|b_i\beta,\sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n_i}{2}} \exp -\frac{1}{2\sigma^2}(y_i - X_i\beta - Z_ib_i)^T(y_i - X_i\beta - Z_ib_i),\\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n_i}{2}} \exp -\frac{\left[(y_i - X_i\beta - Z_ib_i)^T(y_i - X_i\beta - Z_ib_i)\right]}{2\sigma^2}.\end{aligned} \tag{2.15}$$

Again, since $b_i \sim N(0, \psi)$ its distribution becomes

$$\begin{aligned} f(b_i|\theta, \sigma^2) &= \frac{1}{(2\pi)^{q/2}|\psi|^{1/2}} \exp\left\{-\frac{(b_i - \mathbf{0})^T(\psi)^{-1}(b_i - \mathbf{0})}{2}\right\}, \\ &= \frac{1}{(2\pi)^{q/2}|\psi|^{1/2}} \exp\left\{-\frac{(b_i^T(\psi)^{-1}b_i)}{2}\right\} \end{aligned} \tag{2.16}$$

(2.4) implies that

$$\begin{aligned} b_i^T\psi^{-1}b_i &= \left(\frac{b_i^T\Delta^T\Delta b_i}{\sigma^2}\right) \\ &= \left(\frac{(\Delta b_i)^T(\Delta b_i)}{\sigma^2}\right) \\ &= \left(\frac{\|\Delta b_i\|^2}{\sigma^2}\right) \end{aligned} \tag{2.17}$$

Hence the equation now becomes,

$$f(b_i|\theta, \sigma^2) = \frac{1}{(2\pi)^{q/2}|\psi|^{1/2}} \exp\left(-\frac{\|\Delta b_i\|^2}{2\sigma^2}\right) \tag{2.18}$$

Again we have

$$\begin{aligned} |\psi| &= |\left(\frac{\Delta^T\Delta}{\sigma^2}\right)^{-1}| \\ &= \sigma^{2q}|(\Delta^T\Delta)^{-1}| \\ &\Rightarrow |\psi|^{1/2} = \sigma^q abs|\Delta|^{-1} \end{aligned}$$

Thus our equation becomes

$$f(b_i|\theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{q/2} abs|\Delta|^{-1}} \exp\left(-\frac{\|\Delta b_i\|^2}{2\sigma^2}\right) \tag{2.19}$$

Substituting (2.5) and (2.19) into (2.13) we get the likelihood as

$$\begin{aligned} L(\beta, \theta, \sigma^2|\mathbf{y}) &= \prod_{i=1}^{M} \int f(\mathbf{y_i}|b_i\beta, \sigma^2).f(b_i|\theta, \sigma^2)dbi \\ &= \prod_{i=1}^{M} \int \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n_i}{2}} \exp\left\{-\frac{\|(y_i - X_i\beta - Z_ib_i)\|^2}{2\sigma^2}\right\}.\frac{1}{(2\pi\sigma^2)^{q/2}abs|\Delta|^{-1}} \exp\left(-\frac{\|\Delta b_i\|^2}{2\sigma^2}\right) dbi \end{aligned}$$

Which can now be written as

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \int \left\{\frac{\exp -[\|(y_i - X_i\beta - Z_ib_i)\|^2 + \|\Delta b_i\|^2]/2\sigma^2}{(2\pi\sigma^2)^{q/2}}\right\} dbi \tag{2.20}$$

Part of the exponent in the above equation can be written as

$$\begin{aligned}\|(y_i - X_i\beta - Z_i b_i)\|^2 + \|\Delta b_i\|^2 &= (y_i - X_i\beta - Z_i b_i)^T(y_i - X_i\beta - Z_i b_i) + (-\Delta b_i)^T(-\Delta b_i)\\ &= (y_i - X_i\beta - Z_i b_i)^T(y_i - X_i\beta - Z_i b_i)\\ &\quad + (y_i\mathbf{0} - X_i\mathbf{0} - \Delta b_i)^T(y_i\mathbf{0} - X_i\mathbf{0} - \Delta b_i)\\ &= (\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)^T(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)\\ &= \left\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\right\|^2\end{aligned}$$

where

$$\bar{X}_i = \begin{bmatrix} X_i \\ 0 \end{bmatrix}, \bar{y}_i = \begin{bmatrix} y_i \\ 0 \end{bmatrix}, \bar{Z}_i = \begin{bmatrix} Z_i \\ \Delta \end{bmatrix} \tag{2.21}$$

In this case the contribution of the marginal distributions of the random effects is changed into extra rows for the response and the design matrices. This is known as a pseudo-data approach due to the addition of "pseudo" observations. Thus equation (2.20) becomes

$$\begin{aligned}L(\beta,\theta,\sigma^2|\mathbf{y}) = \prod_{i=1}^{M} &\frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}}\\ &\times \int \left\{ \frac{\exp -[\left\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\right\|^2]/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} db_i\end{aligned} \tag{2.22}$$

In this integral, the exponent is in the form of a squared norm or a residual sum of squares. To determine the conditional modes b_i given the data we minimize this residual sum of squares,

$$\left\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\right\|^2 = (\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)^T(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)$$

We minimize by differentiating the above equation with respect to b_i and equalizing the resultant by zero. Thus we have

$$\frac{\partial}{\partial b_i}[(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)^T(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)] = 0$$

This implies that

$$-2\bar{Z}_i^T(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i) = 0 \Rightarrow \bar{Z}_i^T(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i) = 0,$$

which in turn also implies that

$$\bar{Z}_i^T\bar{Z}_i b_i = \bar{Z}_i^T\bar{y}_i - \bar{Z}_i^T\bar{X}_i\beta \Rightarrow \bar{Z}_i^T\bar{Z}_i b_i = \bar{Z}_i^T(\bar{y}_i - \bar{X}_i\beta)$$

and so

$$\hat{b}_i = (\bar{Z}_i^T \bar{Z}_i)^{-1} \bar{Z}_i^T (\bar{y}_i - \bar{X}_i \beta)$$

The squared norm can then be expressed as

$$\begin{aligned} \left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i \right\|^2 &= \left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2 + \left\| \bar{Z}_i (b_i - \hat{b}_i) \right\|^2 \\ &= \left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2 + (b_i - \hat{b}_i)^T \bar{Z}_i^T \bar{Z}_i (b_i - \hat{b}_i) \end{aligned} \tag{2.23}$$

Substituting (2.23) into (2.22) we get

$$L(\beta, \theta, \sigma^2 | \mathbf{y}) = \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \int \left\{ \frac{\exp -[\left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2 + (b_i - \hat{b}_i)^T \bar{Z}_i^T \bar{Z}_i (b_i - \hat{b}_i)]/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} dbi \tag{2.24}$$

The first part in (2.23) does not depend on b_i and thus its exponential can be factored out of (2.24) to get

$$\begin{aligned} L(\beta, \theta, \sigma^2 | \mathbf{y}) &= \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \exp - \frac{\left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2}{2\sigma^2} \int \left\{ \frac{\exp -[(b_i - \hat{b}_i)^T \bar{Z}_i^T \bar{Z}_i (b_i - \hat{b}_i)]/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} dbi \\ &= \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \exp - \frac{\left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2}{2\sigma^2} \int \frac{(\sqrt{|\bar{Z}_i^T \bar{Z}_i|})^{-1}}{(\sqrt{|\bar{Z}_i^T \bar{Z}_i|})^{-1} (2\pi\sigma^2)^{q/2}} \exp \left\{ \frac{-(b_i - \hat{b}_i)^T \bar{Z}_i^T \bar{Z}_i (b_i - \hat{b}_i)}{2\sigma^2} \right\} db_i \\ &= \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \exp - \frac{\left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2}{2\sigma^2} . (\sqrt{|\bar{Z}_i^T \bar{Z}_i|})^{-1} \\ &\quad \times \int \frac{1}{(2\pi\sigma^2)^{q/2} (\sqrt{|\bar{Z}_i^T \bar{Z}_i|})^{-1}} \exp \left\{ \frac{-(b_i - \hat{b}_i)^T \bar{Z}_i^T \bar{Z}_i (b_i - \hat{b}_i)}{2\sigma^2} \right\} db_i \end{aligned} \tag{2.25}$$

The integral part in the above equation is multivariate normal with mean 0 and variance 1. Hence the equation above becomes

$$\begin{aligned} L(\beta, \theta, \sigma^2 | \mathbf{y}) &= \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \frac{\exp - \left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2 / 2\sigma^2}{\sqrt{|\bar{Z}_i^T \bar{Z}_i|}} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} . \exp\{-\Sigma_{i=1}^{M} \left\| \bar{y}_i - \bar{X}_i\beta - \bar{Z}_i \hat{b}_i \right\|^2 / 2\sigma^2\} \prod_{i=1}^{M} \frac{abs|\Delta|}{\sqrt{|\bar{Z}_i^T \bar{Z}_i|}} \end{aligned} \tag{2.26}$$

□

This equation could be used directly in an optimization routine to calculate the maximum likelihood estimates for β, θ and σ^2. The optimization is much simpler if we first concentrate or profile the likelihood so it is a function of θ only. This means calculating the conditional estimates $\hat{\beta}(\theta)$ and $\hat{\sigma}(\theta)$ as the values that maximize $L(\beta, \theta, \sigma^2)$ for a given θ.

2.1.2 Evaluating the Likelihood Through Decomposition

From our augmented model matrices (2.21) and out of the linear mixed-effects model, the orthogonal triangular decomposition of the augmented model matrix $\bar{Z}_i$ is

$$\bar{Z}_i = Q_i \begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix} \tag{2.27}$$

where $Q_{(i)}$ is $(n_i + q) \times (n_i + q)$ and R_{11} is $q \times q$.

Lemma 2. *The likelihood function of the model (2.1) in orthogonal triangular decomposition is given as*

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-N/2} \exp\left\{ \frac{\|c_{-1}\|^2 + \|c_0 - R_{00}\beta\|^2}{-2\sigma^2} \right\} \prod_{i=1}^{M} abs \left\{ \frac{|\Delta|}{|R_{11(i)}|} \right\},$$

where c is as defined in the Appendix A.5 and c_{-1}, c_0, R_{00} and $R_{11(i)}$ are as applied in the following definitions:

1. *$R_{10(i)}$ is a $q \times p$ matrix and $R_{00(i)}$ is a $n_i \times p$ matrix, defined as*

$$\begin{bmatrix} R_{10(i)} \\ R_{oo(i)} \end{bmatrix} = Q_{(i)}^T \bar{X}_i \tag{2.28}$$

and the q-vector $c_{1(i)}$ and the n_i-vector $c_{0(i)}$ may be defined as

$$\begin{bmatrix} c_{1(i)} \\ c_{o(i)} \end{bmatrix} = Q_{(i)}^T \bar{y}_i \tag{2.29}$$

2. *or we may think of these matrices as components in an orthogonal triangular decomposition of the augmented matrix*

$$\begin{bmatrix} Z_i & X_i & y_i \\ \Delta & 0 & 0 \end{bmatrix} = Q_{(i)} \begin{bmatrix} R_{11(i)} & R_{10(i)} & c_{1(i)} \\ 0 & R_{00(i)} & c_{0(i)} \end{bmatrix} \tag{2.30}$$

in which case reduction to triangular form is stopped after the first q columns,

3. *and*

$$\begin{bmatrix} R_{00(i)} & c_{0(i)} \\ \vdots & \vdots \\ R_{00(M)} & c_{0(M)} \end{bmatrix} = Q_0 \begin{bmatrix} R_{00} & c_0 \\ 0 & c_{-1} \end{bmatrix} \tag{2.31}$$

which is a **peculiar numbering scheme** *for the* **sub matrices** *and* **sub vectors** *designed to allow easy extension to more than one level of random effects.*

Proof. From Appendix A.4, we have

$$\begin{aligned}
\left\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\right\|^2 &= \left\|Q^T_{(i)}(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)\right\|^2 \\
&= \left\|Q^T_{(i)}\bar{y}_i - Q^T_{(i)}\bar{X}_i\beta - Q^T_{(i)}\bar{Z}_i b_i\right\|^2 \\
&= \left\|c_{(i)} - Q^T_{(i)}Q_{(i)}\begin{bmatrix} R_{(i)} \\ 0 \end{bmatrix}\beta - Q^T_{(i)}Q_{(i)}\begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix} b_i\right\|^2 \qquad (2.32) \\
&= \left\|c_{(i)} - \begin{bmatrix} R_{(i)} \\ 0 \end{bmatrix}\beta - \begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix} b_i\right\|^2 \\
&= \left\|c_{1(i)} - R_{10(i)}\beta - R_{11(i)}b_i\right\|^2 + \left\|c_{0(i)} - R_{00(i)}\beta\right\|^2
\end{aligned}$$

The integral in (2.22),

$$I = \int \left\{ \frac{\exp -[\left\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\right\|^2]/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} db_i \qquad (2.33)$$

now becomes

$$\begin{aligned}
I &= \int \left\{ \frac{\exp -[\left\|c_{1(i)} - R_{10(i)}\beta - R_{11(i)}b_i\right\|^2 + \left\|c_{0(i)} - R_{00(i)}\beta\right\|^2]/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} dbi \\
&= \exp\left\{ \frac{\left\|c_{0(i)} - R_{00(i)}\beta\right\|^2}{-2\sigma^2} \right\} \int \frac{\left\{\exp \frac{\left\|c_{1(i)} - R_{10(i)}\beta - R_{11(i)}b_i\right\|^2}{-2\sigma^2}\right\}}{(2\pi\sigma^2)^{q/2}} db_i
\end{aligned} \qquad (2.34)$$

where $R_{11(i)}$ is nonsingular. Performing a change of variable,

$$\phi_i = (c_{1(i)} - R_{10(i)}\beta - R_{11(i)}b_i)/\sigma \Rightarrow d\phi_i = -\sigma^{-q}R_{11(i)}db_i = \sigma^{-q}abs|R_{11(i)}|db_i$$

This implies that

$$d\phi_i = \sigma^{-q}abs|R_{11(i)}|db_i \Rightarrow db_i = \left\{\frac{d\phi_i}{\sigma^{-q}abs|R_{11(i)}|}\right\} = \left\{\frac{\sigma^q db_i}{abs|R_{11(i)}|}\right\} \qquad (2.35)$$

This implies that the last integral in (2.34) now becomes

$$\begin{aligned}
\int \frac{\left\{\exp \frac{\left\|c_{1(i)} - R_{10(i)}\beta - R_{11(i)}b_i\right\|^2}{-2\sigma^2}\right\}}{(2\pi\sigma^2)^{q/2}} db_i &= \frac{1}{abs|R_{11(i)|}} \int \frac{\exp(-\left\|\phi\right\|^2/2)}{(2\pi)^{q/2}} d\phi_i \qquad (2.36) \\
&= \frac{1}{abs|R_{11(i)}|}
\end{aligned}$$

It is important to note that the result in (2.36) is the same as the denominator in the product of (2.26), that is,

$$\begin{aligned}
\sqrt{|\bar{Z}_i^T \bar{Z}_i|} &= \sqrt{|[R_{11(i)}^T 0] Q_{(i)}^T Q_{(i)} \begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix}|} \\
&= \sqrt{|R_{11(i)}^T R_{11(i)}|} \\
&= \sqrt{|R_{11(i)}^T||R_{11(i)}|} \\
&= \sqrt{(|R_{11(i)}^T|)^2} \\
&= abs|R_{11(i)}|
\end{aligned} \tag{2.37}$$

$R_{11(i)}$ is a triangular matrix whose determinant is simply the product of its diagonal elements. Substituting (2.36) and (2.34) into (2.20) we get

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = \frac{\exp(-\sum_{i=1}^{M} \left\| c_{0(i)} - R_{00(i)}\beta \right\|^2 / 2\sigma^2)}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} abs\left\{\frac{|\Delta|}{|R_{11(i)}|}\right\} \tag{2.38}$$

It is worth noting that the term in the exponent has the form of a **residual sum of squares** for β pooled over all the groups. From (2.31), we have

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-N/2} \exp\left\{\frac{\|c_{-1}\|^2 + \|c_0 - R_{00}\beta\|^2}{-2\sigma^2}\right\} \prod_{i=1}^{M} abs\left\{\frac{|\Delta|}{|R_{11(i)}|}\right\} \tag{2.39}$$

since

$$\begin{aligned}
\left\| c_{0(i)} - R_{00(i)}\beta \right\|^2 &= \left\| Q_0^T (c_{0(i)} - R_{00(i)}\beta) \right\|^2 \\
&= \left\| Q_0^T c_{0(i)} - Q_0^T R_{00(i)}\beta \right\|^2 \\
&= \left\| Q_0^T Q_0 \begin{bmatrix} c_0 \\ c_{-1} \end{bmatrix} - Q_0^T Q_0 \begin{bmatrix} R_{00} \\ 0 \end{bmatrix} \beta \right\|^2 \\
&= \left\| \begin{bmatrix} c_0 \\ c_{-1} \end{bmatrix} - \begin{bmatrix} R_{00} \\ 0 \end{bmatrix} \beta \right\|^2 \\
&= \|c_0 - R_{00}\beta\|^2 + \|c_{-1}\|^2
\end{aligned} \tag{2.40}$$

□

For a given θ, values of β and σ^2 that maximize (2.39) are

$$\begin{aligned}\frac{\partial L(\beta,\theta,\sigma^2|\mathbf{y})}{\partial\beta} &= 0 \\ &\Rightarrow \frac{-2R_{00}(c_0 - R_{00}\beta)}{-2\sigma^2} = 0 \\ &\Rightarrow R_{00}c_o - R_{00}^2\beta = 0 \\ &\Rightarrow \hat{\beta}(\theta) = R_{00}^{-1}c_0\end{aligned}$$

and by writing (2.39) as a log-likelihood, that is,

$$logL(\beta,\theta,\sigma^2|\mathbf{y}) = (-\frac{N}{2})log(2\pi) + (-\frac{N}{2})log(\sigma^2) - \frac{1}{2\sigma^2}(\|c_{-1}\|^2 + \|c_0 - R_{00}\beta\|^2)$$

then we have,

$$\begin{aligned}\frac{\partial logL(\beta,\theta,\sigma^2|\mathbf{y})}{\partial\sigma^2} &= 0 \\ &\Rightarrow \frac{-N}{2\sigma^2} + \frac{\|c_{-1}\|^2 + \|c_0 - R_{00}\beta\|^2}{2\sigma^4} = 0 \\ &\Rightarrow \hat{\sigma}^2(\theta) = \frac{\|c_{-1}\|^2 + \left\|c_0 - R_{00}\hat{\beta}\right\|^2}{N}\end{aligned}$$

but

$$\hat{\beta}(\theta) = R_{00}^{-1}c_0 \tag{2.41}$$

which implies that

$$\begin{aligned}\hat{\sigma}^2(\theta) &= \frac{\|c_{-1}\|^2}{N} + \frac{\left\|c_0 - R_{00}R_{00}^{-1}c_0\right\|^2}{N} \\ &= \frac{\|c_{-1}\|^2}{N}\end{aligned} \tag{2.42}$$

which leads to the likelihood function as

$$\begin{aligned}L(\theta|\mathbf{y}) &= L(\hat{\beta}(\theta),\theta,\hat{\sigma}^2(\theta)|\mathbf{y}) \\ &= \frac{1}{(2\pi\hat{\sigma}^2(\theta))^{N/2}}\exp\left\{\frac{\|c_{-1}\|^2 + \left\|c_0 - R_{00}\hat{\beta}(\theta)\right\|^2}{-2\hat{\sigma}^2(\theta)}\right\}\prod_{i=1}^{M} abs\left\{\frac{|\Delta|}{|R_{11(i)}|}\right\} \qquad (2.43) \\ &= \left\{\frac{N}{2\pi\left\|c_{-1}\right\|^2}\right\}^{N/2}\exp\left\{-\frac{N}{2}\right\}\prod_{i=1}^{M} abs\left\{\frac{|\Delta|}{|R_{11(i)}|}\right\}\end{aligned}$$

or the profiled log-likelihood function as

$$
\begin{aligned}
l(\theta|\mathbf{y}) &= logL(\theta|\mathbf{y}) \\
&= \frac{N}{2}log(N) - \frac{N}{2}log(2\pi) - \frac{N}{2}log(\|c_{-1}\|^2) - \frac{N}{2} + \sum_{i=1}^{M} log\left\{abs(\frac{|\Delta|}{|R_{11(i)}|})\right\} \\
&= \frac{N}{2}[log(N) - log(2\pi) - 1] - Nlog\,\|c_{-1}\| + \sum_{i=1}^{M} log\left\{abs(\frac{|\Delta|}{|R_{11(i)}|})\right\}
\end{aligned}
\tag{2.44}
$$

This profiled log-likelihood is maximized with respect to θ, producing the maximum likelihood estimator $\hat{\theta}$. The maximum likelihood estimators $\hat{\theta}$ and $\hat{\sigma}^2$ are obtained by setting $\theta = \hat{\theta}$ in (2.41) and in (2.42).
Technically the random effects b_i **are not parameters** for the statistical model although they do behave in some way s like parameters .
The conditional modes of the random effects, evaluated at the conditional estimate of β, are the **Best Linear Unbiased Predictors or BLUPs** of the b_i, i=1,$\cdots$,M.
These conditional modes can be evaluated using the matrices from the orthogonal triangular decomposition.
From the exponent value in (2.34), which is the residual sum-of-squares, we can determine the conditional modes of the random effects at the conditional estimate of β, by minimizing these residual sum-of-squares, that is,

$$
\begin{aligned}
&\frac{\partial}{\partial b_i}\left\|c_{1(i)} - R_{10(i)}\beta - R_{11(i)}b_i\right\|^2 = 0 \\
&\qquad \Rightarrow -2R_{11(i)}(c_{1(i)} - R_{10(i)}\beta - R_{11(i)}b_i) = 0 \\
&\qquad \Rightarrow \hat{b}_i(\theta) = R_{11(i)}^{-1}(c_{1(i)} - R_{10(i)}\hat{\beta}(\theta))
\end{aligned}
\tag{2.45}
$$

Partially, the unknown vector θ is replaced by its maximum likelihood estimator $\hat{\theta}$, producing estimated BLUPs $\hat{b}_i(\hat{\theta})$

2.1.3 Restricted Likelihood Estimation (REML)

Maximum likelihood estimates (MLE) of 'variance components' tend to underestimate these parameters. An alternative to this is the **restricted** or **residual maximum likelihood (REML)** and very much prefered by Patterson and Thompson, (1971); Harville, (1977), for estimation of variance components.
Laird and Ware (1982) defined the REML estimation criterion as

$$
L_R(\theta, \sigma^2|\mathbf{y}) = \int L(\beta, \theta, \sigma^2|\mathbf{y})d\beta \tag{2.46}
$$

These, within a Bayesian framework, corresponds to assuming a locally uniform prior distribution for the **fixed effects** β and integrating them out of the likelihood, hence we have the following Lemma:

Lemma 3. *The restricted (residual) maximum likelihood (REML) function of the model (2.1) in orthogonal triangular decomposition is given as*

$$L_R(\theta, \sigma^2|\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{N-p}{2}}} \frac{\exp - \|c_{-1}\|^2/(2\sigma^2)}{abs|R_{00}|} \prod_{i=1}^{M} abs\left\{\frac{|\Delta|}{|R_{11(i)}|}\right\}$$

where c_{-1}, R_{00} and $R_{11(i)}$ are as defined above.

Proof. As earlier found, the likelihood function was given as

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \int \left\{\frac{\exp -[\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\|^2]/2\sigma^2}{(2\pi\sigma^2)^{q/2}}\right\} db_i \tag{2.47}$$

where $\bar{y}_i$, $\bar{X}_i$ and $\bar{X}_i$ are the augmented data vectors and model matrices. This likelihood function leads to

$$\int L(\beta, \theta, \sigma^2|\mathbf{y}) d\beta = \int \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \int \left\{\frac{\exp -[\|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\|^2]/2\sigma^2}{(2\pi\sigma^2)^{q/2}}\right\} db_i d\beta \tag{2.48}$$

Through Orthogonal triangular decomposition where

$$\begin{bmatrix} Z_i & X_i & y_i \\ \Delta & 0 & 0 \end{bmatrix} = Q_{(i)} \begin{bmatrix} R_{11(i)} & R_{10(i)} & c_{1(i)} \\ 0 & R_{00(i)} & c_{0(i)} \end{bmatrix} \tag{2.49}$$

and

$$\begin{aligned} \|\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i\|^2 &= \|Q_{(i)}^T(\bar{y}_i - \bar{X}_i\beta - \bar{Z}_i b_i)\|^2 \\ &= \|Q_{(i)}^T\bar{y}_i - Q_{(i)}^T\bar{X}_i\beta - Q_{(i)}^T\bar{Z}_i b_i\|^2 \\ &= \left\| c_{(i)} - Q_{(i)}^T Q_{(i)} \begin{bmatrix} R_{(i)} \\ 0 \end{bmatrix} \beta - Q_{(i)}^T Q_{(i)} \begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix} b_i \right\|^2 \\ &= \left\| c_{(i)} - \begin{bmatrix} R_{(i)} \\ 0 \end{bmatrix} \beta - \begin{bmatrix} R_{11(i)} \\ 0 \end{bmatrix} b_i \right\|^2 \\ &= \|c_{1(i)} - R_{10(i)}\beta - R_{11(i)} b_i\|^2 + \|c_{0(i)} - R_{00(i)}\beta\|^2 \end{aligned} \tag{2.50}$$

we have

$$\int L(\beta, \theta, \sigma^2|\mathbf{y}) d\beta = \int \left\{ \prod_{i=1}^{M} \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \exp\left\{\frac{\|c_{0(i)} - R_{00(i)}\beta\|^2}{-2\sigma^2}\right\} \int \frac{\left\{\exp \frac{\|c_{1(i)} - R_{10(i)}\beta - R_{11(i)} b_i\|^2}{-2\sigma^2}\right\}}{(2\pi\sigma^2)^{q/2}} db_i \right\} d\beta \tag{2.51}$$

Using the same change-of-variable technique as in (2.35) and same orthogonal triangular decomposition as in (2.31) which led to (2.40) we have

$$\int L(\beta,\theta,\sigma^2|\mathbf{y})d\beta = \int \prod_{i=1}^{M} \exp\frac{\|c_{-1}\|^2}{-2\sigma^2}\exp\frac{\|c_0 - R_{00}\beta\|^2/(-2\sigma^2)}{(2\pi\sigma^2)^{n_i/2}} abs\left\{\frac{|\Delta|}{|R_{11(i)}|}\right\} d\beta \quad (2.52)$$

Applying same change-of-variable technique as in (2.35);

Let

$$\begin{aligned} u &= (c_0 - R_{00}\beta)/\sigma \\ \Rightarrow du &= \frac{-R_{00}d\beta}{\sigma^p} \\ \Rightarrow d\beta &= \frac{\sigma^p du}{abs|R_{00}|} \end{aligned} \quad (2.53)$$

Thus we have

$$\begin{aligned} \prod_{i=1}^{M}\int \exp\frac{\|c_0 - R_{00}\beta\|^2/(-2\sigma^2)}{(2\pi\sigma^2)^{n_i/2}}d\beta &= \int \exp\frac{-\|u\|^2/2}{(2\pi\sigma^2)^{N/2}} \times \frac{\sigma^p du}{abs|R_{00}|} \\ &= \frac{\sigma^p}{abs|R_{00}|}\int \exp\frac{-\|u\|^2/2}{(2\pi\sigma^2)^{p/2}}du \times \frac{1}{(2\pi\sigma^2)^{(N/2-p/2)}} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N-p}{2}}abs|R_{00}|}\int \exp\frac{-\|u\|^2/2}{(2\pi)^{p/2}}du \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N-p}{2}}abs|R_{00}|} \end{aligned} \quad (2.54)$$

Incorporating (2.52) and (2.54) in (2.51) we end up getting the REML as given by Laird and Ware (1982),

$$\begin{aligned} L_R(\theta,\sigma^2|\mathbf{y}) &= \int L(\beta,\theta,\sigma^2|\mathbf{y})d\beta \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N-p}{2}}}\frac{\exp-\|c_{-1}\|^2/(2\sigma^2)}{abs|R_{00}|}\prod_{i=1}^{M} abs\left\{\frac{|\Delta|}{|R_{11(i)}|}\right\} \end{aligned} \quad (2.55)$$

□

Introducing logarithms we get the log-restricted likelihood of the above REML function as:

$$\begin{aligned} l_R(\theta,\sigma^2|\mathbf{y}) &= logL_R(\theta,\sigma^2|\mathbf{y}) \\ &= \frac{-(N-p)}{2}log(2\pi\sigma^2) - \frac{\|c_{-1}\|^2}{2\sigma^2} - log(abs|R_{00}|) + \sum_{i=1}^{M} log\left\{abs(\frac{|\Delta|}{|R_{11(i)}|})\right\} \end{aligned} \quad (2.56)$$

producing the conditional estimate $\hat{\sigma}_R^2(\theta)$ as

$$\begin{aligned}\frac{\partial l_R(\theta,\sigma^2|\mathbf{y})}{\partial\sigma^2} &= 0\\ &\Rightarrow -\frac{N-p}{2\sigma^2}+\frac{\|c_{-1}\|^2}{2\sigma^4}=0\\ &\Rightarrow \hat{\sigma}_R^2(\theta)=\frac{\|c_{-1}\|^2}{N-p}\end{aligned} \tag{2.57}$$

from where we obtain the profiled log-restricted-likelihood by substituting the estimate $\hat{\sigma}_R^2(\theta)$ as

$$\begin{aligned} l_R(\theta|\mathbf{y}) &= l_R(\theta,\hat{\sigma}_R^2(\theta)|\mathbf{y})\\ &= \frac{-(N-p)}{2}log(2\pi\hat{\sigma}_R^2(\theta)) - \frac{\|c_{-1}\|^2}{2\hat{\sigma}_R^2(\theta)} - log(abs|R_{00}|) + \sum_{i=1}^{M} log\left\{abs(\frac{|\Delta|}{|R_{11(i)}|})\right\}\\ &= \frac{-(N-p)}{2}log\left\{\frac{2\pi\|c_{-1}\|^2}{N-p}\right\} - \frac{\|c_{-1}\|^2(N-p)}{2\|c_{-1}\|^2} - log(abs|R_{00}|) + \sum_{i=1}^{M} log\left\{abs(\frac{|\Delta|}{|R_{11(i)}|})\right\}\\ &= \frac{-(N-p)}{2}log\left\{\frac{2\pi}{N-p}\right\} - \frac{(N-p)}{2}.2log\|c_{-1}\|\\ &\quad - \frac{(N-p)}{2} - log(abs|R_{00}|) + \sum_{i=1}^{M} log\left\{abs(\frac{|\Delta|}{|R_{11(i)}|})\right\}\\ &= constant - (N-p)log\|c_{-1}\| - log(abs|R_{00}|) + \sum_{i=1}^{M} log\left\{abs(\frac{|\Delta|}{|R_{11(i)}|})\right\}\end{aligned} \tag{2.58}$$

Components of the profiled log-restricted-likelihood in (2.58) are similar to those in the profiled log-likelihood in (2.44) except that the log of norm of the residual vector has a different multiplier and there is an extra determinant term of

$$log(abs|R_{00}|) = log|\sum_{i=1}^{M} X_i^T\Sigma_i^{-1}X_i|/2 \tag{2.59}$$

Evaluation of the restricted maximum likelihood estimates (REML) is done by optimizing the profiled log-restricted-likelihood (2.58) with respect to θ only.
Using the resulting REML estimate $\hat{\theta}_R$, the REML estimate of σ^2 can be obtained as $\hat{\sigma}_R(\hat{\theta}_R)$. REML estimated BLUPs of the random effects are obtained by re placing θ with $\hat{\theta}_R$ in (2.45) to become

$$\hat{b}_i(\hat{\theta}_R) = R_{11(i)}^{-1}(c_{1(i)} - R_{10(i)}\hat{\beta}(\hat{\theta}_R)) \tag{2.60}$$

REML criterion only depends on θ and σ. In REML criterion we do not speak of REML estimates of β but we can evaluate the "best guess" at β from (2.41) once $\hat{\theta}_R$has been determined.
Major difference between the likelihood function and the restricted likelihood function is that likelihood function is invariant to one-to-one reparametirizations of the fixed effects (for

example, a change in the contrasts representing a categorical variable) while the latter is not.
Changing the X_i matrices results in change in $log(abs|R_{00}|)$ and a corresponding change in $l_R(\theta|\mathbf{y})$, i.e the profiled log-restricted-likelihood. Thus LME models with different fixed effects structures fit using REML cannot be compared on the basis of their restricted likelihoods. Hence under such circumstances the likelihood ratio tests are not valid.

2.2 An alternative derivation to penalized Least-Squares Problem

Consider the general form of a Mixed Effect Model with a single level of random effects labeled (2.1), that is,

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i,\ i = 1, \dots, m \tag{2.61}$$

where β, and ϵ_i are as defined in (2.1). $\mathbf{X}_i$ is a known fixed effects regressor $(n_i \times p)$ matrix for the p-dimensional fixed-effects vector β, and $\mathbf{Z}_i$ is a known random effects regressor $(n_i \times q)$ matrix for the q-dimensional random-effects vector $\mathbf{b}_i$. Noting that Δ is a relative precision factor of the random effects, expressed relative to the precision $1/\sigma^2$, of the ϵ_i, we can have $\mathbf{b}_i \sim N(0, \sigma^2(\Delta^T\Delta)^{-1})$. This is because a relative precision factor is any $q \times q$ matrix that satisfies

$$\psi = \Delta^T\Delta$$

One possible Δ is the Cholesky factor (see Appendix A.2.1) of ψ, which will be non-singular since the later is positive definite. Others can be used.
Generally, Δ (and hence ψ) depend on a k-dimensional parameter vector θ. Typically $\mathbf{Z}$, ψ and Δ are very large and sparse (mostly zeros) while k, the dimension of θ, is small.
$\mathbf{X}_i$ has full column rank and hence $\Delta^T\Delta$ is positive definite. If $\mathbf{X}_i$ is rank deficient meaning that $\Delta^T\Delta$ is singular, then the model can be transformed to an alternative model that fulfills the desired conditions.
The parameters of the model are β,σ^2, and whatever parameters determine Δ. We use θ to represent an unconstrained set of parameters that determine Δ. The likelihood function for the parameters β,θ and σ^2 of the model (2.61) is given as

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{(n+q)}{2}}} \int \exp\left\{\frac{-[\|(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})\|^2 + \|\Delta\mathbf{b}\|^2]}{2\sigma^2}\right\} d\mathbf{b} \tag{2.62}$$

which is just like an earlier given equation defined in (2.20) The restricted (or residual) maximum likelihood estimates (REML) of θ and σ^2 optimize a related criterion that can be written as

$$L_R(\theta, \sigma^2|\mathbf{y}) = \int L(\beta, \theta, \sigma^2|\mathbf{y}) d\beta \tag{2.63}$$

as defined by Laird and Ware (1982).

Expressions (2.62) and (2.63) can be expressed succinctly (expressed briefly and clearly) using the solution to a penalized least-squares problem.
For fixed value of θ, the penalized least-squares problem could be defined by

- the augmented model matrix Ψ and
- the augmented response vector $\bar{\mathbf{y}}$

leading to

$$\min_{\mathbf{b},\beta} \left\| \bar{\mathbf{y}} - \Psi(\theta) \begin{bmatrix} \mathbf{b} \\ \beta \end{bmatrix} \right\|^2$$

where

$$\Psi(\theta) = \begin{bmatrix} \mathbf{Z} & \mathbf{X} \\ \Delta(\theta) & \mathbf{0} \end{bmatrix}$$

and

$$\bar{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$$

Hence we have

$$\begin{aligned} \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z} & \mathbf{X} \\ \Delta(\theta) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \beta \end{bmatrix} \right\|^2 &= \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\mathbf{b} + \mathbf{X}\beta \\ \Delta(\theta)\mathbf{b} + \mathbf{0} \end{bmatrix} \right\|^2 \\ &= \left\| \begin{bmatrix} \mathbf{y} - \mathbf{Z}\mathbf{b} - \mathbf{X}\beta \\ \mathbf{0} - \Delta(\theta)\mathbf{b} + \mathbf{0} \end{bmatrix} \right\|^2 \end{aligned} \tag{2.64}$$

By forming

$$\Psi_e = \begin{bmatrix} \Psi & \bar{\mathbf{y}} \end{bmatrix}$$

and letting R_e be the Cholesky decomposition (see Appendix A.2.1) of $\Psi_e^T \Psi_e$, then we can have,

$$\begin{aligned} \Psi_e &= \begin{pmatrix} \Psi & \bar{\mathbf{y}} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Z} & \mathbf{X} & \mathbf{y} \\ \Delta & \mathbf{0} & \mathbf{0} \end{pmatrix} \end{aligned}$$

which implies

$$\begin{aligned} \Psi_e^T \Psi_e &= \begin{pmatrix} \mathbf{Z}^T & \Delta^T \\ \mathbf{X}^T & \mathbf{0}^T \\ \mathbf{y}^T & \mathbf{0}^T \end{pmatrix} \begin{pmatrix} \mathbf{Z} & \mathbf{X} & \mathbf{y} \\ \Delta & \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta & \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{y} \\ \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{y} \\ \mathbf{y}^T\mathbf{Z} & \mathbf{y}^T\mathbf{X} & \mathbf{y}^T\mathbf{y} \end{pmatrix} \\ &= R_e^T R_e \end{aligned} \tag{2.65}$$

where

$$R_e = \begin{pmatrix} R_{ZZ} & R_{ZX} & r_{Zy} \\ \mathbf{0} & R_{XX} & r_{Xy} \\ \mathbf{0} & \mathbf{0} & r_{yy} \end{pmatrix} \tag{2.66}$$

Matrices R_{ZZ}, a $q \times q$ and R_{XX}, a $p \times p$ are upper triangular. Conditions that ψ is positive definite and $\mathbf{X}$ is of full column rank ensures that Ψ has full column rank, hence R_{ZZ} and R_{XX} are non-singular. Corresponding vectors, r_{Zy} is q-dimensional, r_{Xy} is p-dimensional and r_{yy} is a scalar.
Representation (2.66) is a particular form of the mixed model equations by Henderson (1984).

Lemma 4. *The maximum likelihood deviance is given as*

$$-2l(\beta, \theta, \sigma^2) = log\left\{\frac{|\mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta|}{|\Delta^T\Delta|}\right\} + nlog(2\pi\sigma^2) + \frac{r_{yy}^2 + \|r_{Xy} - R_{XX}\beta\|^2}{\sigma^2}$$

Proof. Writing the blocks in the opposite order from which they are typically written, that is,

$$\mathbf{a} = (-\mathbf{b}^T, -\beta^T, 1)^T,$$

we can write the numerator of the exponent in the integral in (2.62) as

$$\|(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})\|^2 + \|\Delta\mathbf{b}\|^2 = \mathbf{a}^T\Psi_e^T\Psi_e\mathbf{a}$$

Thus,

$$\begin{aligned} \mathbf{a}^T\Psi_e^T\Psi_e\mathbf{a} &= \mathbf{a}^T\begin{pmatrix} \Psi^T \\ \bar{\mathbf{y}}^T \end{pmatrix}\begin{pmatrix} \Psi & \bar{\mathbf{y}} \end{pmatrix}\mathbf{a} \\ &= \mathbf{a}^T\begin{pmatrix} \Psi^T\Psi & \Psi^T\bar{\mathbf{y}} \\ \bar{\mathbf{y}}^T\Psi & \bar{\mathbf{y}}^T\bar{\mathbf{y}} \end{pmatrix}\mathbf{a} \\ &= \mathbf{a}^T\begin{pmatrix} \mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta & \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{y} \\ \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{y} \\ \mathbf{y}^T\mathbf{Z} & \mathbf{y}^T\mathbf{X} & \mathbf{y}^T\mathbf{y} \end{pmatrix}\mathbf{a} \\ &= (-\mathbf{b}^T, -\beta^T, 1)\begin{pmatrix} \mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta & \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{y} \\ \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{y} \\ \mathbf{y}^T\mathbf{Z} & \mathbf{y}^T\mathbf{X} & \mathbf{y}^T\mathbf{y} \end{pmatrix}\begin{pmatrix} -\mathbf{b}^T \\ -\beta^T \\ 1 \end{pmatrix} \end{aligned} \tag{2.67}$$

whose expansion is equivalent to expanding

$$\|(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})\|^2 + \|\Delta\mathbf{b}\|^2$$

But from

$$\Psi_e^T \Psi_e = R_e^T R_e$$

where R_e is as defined in (2.66), we have

$$\begin{aligned}
\mathbf{a}^T \Psi_e^T \Psi_e \mathbf{a} &= \mathbf{a}^T R_e^T R_e \mathbf{a} \\
&= \|R_e \mathbf{a}\|^2 \\
&= \left\| \begin{pmatrix} R_{ZZ} & R_{ZX} & r_{Zy} \\ \mathbf{0} & R_{XX} & r_{Xy} \\ \mathbf{0} & \mathbf{0} & r_{yy} \end{pmatrix} \begin{pmatrix} -\mathbf{b}^T \\ -\beta^T \\ 1 \end{pmatrix} \right\|^2 \\
&= \left\| \begin{pmatrix} -R_{ZZ}\mathbf{b} - R_{ZX}\beta + r_{Zy} \\ \mathbf{0} - R_{XX}\beta + r_{Xy} \\ \mathbf{0} + \mathbf{0} + r_{yy} \end{pmatrix} \right\|^2 \\
&= \|r_{Zy} - R_{ZX}\beta - R_{ZZ}\mathbf{b}\|^2 + \|r_{Xy} - R_{XX}\beta\|^2 + r_{yy}^2
\end{aligned} \tag{2.68}$$

Now, from (2.62), and by a simple change of variable, we get the likelihood function as

$$\begin{aligned}
L(\beta, \theta, \sigma^2|\mathbf{y}) &= \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n}{2}}} \int \frac{1}{(2\pi\sigma^2)^{\frac{q}{2}}} \exp\left\{ \frac{-[\|(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})\|^2 + \|\Delta\mathbf{b}\|^2]}{2\sigma^2} \right\} d\mathbf{b} \\
&= \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{ \frac{\|r_{Xy} - R_{XX}\beta\|^2 + r_{yy}^2}{-2\sigma^2} \right\} \\
&\quad \times \int \frac{1}{(2\pi\sigma^2)^{\frac{q}{2}}} \exp\left\{ \frac{\|r_{Zy} - R_{ZX}\beta - R_{ZZ}\mathbf{b}\|^2}{-2\sigma^2} \right\} d\mathbf{b} \\
&= \frac{abs|\Delta|}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{ \frac{\|r_{Xy} - R_{XX}\beta\|^2 + r_{yy}^2}{-2\sigma^2} \right\} \times \frac{1}{\sqrt{|\mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta|}}
\end{aligned} \tag{2.69}$$

The simple change of variable on the integral part leading us to the above result is solved by letting

$$\begin{aligned}
\Omega &= \frac{(r_{Zy} - R_{ZX}\beta - R_{ZZ}\mathbf{b})}{\sigma} \\
&\Rightarrow d\Omega = -\sigma^{-q} R_{ZZ} d\mathbf{b} \\
&\Rightarrow d\Omega = \sigma^{-q} abs|R_{ZZ}| d\mathbf{b} \\
&\Rightarrow d\mathbf{b} = \frac{\sigma^q d\Omega}{abs|R_{ZZ}|}
\end{aligned}$$

resulting in

$$\int \frac{1}{(2\pi\sigma^2)^{\frac{q}{2}}} \exp\left\{\frac{\|r_{Zy} - R_{ZX}\beta - R_{ZZ}\mathbf{b}\|^2}{-2\sigma^2}\right\} d\mathbf{b} = \frac{1}{abs|R_{ZZ}|}\int \frac{-\|\Omega\|^2/2}{(2\pi)^{q/2}} d\Omega$$
$$= \frac{1}{abs|R_{ZZ}|}$$
$$= \frac{1}{\sqrt{|\mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta|}}$$

Taking logarithms of (2.69) and writing it in the form of a deviance gives

$$\begin{aligned} l(\beta,\theta,\sigma^2) &= logL(\beta,\theta,\sigma^2|\mathbf{y}) \\ &= \frac{-n}{2}log(2\pi\sigma^2) - \frac{1}{2}log\left\{\frac{|\mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta|}{|\Delta^T\Delta|}\right\} - \frac{1}{2}\frac{r_{yy}^2 + \|r_{Xy} - R_{XX}\beta\|^2}{\sigma^2} \\ &\Rightarrow -2l(\beta,\theta,\sigma^2) = log\left\{\frac{|\mathbf{Z}^T\mathbf{Z} + \Delta^T\Delta|}{|\Delta^T\Delta|}\right\} + nlog(2\pi\sigma^2) + \frac{r_{yy}^2 + \|r_{Xy} - R_{XX}\beta\|^2}{\sigma^2.} \end{aligned} \tag{2.70}$$

□

Same results were obtained by Bates and DebRoy (2004) This leads to the following results for the maximum likelihood estimates (MLE):

- The conditional MLE of the fixed-effects, $\hat{\beta}(\theta)$,

$$\begin{aligned} \frac{\partial - 2l(\beta,\theta,\sigma^2)}{\partial\beta} &= 0 \\ &\Rightarrow \frac{-2R_{XX}(r_{Xy} - R_{XX}\beta)}{\sigma^2} = 0 \\ &\Rightarrow R_{XX}r_{Xy} - R_{XX}^2\beta = 0 \\ &\Rightarrow \hat{\beta}(\theta) = R_{XX}^{-1}r_{Xy} \end{aligned} \tag{2.71}$$

  which could be equal to $(= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)$ in classical linear modeling.

- The conditional MLE of the variance, $\hat{\sigma}^2(\theta)$,

$$\begin{aligned} \frac{\partial - 2l(\beta,\theta,\sigma^2)}{\partial\sigma^2} &= 0 \\ &\Rightarrow \frac{n}{\sigma^2} - \frac{r_{yy}^2 + \left\|r_{Xy} - R_{XX}\hat{\beta}(\theta)\right\|^2}{\sigma^4} = 0 \\ &\Rightarrow \hat{\sigma}^2(\theta) = \frac{r_{yy}^2 + \left\|r_{XX} - R_{XX}(R_{XX}^{-1}r_{Xy})\right\|^2}{n} \\ &\Rightarrow \hat{\sigma}^2(\theta) = \frac{r_{yy}^2}{n} \end{aligned} \tag{2.72}$$

- The profiled log-likelihood, $\hat{l}(\theta)$, a function of θ only, is given by

$$\begin{aligned} -2\hat{l}(\theta) &= log\left\{\frac{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta|}{|\Delta^T\Delta|}\right\} + nlog(\frac{2\pi r_{yy}^2}{n}) + \frac{r_{yy}^2 + \left\|r_{Xy} - R_{XX}R_{XX}^{-1}r_{Xy}\right\|^2}{\frac{r_{yy}^2}{n}} \\ &= log\left\{\frac{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta|}{|\Delta^T\Delta|}\right\} + n(1+log(\frac{2\pi r_{yy}^2}{n})) \end{aligned} \tag{2.73}$$

- The conditional modes of random effects, evaluated at the conditional estimates of β are evaluated from the exponent value in the integrand value above (which are the residual sum of squares). It is arrived at through minimizing these residual sum of squares;

$$\begin{aligned} \frac{\partial\left\|r_{Zy} - R_{ZX}\beta - R_{ZZ}\mathbf{b}\right\|^2}{\partial\mathbf{b}} &= 0 \\ &\Rightarrow -2R_{ZZ}(r_{Zy} - R_{ZX}\beta - R_{ZZ}\mathbf{b}) = 0 \\ &\Rightarrow \hat{\mathbf{b}}(\beta,\theta) = R_{ZZ}^{-1}(r_{Zy} - R_{ZX}\hat{\beta}(\theta)) \end{aligned} \tag{2.74}$$

  Thus it satisfies

$$R_{ZZ}\hat{\mathbf{b}}(\beta,\theta) = r_{Zy} - R_{ZX}\hat{\beta}(\theta)$$

- The conditional distribution of $\mathbf{b}$ is

$$\mathbf{b}|\mathbf{y},\beta,\theta,\sigma^2 \sim N(\hat{\mathbf{b}}(\beta,\theta), \sigma^2(\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta)^{-1})$$

Lemma 5. *The restricted maximum likelihood deviance is given as*

$$-2l_R(\theta,\sigma^2) = log\left\{\frac{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta||R_{XX}|^2}{|\Delta^T\Delta|}\right\} + (n-p)log(2\pi\sigma^2) + \frac{r_{yy}^2}{\sigma^2}$$

Proof. As in (2.69), we can apply a simple change of variable technique to evaluate the REML by letting

$$\begin{aligned} \omega &= \frac{r_{Xy} - R_{XX}\beta}{\sigma} \\ &\Rightarrow d\omega = \frac{-R_{XX}d\beta}{\sigma^p} \\ &\Rightarrow d\omega = \frac{abs|R_{XX}|d\beta}{\sigma^p} \\ &\Rightarrow d\beta = \frac{\sigma^p d\omega}{abs|R_{XX}|} \end{aligned}$$

Thus (2.63) becomes

$$
\begin{aligned}
L_R(\theta,\sigma^2|\mathbf{y}) &= \int L(\beta,\theta,\sigma^2|\mathbf{y})d\beta \\
&= \frac{abs|\Delta|}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{\|r_{Xy}-R_{XX}\beta\|^2+r_{yy}^2}{-2\sigma^2}\right\} \times \frac{1}{\sqrt{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta|}}d\beta \\
&= \frac{abs|\Delta|}{(2\pi\sigma^2)^{(n-p)/2}} \exp\left\{\frac{r_{yy}^2}{-2\sigma^2\times\sqrt{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta|}}\right\}\int \frac{1}{(2\pi\sigma^2)^{p/2}}\exp\left\{\frac{\|r_{Xy}-R_{XX}\beta\|^2}{-2\sigma^2}\right\}d\beta \\
&= \left\{\frac{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta|}{|\Delta^T\Delta|}\right\}^{-1/2} \exp\left\{\frac{r_{yy}^2}{-2\sigma^2\times abs|R_{XX}|}\right\}\cdot\frac{1}{(2\pi\sigma^2)^{(n-p)/2}} \\
&= \left\{\frac{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta||R_{XX}|^2}{|\Delta^T\Delta|}\right\}^{-1/2} \exp\left\{\frac{r_{yy}^2}{-2\sigma^2}\right\}\times\frac{1}{(2\pi\sigma^2)^{(n-p)/2}}
\end{aligned}
\tag{2.75}
$$

Taking logarithms and giving the result in form of a deviance gives

$$
\begin{aligned}
l_R(\beta,\theta,\sigma^2) &= logL_R(\beta,\theta,\sigma^2|\mathbf{y}) \\
&= -\frac{1}{2}log\left\{\frac{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta||R_{XX}|^2}{|\Delta^T\Delta|}\right\} - \frac{(n-p)}{2}log(2\pi\sigma^2)+\frac{r_{yy}^2}{-2\sigma^2} \\
\Rightarrow -2l_R(\theta,\sigma^2) &= log\left\{\frac{|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta||R_{XX}|^2}{|\Delta^T\Delta|}\right\}+(n-p)log(2\pi\sigma^2)+\frac{r_{yy}^2}{\sigma^2}
\end{aligned}
\tag{2.76}
$$

□

Noting that

$$|\mathbf{Z}^T\mathbf{Z}+\Delta^T\Delta||R_{XX}|^2 = |\Psi^T\Psi|,$$

then we have

- The conditional REML estimate of the variance, $\hat{\sigma}_R^2(\theta)$,

$$
\begin{aligned}
\frac{\partial[-2l_R(\theta,\sigma^2)]}{\partial\sigma^2} &= 0 \\
&\Rightarrow \frac{(n-p)}{\sigma^2}-\frac{r_{yy}^2}{\sigma^4}=0 \\
&\Rightarrow \hat{\sigma}_R^2(\theta) = \frac{r_{yy}^2}{(n-p)}
\end{aligned}
$$

- The profiled log-restricted likelihood is given by

$$
\begin{aligned}
-2l_R(\theta) &= log\left\{\frac{|\Psi^T\Psi|}{|\Delta^T\Delta|}\right\}+(n-p)log\left\{\frac{2\pi r_{yy}^2}{(n-p)}\right\}+\frac{r_{yy}^2}{(r_{yy}^2)/(n-p)} \\
&= log\left\{\frac{|\Psi^T\Psi|}{|\Delta^T\Delta|}\right\}+(n-p)\left\{1+log\left\{\frac{2\pi r_{yy}^2}{(n-p)}\right\}\right\}
\end{aligned}
$$

2.3 Review of Nonlinear Mixed Effects (NLME) Models

As it has been explained by Pinheiro and Bates (1995) we use NLME models because of their interpretability, parsimony and validity beyond the observed range of the data. The most common application of NLME models is for repeated measures data and in particular, longitudinal data.
The NLME model for repeated measures proposed by Lindstrom and Bates (1990) can be thought of as a hierarchical model.
Pinheiro and Bates (1995) used a generalization of Lindstrom and Bates (1990) that in some ways generalizes both the linear mixed-effects model of Laird and Ware (1982) and the usual nonlinear model for the independent data, Bates and Watts (1988). According to Pinheiro and Bates (1995), at one level the j^{th} observation on the i^{th} group is modeled as

$$y_{ij} = f(\phi_{ij}, \vartheta_{ij}) + \epsilon_{ij}, i = 1, \cdots, M; j = 1, \cdots, n_i \tag{2.77}$$

where M is the number of groups, n_i is the number of observations on the i^{th} group and f is a general, real-valued, differentiable function of a group-specific parameter vector ϕ_{ij} and a covariate vector ϑ_{ij}, and ϵ_{ij} is a normally distributed within group error term.
The function f is nonlinear in at least one component of the group-specific parameter vector ϕ_{ij}, which is modeled as

$$\phi_{\mathbf{ij}} = A_{ij}\beta + B_{ij}b_i, b_i \sim N(0, \phi) \tag{2.78}$$

where β is p-dimensional vector of fixed effects, b_i is a q-dimensional random effects vector associated with i^{th} group (not varying with j) with variance-covariance matrix ϕ.
A_{ij} and B_{ij} are matrices of appropriate dimensions depending on the group and possibly on the values of some covariates at the j^{th} observation.
The model is a slight generalization of that described in Lindstrom and Bates (1990) since $A_{ij}\beta$ and $B_{ij}b_i$ can depend on j and allows the incorporation of "time-varying" covariates in the fixed effects or the random effects for the model.
Two very important assumptions taken into account for this model are

- observations corresponding to different groups are independent.
- within-group errors ϵ_{ij} are independently distributed as $N(0, \sigma^2)$ and are independent of the b_i.

First developments of NLME models appear in Sheiner and Beal (1980) whose estimation methods are widely used in pharmacokinetics. Their model is similar to (2.77). They developed a MLE method based on a first-order Taylor expansion of the model function around 0, the expected value of the random effects vector b.
Mallet, Mentre, Steimer, Lokiek (1988) proposed a nonparametric maximum likelihood method for nonlinear mixed-effects models. They made use of a model similar to (2.77) but made no assumptions about the distribution of the random effects. They assumed the conditional distribution of the response vector given the random effects. Their objective of the estimation procedure was to get probability distribution of the group-specific coefficients, ϕ_{ij}, that maximizes the likelihood of the data.

Mallet (1986) showed that the maximum likelihood solution is a discrete distribution with the number of discontinuity points less than or equal to the number of the groups in the sample.

Davidian and Gallant (1992) introduced a smooth, nonparametric MLE method for NLME. Their model was similar to (2.77), but with a more general definition for the group-specific coefficients

$$\phi_{ij} = g(\beta, b_i, \vartheta_{ij}) \tag{2.79}$$

where g is general, possibly nonlinear function.

Just as Mallet et al. (1988), they assume the conditional distribution of the response vector given the random effects is known (up to the parameters that define it) but the distribution of random effects is free to vary within a class of smooth densities defined by Gallant and Nychka (1987)

Bennett and Wakefield (1993) and Wakefield (1996) described a Bayesian approach using hierarchical models for nonlinear mixed-effects. First stage of their model was similar to (2.77). The distributions of random effects and errors ϵ_{ij} were assumed to be known upto population parameter. Prior distributions for the population parameters must be provided. To approximate the posterior density of the random effects, Markov-chain Monte Carlo methods were used, for example, Gibbs Sampler by Geman and Geman (1984) or the Metropolis algorithm by Hastings (1970).

Vonesh and Carter (1992) developed a mixed-effects model that is nonlinear in the fixed effects but linear in the random effects. Their model was

$$y_i = f(\beta, \vartheta_i) + Z_i(\beta) b_i + \epsilon_i \tag{2.80}$$

where as before

β is the fixed effects, b_i is the random effects and ϵ is the within-group error term.

ϑ is a matrix of covariates while Z_i is a full-rank matrix of known functions of the fixed effects β. As usual $b_i \sim N(0, \phi)$ and $\epsilon_i \sim N(0, \sigma^2 I)$ and the two vectors are independent of each other.

Vonesh and Carter (1992) in some way incorporated in their model the approximation suggested by Sheiner and Beal (1980) and Lindstrom and Bates (1990) though their approach concentrates more on inferences about the fixed effects and less on the variance-covariance components of the random effects.

2.3.1 Estimation in NLME Models

Different methods have been proposed to estimate parameters in the NLME model in (2.77). In our review we restrict ourselves on methods discussed by Pinheiro and Bates (1995) which are strictly based on the likelihood function. Pinheiro and Bates (1995) based their maximum likelihood estimate in (2.77) on the marginal density of $\mathbf{y}$

$$f(\mathbf{y}|\beta, \sigma^2, \phi_1, \cdots, \phi_Q) = \int f(\mathbf{y}|\mathbf{b}, \beta, \sigma^2)\ f(\mathbf{b})\ d\mathbf{b} \tag{2.81}$$

where $f(\mathbf{y}|\beta, \sigma^2, \phi_1, \cdots, \phi_Q)$ is the marginal density of $\mathbf{y}$, $f(\mathbf{y}|\mathbf{b}, \beta, \sigma^2)$ is the conditional density of $\mathbf{y}$ given the random effects $\mathbf{b}$ and $f(\mathbf{b})$ is the marginal distribution of $\mathbf{b}$.

Generally, the integral above does not have a closed-form expression when the model function f is nonlinear in $\mathbf{b}_i$, so different approximations have been proposed for estimating it.
We can therefore get the marginal density of responses $\mathbf{y}$ by first getting the conditional density of $\mathbf{y}$ given the random effects $\mathbf{b}$ as

$$f(\mathbf{y}|\mathbf{b}, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left\{ \frac{-\|y_i - f(\phi_i, \vartheta_i)\|^2}{2\sigma^2} \right\} \tag{2.82}$$

where

$$E(\mathbf{y}_i) = f(\phi_i, \vartheta_i)$$

and

$$var(\mathbf{y}_i) = \sigma^2 \mathbf{I}$$

implying that $\mathbf{y}_i \sim N(f(\phi_i, \vartheta_i), \sigma^2\mathbf{I})$.
Also the marginal distribution of $\mathbf{b}$ becomes

$$\begin{aligned} f(\mathbf{b}|\psi_1, \cdots, \psi_Q) &= \frac{1}{(2\pi)^{q/2}|\psi|^{1/2}} \exp\left\{ \frac{-(\mathbf{b}_i^T \psi^{-1} \mathbf{b}_i)}{2} \right\} \\ &= \frac{1}{(2\pi)^{q/2}\sigma^q abs|\Delta|^{-1}} \exp\left\{ \frac{-\|\Delta \mathbf{b}_i\|^2}{2\sigma^2} \right\} \end{aligned} \tag{2.83}$$

in which case

$$E(\mathbf{b}_i|\psi_1, \cdots, \psi_Q) = 0$$

and

$$var(\mathbf{b}_i|\psi_1, \cdots, \psi_Q)) = \psi_1, \cdots, \psi_Q$$

Thus distribution of $\mathbf{b}_i \sim N(0, \psi_1, \cdots, \psi_Q)$.
Again

$$\begin{aligned} \psi^{-1} &= \sigma^{-2}\Delta^T\Delta \\ &\Rightarrow |\psi| = \sigma^{2q}|(\Delta^T\Delta)^{-1}| \\ &\Rightarrow |\psi|^{1/2} = \sigma^q abs|\Delta|^{-1} \end{aligned}$$

Thus from (2.82) and (2.83), the marginal density of $\mathbf{y}$ becomes

$$\begin{aligned} f(\mathbf{y}|\beta, \sigma^2, \psi_1, \cdots, \psi_Q) &= \int f(\mathbf{y}|\mathbf{b}, \beta, \sigma^2)\ f(\mathbf{b})\ d\mathbf{b} \\ &= L(\beta, \sigma^2, \psi|\mathbf{y}) \\ &= \prod_{i=1}^{M} \int f(\mathbf{y}_i|\mathbf{b}_i, \beta, \sigma^2)\ f(\mathbf{b}_i|\psi_i)\ d\mathbf{b}_i \\ &= \prod_{i=1}^{M} \int \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left\{ \frac{-\|y_i - f_i(\phi_i, \vartheta_i)\|^2}{2\sigma^2} \right\} \\ &\quad \times \frac{1}{(2\pi)^{q/2}\sigma^q abs|\Delta|^{-1}} \exp\left\{ \frac{-\|\Delta \mathbf{b}_i\|^2}{2\sigma^2} \right\}\ d\mathbf{b}_i \end{aligned} \tag{2.84}$$

Now the likelihood function becomes

$$
\begin{aligned}
L(\beta, \sigma^2, \psi|\mathbf{y}) &= f(\mathbf{y}|\beta, \sigma^2, \Delta) \\
&= \prod_{i=1}^{M} \frac{1}{(2\pi\sigma^2)^{n_i/2}} \frac{1}{(2\pi)^{q/2}\sigma^q abs|\Delta|^{-1}} \int \exp\left\{\frac{-\|y_i - f_i(\phi_i, \vartheta_i)\|^2 + \|\Delta \mathbf{b}_i\|^2}{2\sigma^2}\right\} d\mathbf{b}_i \\
&= \prod_{i=1}^{M} \frac{1}{(2\pi\sigma^2)^{[n_i+q]/2} abs|\Delta|^{-1}} \int \exp\left\{\frac{-\|y_i - f_i(\phi_i, \vartheta_i)\|^2 + \|\Delta \mathbf{b}_i\|^2}{2\sigma^2}\right\} d\mathbf{b}_i \\
&= \frac{abs|\Delta|^M}{(2\pi\sigma^2)^{[N+Mq]/2}} \prod_{i=1}^{M} \int \exp\left\{\frac{-\|y_i - f_i(\phi_i, \vartheta_i)\|^2 + \|\Delta \mathbf{b}_i\|^2}{2\sigma^2}\right\} d\mathbf{b}_i \\
&= \frac{abs|\Delta|^M}{(2\pi\sigma^2)^{[N+Mq]/2}} \prod_{i=1}^{M} \int \exp\left\{\frac{-\|y_i - f_i(\beta, \mathbf{b}_i)\|^2 + \|\Delta \mathbf{b}_i\|^2}{2\sigma^2}\right\} d\mathbf{b}_i
\end{aligned}
\tag{2.85}
$$

where

$$
\begin{aligned}
f_i(\beta, \mathbf{b}_i) &= f_i(\phi_i, \vartheta_i) \\
&= f_i[\phi_i(\beta, \mathbf{b}_i), \vartheta_i]
\end{aligned}
$$

Linear Mixed Effects Approximation

Lindstrom and Bates (1990) Algorithm alternates between two steps

- a penalized nonlinear least squares (PNLS) step, and
- a linear mixed effects (LME) step

In short their estimation approximation moves from the single-level linear mixed effects model given in (2.1) to the single-level nonlinear mixed effect model in (2.77), out of which we get the penalized nonlinear least squares objective function as

$$
\sum_{i=1}^{M} [\|y_i - f_i(\beta, \mathbf{b}_i)\|^2 + \|\Delta \mathbf{b}_i\|^2] \tag{2.86}
$$

From (2.1), if $\sigma^2\mathbf{I}$ and $\psi = \sigma^2(\Delta^T\Delta)^{-1}$ are known, and ϕ, a linear function of β and $\mathbf{b}(\phi(A_i\beta + B\mathbf{b}_i) = X_i\beta + Z_i\mathbf{b}_i)$, then with the usual definition of X and Z, the standard estimators for β and $\mathbf{b}$ are the generalized least squares estimators

$$
\hat{\beta}_{lin}(\theta) = (X^T(\Sigma(\Delta))^{-1}X)^{-1}X^T(\Sigma(\Delta))^{-1}y
$$

where

$$
\Sigma(\Delta) = \mathbf{I} + \bar{Z}(\Delta^T\Delta)^{-1}\bar{Z}^T \tag{2.87}
$$

and

$$
\hat{\mathbf{b}}_{lin} = \hat{\mathbf{b}}_{lin}(\theta) = (\Delta^T\Delta)^{-1}\bar{Z}^T[\Sigma(\Delta)]^{-1}(y - X\hat{\beta}_{lin}(\theta))
$$

Note that θ contains the unique elements of Δ.
Estimates $\hat{\beta}_{lin}$ and $\hat{\mathbf{b}}_{lin}$, jointly maximize the function

$$\begin{aligned} g_{lin}(\beta, \mathbf{b}|y) &= -\frac{1}{2}\sigma^{-2}(y - X\beta - Z\mathbf{b})^2 - \frac{1}{2}\sigma^{-2}\mathbf{b}^T\Delta^T\Delta\mathbf{b} \\ &= -\frac{1}{2}\sigma^{-2}[\|y - X\beta - Z\mathbf{b}\|^2 + \|\Delta\mathbf{b}\|^2] \end{aligned}$$

which

- for fixed β, is the logarithm of the posterior density of $\mathbf{b}$ (up to a constant)
- for fixed $\mathbf{b}$, is the log-likelihood for β (up to a constant)

The two terms in the above equation are **a sum of squares** and **a quadratic term** in $\mathbf{b}$. By transforming the quadratic term in $\mathbf{b}$ to **an equivalent sum of squares term**, we can **treat the optimization purely as a least squares problem**, which is **easy to translate into nonlinear setting.**

Lemma 6. *The Linear Mixed Effects estimation problem for obtaining the estimates in NLME models can be defined as*

$$\begin{aligned} \bar{l}(\beta, \sigma, \Delta|y) &= \bar{l}(\beta, \sigma, \theta|\theta^{(w)}, y) \\ &= -\frac{1}{2}log|\sigma^2(\mathbf{I} + \mathbf{Z}^{(w)}\Delta^{-1}\Delta^{-1^T}\mathbf{Z}^{w^T})| \\ &\quad - \frac{1}{2}\sigma^{-2}[\mathbf{w}^{(w)} - \mathbf{X}^{(w)}\beta]^T(\mathbf{I} + \mathbf{Z}^{(w)}\Delta^{-1}\Delta^{-1^T}\mathbf{Z}^{w^T})^{-1}[\mathbf{w}^{(w)} - \mathbf{X}^{(w)}\beta] \end{aligned}$$

where the desired estimates are $\beta^{(ML)}$, $\sigma^{(ML)}$ and $\theta^{(ML)}$ which maximize $\bar{l}$.

Proof. The least squares problem is created by augmenting the **data vector** with **"pseudo-data"** as

$$\bar{y}_i = \bar{X}_i\beta + \bar{Z}_i\mathbf{b}_i + \bar{e}_i$$

where

$$\bar{X}_i = \begin{bmatrix} X_i \\ 0 \end{bmatrix}, \bar{y}_i = \begin{bmatrix} y_i \\ 0 \end{bmatrix}, \bar{Z}_i = \begin{bmatrix} Z_i \\ \Delta \end{bmatrix}$$

and

$$\bar{e}_i \sim N(0, \sigma^2\mathbf{I})$$

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

$$\begin{aligned} g_{lin}(\beta, \mathbf{b}|y_i) &= -\frac{1}{2}(y_i - f_i(\beta, \mathbf{b}_i))^T(y_i - f_i\beta, \mathbf{b}_i) - \frac{1}{2}\sigma^{-2}\mathbf{b}^T\Delta^T\Delta\mathbf{b} \\ &= -\frac{1}{2}\sigma^{-2}[\|y_i - f_i\beta, \mathbf{b}_i\|^2 + \|\Delta\mathbf{b}\|^2] \end{aligned} \quad (2.88)$$

in which case, for fixed β, g is a constant plus the log of the posterior density of $\mathbf{b}$.

Thus it is clear that the $\mathbf{b}$ that maximizes g for a given value of β (fixed β) is the **posterior mode.** $\hat{\beta}$ is a MLE relative to an approximate marginal density of y.

Just as in the linear case, the estimates $\hat{\beta}$ and $\mathbf{b}$ could be calculated as a solution to a nonlinear least squares problem formed by augmenting the data vector with **"pseudo-data"** as

$$\bar{y} = \bar{f}(\beta, \mathbf{b}) + \hat{\epsilon} \tag{2.89}$$

where

$$\bar{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}, \bar{f}(\beta, \mathbf{b}) = \begin{bmatrix} f(\beta, \mathbf{b}) \\ \Delta \mathbf{b} \end{bmatrix}$$

and

$$\bar{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$$

We wish to define the maximum likelihood estimators for θ with respect to the marginal density of y, that is,

$$f(\mathbf{y}|\beta, \sigma^2, \psi_1, \cdots, \psi_Q) = \int f(\mathbf{y}|\mathbf{b}, \beta, \sigma^2)\ f(\mathbf{b})\ d\mathbf{b} \tag{2.90}$$

The expectation function $f(\beta, \mathbf{b})$ is nonlinear in $\mathbf{b}$. Thus **there is no closed form expression for this density** and **calculation of such estimates is very difficult.**

Hence, instead, we approximated the conditional distributions of y for $\mathbf{b}$ near $\hat{\mathbf{b}}(\theta)$, that is,

$$y|\mathbf{b} \sim N(f(\beta, \mathbf{b}), \sigma^2 \mathbf{I})$$

where

$$f(\beta, \mathbf{b}) = f[\phi(A\beta + B\mathbf{b}), \vartheta]$$

with a multivariate normal with expectation that is linear in $\mathbf{b}$.

To accomplish this we approximated the residual $y - f(\beta, \mathbf{b})$ near $\hat{\mathbf{b}}$ as

$$y - f(\beta, \mathbf{b}) \approx y - [f(\beta, \hat{\mathbf{b}}) + \hat{Z}\mathbf{b} - \hat{Z}\hat{\mathbf{b}}]$$

where

$$\begin{aligned}
\hat{Z}_i &= \hat{Z}_i(\theta) \\
&= \frac{\partial f_i}{\partial \mathbf{b}_i^T}|_{\hat{\beta},\hat{\mathbf{b}}} \\
&= (\frac{\partial f_i}{\partial \phi_i^T}|_{\hat{\beta},\hat{\mathbf{b}}})B_i \\
&= f_i'(A\beta + B\mathbf{b}_i)B_i
\end{aligned}$$

and thus

$$\begin{aligned}
\hat{Z} &= \hat{Z}(\theta) \\
&= diag(\hat{Z}_1, \cdots, \hat{Z}_M) \\
&= \frac{\partial f}{\partial \mathbf{b}^T}|_{\hat{\beta},\hat{\mathbf{b}}} \\
&= f'(A\beta + B\mathbf{b})B
\end{aligned}$$

Note that $\hat{Z}$ is a function of θ, the reason being that $\hat{\beta}$ and $\hat{\mathbf{b}}$ are functions of θ.
Then we have that

$$y - f(\beta, \hat{\mathbf{b}}) + \hat{Z}\hat{\mathbf{b}} - \hat{Z}\mathbf{b}|\mathbf{b} \sim N(0, \sigma^2\mathbf{I})$$

Hence the approximate conditional distribution of y becomes

$$y|\mathbf{b} \sim N(f(\beta, \hat{\mathbf{b}}) - \hat{Z}\hat{\mathbf{b}} + \hat{Z}\mathbf{b}, \sigma^2\mathbf{I})$$

and the distribution of $\mathbf{b}_i$ is given as

$$\mathbf{b}_i \sim N(0, \sigma^2[\Delta^T\Delta]^{-1})$$

Thus the two expressions above allows us to approximate the marginal distribution of y as

$$y_i \sim N(f(\beta, \hat{\mathbf{b}}) - \hat{Z}\hat{\mathbf{b}}, \sigma^2\Sigma_i(\Delta))$$

where $\Sigma(\Delta)$ is as defined in (2.87)
This kind of approximation has been used in a similar setting by Stiratelli et al. (1984). Algorithm of Sheiner and Beal (1980) use similar approximation but evaluated at the expectation of the random effects (at $\mathbf{b} = 0$ in this model) rather than at the current estimates. Simplication reduces the computational burden if a Newton-Raphson type algorithm is used but may result in poor estimates.

More accurate method of approximating the marginal distribution of y at the current estimates of the random effects as discussed by Wolf in a proposed application of the Estimation Maximization (EM) algorithm to maximum likelihood (ML) estimation for the nonlinear random effects model.

The log-likelihood corresponding to the approximate marginal distribution above is

$$l(\beta, \sigma, \Delta|y) = -\frac{1}{2}log|\sigma^2\Sigma_i(\Delta)| - \frac{1}{2}\sigma^{-2}[y - f_i(\beta, \hat{\mathbf{b}}) + \hat{Z}\hat{\mathbf{b}}]^T(\Sigma_i(\Delta))^{-1}[y - f_i(\beta, \hat{\mathbf{b}}) + \hat{Z}\hat{\mathbf{b}}] \tag{2.91}$$

where $\mathbf{b}$ and $\hat{Z}$ depend on θ.

Define $\beta^{(ML)}$, $\sigma^{(ML)}$ and $\theta^{(ML)}$ to be maximum likelihood estimators for β, σ and θ with respect to l, (2.91.) As in the linear case, the two estimators for β are equivalent, that is, $\beta^{(ML)} = \hat{\beta}(\theta^{(ML)})$, from the fact that, they both maximize $l(\beta, \sigma^{(ML)}, \theta^{(ML)}|y)$.

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

Method of defining **restricted maximum likelihood (REML)** estimators is the same as that for the MLE except that l above becomes

$$l_R(\beta, \sigma, \Delta|y) = -\frac{1}{2}log|\sigma^{-2}\hat{X}^T(\Sigma_i(\Delta))^{-1}\hat{X}| + l(\beta, \sigma, \Delta|y) \tag{2.92}$$

where

$$\begin{aligned}\hat{X}_i &= \hat{X}_i(\theta)\\ &= \frac{\partial f_i}{\partial \beta^T}|_{\hat{\beta},\hat{\mathbf{b}}}\\ &= (\frac{\partial f_i}{\partial \phi_i^T}|_{\hat{\beta},\hat{\mathbf{b}}})A_i\\ &= f_i'(A\beta + B\mathbf{b}_i)A_i\end{aligned}$$

and thus

$$\begin{aligned}\hat{X} &= \hat{X}(\theta)\\ &= \begin{bmatrix}\hat{X}_1\\ \vdots\\ \hat{X}_M\end{bmatrix}\\ &= \frac{\partial f}{\partial \beta^T}|_{\hat{\beta},\hat{\mathbf{b}}}\\ &= f'(A\beta + B\mathbf{b})A\end{aligned}$$

We define estimators $\beta^{(RML)}$, $\sigma^{(RML)}$ and $\theta^{(RML)}$ as those that maximize l_R.

REML is based on $N - p$ linearly independent error contrasts.

In the nonlinear model, the derivative matrix $\hat{X}$, which defines these errors contrast depends on $\hat{\beta}$ and $\hat{\mathbf{b}}$.

However, it has been noted that since the subspace spanned by the columns of this matrix depends only on intrinsic nonlinearity and not a parameter effects nonlinearity (Bates and Watts, (1980)), it will be nearly constant near the estimates.

The two step algorithm proposed for finding $\theta^{(ML)}$ by Lindstrom and Bates (1990) are

1. Pseudo-data (PD) step

   Given the current estimate $\theta^{(w)}$ for θ, then $\beta^{(w)} = \hat{\beta}(\theta^{(w)}), \mathbf{b}^w = \hat{\mathbf{b}}(\theta^(w)), \mathbf{X}^w = \hat{\mathbf{X}}(\theta^{(w)}), \mathbf{Z}^w = \hat{\mathbf{Z}}(\theta^{(w)})$ can be estimated.

2. Linear mixed effects (LME) step

   Given $\mathbf{b}^w$ and $\mathbf{Z}^w$, let $\beta^{(w+1)}$, $\sigma^{(W+1)}$ and $\theta^{(w+1)}$ be the values that maximize

$$\begin{aligned} l(\beta,\sigma,\Delta|y) &= l(\beta,\sigma,\theta|\theta^{(w)},y) \\ &= -\frac{1}{2}log|\sigma^2\Sigma_i(\Delta)| - \frac{1}{2}\sigma^{-2}[y - f_i(\beta,\hat{\mathbf{b}}) + \hat{Z}\hat{\mathbf{b}}]^T(\Sigma_i(\Delta))^{-1}[y - f_i(\beta,\hat{\mathbf{b}}) + \hat{Z}\hat{\mathbf{b}}] \\ &= -\frac{1}{2}log|\sigma^2(\mathbf{I} + \mathbf{Z}^{(w)}\Delta^{-1}\Delta^{-1^T}\mathbf{Z}^{w^T})| - \frac{1}{2}\sigma^{-2}[y - f_i(\beta,\mathbf{b}^{(w)}) + \mathbf{Z}^{(w)}\mathbf{b}^{(w)}]^T \\ &\quad \times (\mathbf{I} + \mathbf{Z}^{(w)}\Delta^{-1}\Delta^{-1^T}\mathbf{Z}^{w^T})^{-1}[y - f_i(\beta,\mathbf{b}^{(w)}) + \mathbf{Z}^{(w)}\mathbf{b}^{(w)}] \end{aligned} \tag{2.93}$$

   We write the dependence of l on $\theta^{(w)}$ explicitly to emphasize the dependence on the value of θ at which $\hat{\mathbf{b}}$ and $\hat{Z}$ are evaluated.

   The algorithm consists of iterating between these two steps until convergence.

The LME above does not quite respond to a LME estimation problem. This is because the residual

$$y - f_i(\beta,\mathbf{b}^{(w)}) + \mathbf{Z}^{(w)}\mathbf{b}^{(w)} \tag{2.94}$$

is nonlinear in β.

This would require second derivatives of the model function, f with respect to fixed effects, β.

To avoid this, we approximate the residual near $\beta^{(w)}$ by:

$$\begin{aligned} y - f_i(\beta, \mathbf{b}^{(w)}) + \mathbf{Z}^{(w)}\mathbf{b}^{(w)} &\approx y - [f_i(\beta, \mathbf{b}^{(w)}) + \mathbf{X}^{(w)}(\beta - \beta^{(w)}) - \mathbf{Z}^{(w)}\mathbf{b}^{(w)}] \\ &= y - [f_i(\beta, \mathbf{b}^{(w)}) + \mathbf{X}^{(w)}\beta^{(w)} + \mathbf{Z}^{(w)}\mathbf{b}^{(w)} - \mathbf{X}^{(w)}\beta] \\ &= \mathbf{w}^{(w)} - \mathbf{X}^{(w)}\beta \end{aligned} \tag{2.95}$$

where

$$\mathbf{X}^{(w)} = \hat{\mathbf{X}}(\theta^{(w)})$$

$$\mathbf{w}^{(w)} = y - [f_i(\beta, \mathbf{b}^{(w)}) + \mathbf{X}^{(w)}\beta^{(w)} + \mathbf{Z}^{(w)}\mathbf{b}^{(w)}$$

Thus, we can define

$$\begin{aligned} \bar{l}(\beta, \sigma, \Delta|y) = \bar{l}(\beta, \sigma, \theta|\theta^{(w)}, y) \\ &= -\frac{1}{2} log|\sigma^2(\mathbf{I} + \mathbf{Z}^{(w)}\Delta^{-1}\Delta^{-1^T}\mathbf{Z}^{w^T})| \\ &\quad - \frac{1}{2}\sigma^{-2}[\mathbf{w}^{(w)} - \mathbf{X}^{(w)}\beta]^T(\mathbf{I} + \mathbf{Z}^{(w)}\Delta^{-1}\Delta^{-1^T}\mathbf{Z}^{w^T})^{-1}[\mathbf{w}^{(w)} - \mathbf{X}^{(w)}\beta] \end{aligned} \tag{2.96}$$

□

Then the LME step with $\bar{l}$ is a **linear mixed effects estimation problem** of the type discussed in Laird and Ware (1982).
Lindstrom and Bates (1990) found that this new LME step will result in the **desired estimates** since $\beta^{(ML)}$, $\sigma^{(ML)}$ and $\theta^{(ML)}$ maximize $\bar{l}$.
Method for obtaining the REML estimates is exactly the same as that for the MLE except $\bar{l}$ is replaced by

$$\bar{l}_R(\beta, \sigma, \theta|\theta^{(w)}, y) = -\frac{1}{2} log|\sigma^{-2}\mathbf{X}^{(w)^T}(\mathbf{I} + \mathbf{Z}^{(w)}\Delta^{-1}\Delta^{-1^T}\mathbf{Z}^{w^T})^{-1}\mathbf{X}^{(w)}| + \bar{l}(\beta, \sigma, \theta|\theta^{(w)}, y) \tag{2.97}$$

At this point it is worth noting that $\mathbf{X}^{(w)}$ depends on both $\hat{\beta}^{(w)}$ and $\hat{\mathbf{b}}^{(w)}$).
Therefore changes in the fixed effects model or random effects model imply changes in the penalty factor for log-restricted likelihood. Thus the log-restricted likelihoods from NLME models with different fixed or random effects models are not comparable.

Laplacian Approximation

Has a number of uses,

- used frequently in Bayesian inference to estimate marginal posterior densities and predictive distributions as in Tierney and Kadane (1986) and Leonard, Hsu and Tsui (1989) papers.

- used for approximating the likelihood function in NLME models.

Considering the single-level NLME model (2.77), to obtain the marginal distribution of $\mathbf{y}$ in (2.85), Pinheiro and Bates (1995) estimated the integral

$$f(\mathbf{y}_i|\beta, \sigma^2, \Delta) = \int (2\pi\sigma^2)^{-(n_i+q)/2}|\Delta| \exp[\frac{-g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i)}{2\sigma^2}] \, d\mathbf{b}_i \tag{2.98}$$

where

$$g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i) = \|y_i - f_i(\beta, \mathbf{b}_i)\|^2 + \|\Delta \mathbf{b}_i\|^2 \tag{2.99}$$

Hence

$$\sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i) = \sum_{i=1}^{M} [\|y_i - f_i(\beta, \mathbf{b}_i)\|^2 + \|\Delta \mathbf{b}_i\|^2]$$

is the objective function defined for the PNLS step of the alternating algorithm defined in (2.86).
Letting

$$\begin{aligned} \hat{\mathbf{b}}_i &= \hat{\mathbf{b}}_i(\beta, \Delta, \mathbf{y}_i) \\ &= \underset{bi}{\operatorname{argmin}}\, g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i) \\ &\Rightarrow g'(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i) = \frac{\partial g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i} \\ &\Rightarrow g''(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i) = \frac{\partial^2 g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T} \end{aligned} \tag{2.100}$$

Considering a Second Order Taylor expansion of g around $\hat{\mathbf{b}}_i$, we get

$$\begin{aligned} g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i) &= g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) + (\mathbf{b}_i - \hat{\mathbf{b}}_i) g'(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) + \frac{1}{2}(\mathbf{b}_i - \hat{\mathbf{b}}_i)^T g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)(\mathbf{b}_i - \hat{\mathbf{b}}_i) \\ &\simeq g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) + \frac{1}{2}(\mathbf{b}_i - \hat{\mathbf{b}}_i)^T g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)(\mathbf{b}_i - \hat{\mathbf{b}}_i) \end{aligned} \tag{2.101}$$

Linear term in the above expansion vanishes since $g'(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) = 0$

Lemma 7. *The modified Laplacian approximation of the single-level NLME model (2.77) is given as*

$$L(\beta, \sigma^2, \Delta|\mathbf{y}) \simeq \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)] \prod_{i=1}^{M} |G(\beta, \Delta, \mathbf{y}_i)|^{-1/2}$$

where

$$G(\beta, \Delta, \mathbf{y}_i) \simeq g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)$$

Proof. From the likelihood function in (2.85) we have

$$
\begin{aligned}
L(\beta, \sigma^2, \Delta|\mathbf{y}) &= f(\mathbf{y}|\beta, \sigma^2, \Delta) \\
&= \prod_{i=1}^{M} \frac{1}{(2\pi\sigma^2)^{n_i/2}} \frac{1}{(2\pi)^{q/2}\sigma^q abs|\Delta|^{-1}} \int \exp\left\{ \frac{-\|y_i - f_i(\phi_i, \vartheta_i)\|^2 + \|\Delta \mathbf{b}_i\|^2}{2\sigma^2} \right\} d\mathbf{b}_i \\
&= \prod_{i=1}^{M} \frac{|\Delta|}{(2\pi\sigma^2)^{n_i/2}} \int \exp\left\{ \frac{-\|y_i - f_i(\phi_i, \vartheta_i)\|^2 + \|\Delta \mathbf{b}_i\|^2 / 2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} d\mathbf{b}_i
\end{aligned}
\tag{2.102}
$$

From (2.99) and (2.102) and also incorporating the results of (2.101) we have

$$
\begin{aligned}
f(\mathbf{y}|\beta, \sigma^2, \Delta) &= \prod_{i=1}^{M} \frac{|\Delta|}{(2\pi\sigma^2)^{n_i/2}} \int \exp\left\{ \frac{-g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i)/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} d\mathbf{b}_i \\
&\simeq \prod_{i=1}^{M} \frac{|\Delta|}{(2\pi\sigma^2)^{n_i/2}} \int \exp\left\{ \frac{[g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) + \frac{1}{2}(\mathbf{b}_i - \hat{\mathbf{b}}_i)^T g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)(\mathbf{b}_i - \hat{\mathbf{b}}_i)]/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} d\mathbf{b}_i \\
&= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)] \prod_{i=1}^{M} \\
&\quad \times \int \exp\left\{ \frac{[-\frac{1}{2\sigma^2}(\mathbf{b}_i - \hat{\mathbf{b}}_i)^T g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)(\mathbf{b}_i - \hat{\mathbf{b}}_i)]}{(2\pi\sigma^2)^{q/2}} \right\} d\mathbf{b}_i \\
&= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)] \prod_{i=1}^{M} \\
&\quad \times \int \frac{|g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)|^{-1/2}}{(2\pi\sigma^2)^{q/2}|g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)|^{-1/2}} \exp -\frac{1}{2\sigma^2}(\mathbf{b}_i - \hat{\mathbf{b}}_i)^T g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)(\mathbf{b}_i - \hat{\mathbf{b}}_i)\ d\mathbf{b}_i \\
&= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)] \prod_{i=1}^{M} |g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)|^{-1/2}
\end{aligned}
\tag{2.103}
$$

Pinheiro and Bates (1995) considered an approximation to $g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)$ similar to the one used in Gauss-Newton optimization, that is, the Hessian

$$
\begin{aligned}
g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) &= \frac{\partial^2 g(\beta, \Delta, \mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} \\
&= \frac{\partial^2}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T} \left\{ \|y_i - f_i(\beta, \mathbf{b}_i)\|^2 + \|\Delta \mathbf{b}_i\|^2 \right\} |_{\mathbf{b}_i=\hat{\mathbf{b}}_i} \\
&= \frac{\partial^2}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T} \left\{ (y_i - f_i(\beta, \mathbf{b}_i))^T (y_i - f_i(\beta, \mathbf{b}_i)) + \mathbf{b}_i^T \Delta^T \Delta \mathbf{b}_i \right\} |_{\mathbf{b}_i=\hat{\mathbf{b}}_i} \\
&= \frac{\partial}{\partial \mathbf{b}_i^T} \left\{ -f_i'^T|_{\hat{\mathbf{b}}_i} (y_i - f_i(\beta, \mathbf{b}_i)) - f_i'|_{\hat{\mathbf{b}}_i} [y_i - f_i(\beta, \mathbf{b}_i)]^T + 2\mathbf{b}_i^T \Delta^T \Delta \right\} \\
&= -f_i''^T|_{\hat{\mathbf{b}}_i} [y_i - f_i(\beta, \mathbf{b}_i)] + f_i'^T|_{\hat{\mathbf{b}}_i} f_i'|_{\hat{\mathbf{b}}_i} - f_i''|_{\hat{\mathbf{b}}_i} [y_i - f_i(\beta, \mathbf{b}_i)]^T + f_i'|_{\hat{\mathbf{b}}_i} f_i'^T|_{\hat{\mathbf{b}}_i} + 2\Delta^T \Delta \\
&= -2f_i''|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} [y_i - f_i(\beta, \mathbf{b}_i)] + 2f_i'|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} f_i'|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} + 2\Delta^T \Delta
\end{aligned}
\tag{2.104}
$$

where

$$
f_i'|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} = \frac{\partial f_i(\beta, \mathbf{b}_i)}{\partial \mathbf{b}_i}|_{\mathbf{b}_i=\hat{\mathbf{b}}_i}
$$

and

$$
f_i''|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} = \frac{\partial^2 f_i(\beta, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}|_{\mathbf{b}_i=\hat{\mathbf{b}}_i}
$$

Since $g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) = 0$

then

$$
g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) = -f_i''(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} [y_i - f_i(\beta, \mathbf{b}_i)] + f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} + \Delta^T \Delta
\tag{2.105}
$$

According to Bates and Watts (1980), the Hessian above involves Second derivatives of f but, at $\hat{b}_i$ the contribution

$$
f_i''(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} [y_i - f_i(\beta, \mathbf{b}_i)]
$$

is usually negligible compared to that of

$$
f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i}
$$

Therefore the approximation becomes

$$
\begin{aligned}
g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) &\simeq G(\beta, \Delta, \mathbf{y}_i) \\
&= f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} + \Delta^T \Delta \qquad (2.106) \\
&= \frac{\partial f_i(\beta, \mathbf{b}_i)}{\partial \mathbf{b}_i}|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} \frac{\partial f_i(\beta, \mathbf{b}_i)}{\partial \mathbf{b}_i^T}|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} + \Delta^T \Delta
\end{aligned}
$$

Pinheiro and Bates (1995) found their approximation to be similar to the one used in the Gauss-Newton algorithm for nonlinear least squares with an advantage of requiring only the first-order partial derivatives of f with respect to the random effects. These first-order partial derivatives are usually available as a by-product of the estimation of $\hat{\mathbf{b}}_i$, which is a penalized least squares problem.

Combining the likelihood (2.103) with the approximation (2.106), we have

$$\begin{aligned} L(\beta, \sigma^2, \Delta|\mathbf{y}) &= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)] \prod_{i=1}^{M} |g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)|^{-1/2} \\ &\simeq \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)] \prod_{i=1}^{M} |G(\beta, \Delta, \mathbf{y}_i)|^{-1/2} \end{aligned} \tag{2.107}$$

□

Thus the modified Laplacian approximation to the log-likelihood of the single-level NLME model (2.77) becomes

$$\begin{aligned} logL(\beta, \sigma^2, \Delta|\mathbf{y}) &= l_{LA}(\beta, \sigma^2, \Delta|\mathbf{y}) \\ &= \frac{-N}{2} log(2\pi\sigma^2) + Mlog|\Delta| - \frac{1}{2}\left\{\sum_{i=1}^{M} log|G(\beta, \Delta, \mathbf{y}_i)| + \sigma^{-2}\sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)\right\} \end{aligned} \tag{2.108}$$

Since $\hat{\mathbf{b}}_i$ does not depend on σ^2, for given β and Δ the maximum likelihood estimate of σ^2 based on the Laplacian approximation to the log-likelihood (l_{LA}) in (2.108) is

$$\begin{aligned} \frac{\partial(l_{LA}(\beta, \sigma^2, \Delta|\mathbf{y}))}{\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}\left\{\frac{-N}{2} log(2\pi\sigma^2) + Mlog|\Delta| - \Theta\right\} = 0 \\ &\Rightarrow \frac{-N}{2\sigma^2} + \frac{\sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)}{2\sigma^4} = 0 \\ &\Rightarrow \hat{\sigma}^2 = \hat{\sigma}^2(\beta, \Delta, \mathbf{y}) = \frac{\sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)}{N} \end{aligned} \tag{2.109}$$

after letting

$$\Theta = \frac{1}{2}\left\{\sum_{i=1}^{M} log|G(\beta, \Delta, \mathbf{y}_i)| + \sigma^{-2}\sum_{i=1}^{M} g(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i)\right\} \tag{2.110}$$

Pinheiro and Bates (1995) obtained the profiled Laplacian approximation to the log-likelihood (l_{LAp}) on σ^2 to reduce the dimension of the optimization problem, as

$$\begin{aligned} l_{LAp}(\beta,\Delta) &= \frac{-N}{2}log(2\pi\hat{\sigma}^2) + Mlog|\Delta| - \frac{1}{2}\left\{\sum_{i=1}^{M} log|G(\beta,\Delta,\mathbf{y}_i)| + \frac{\sum_{i=1}^{M} g(\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i)}{\hat{\sigma}^2}\right\} \\ &= \frac{-N}{2}log(2\pi\hat{\sigma}^2) + Mlog|\Delta| - \frac{1}{2}\left\{\sum_{i=1}^{M} log|G(\beta,\Delta,\mathbf{y}_i)| + \frac{N\sum_{i=1}^{M} g(\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i)}{\hat{\sigma}^2 \times N}\right\} \\ &= \frac{-N}{2}log(2\pi\hat{\sigma}^2) + Mlog|\Delta| - \frac{1}{2}\left\{\sum_{i=1}^{M} log|G(\beta,\Delta,\mathbf{y}_i)| + N\right\} \\ &= \frac{-N}{2}\left\{1 + log(2\pi) + log(\hat{\sigma}^2)\right\} + Mlog|\Delta| - \frac{1}{2}\sum_{i=1}^{M} log|G(\beta,\Delta,\mathbf{y}_i)| \end{aligned} \tag{2.111}$$

Pinheiro and Bates (1995) noted that if the model function f is linear in the random effects, then the modified Laplacian approximation is exact because the Second-order Taylor expansion in (2.101) is exact when

$$f_i(\beta,\mathbf{b}_i) = f_i(\beta) + Z_i(\beta)\mathbf{b}_i \tag{2.112}$$

They also noted that, there does not seem to be a straightforward generalization of the concept of restricted maximum likelihood (REML) to NLME models since REML depends heavily upon the linearity of the fixed effects in the model function, which does not occur in nonlinear models.
Lindstrom and Bates (1990) overcame this problem by using an approximation to the model function f in which the fixed effects,β, occur linearly.

Importance Sampling

Importance sampling provides simple and efficient way of performing **Monte Carlo integration** as discussed by Geweke (1989).
Critical step for the success of importance sampling is the choice of an **importance distribution** from which the sample is drawn and the **importance weights calculated**.
Ideally this distribution corresponds to the density that we are trying to integrate though in practice one uses an easily sampled approximation.
We saw earlier that from Pinheiro and Bates (1995), the integral that they estimated to obtain the marginal distribution of $\mathbf{y}_i$ in (2.85) could be written as (2.98). From our nonlinear mixed-effects model the function that we want to integrate is, upto a multiplicative constant equal to

$$\exp[-g(\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i)/2\sigma^2] \tag{2.113}$$

as can be observed from (2.98).
As shown in (2.101), by taking a Second-order Taylor expansion of $g(\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i)$ around $\hat{\mathbf{b}}_i$ the integrand is, up to a multiplicative constant, approximately $N(\hat{\mathbf{b}}_i, \sigma^2[G(\beta,\Delta,\mathbf{y}_i)]^{-1})$ density, which gives us a natural choice for the **importance distribution**.

Lemma 8. *The importance sampling approximation to the log likelihood of* $\mathbf{y}$ *from the above distribution is*

$$\begin{aligned} logL_{IS}(\beta,\sigma^2,\Delta|\mathbf{y}) &= l_{IS}(\beta,\sigma^2,\Delta|\mathbf{y}) \\ &= \frac{-N}{2}log(2\pi\sigma^2) + Mlog|\Delta| - \frac{1}{2}\sum_{i=1}^{M} log|G(\beta,\Delta,\mathbf{y}_i)| \\ &+ \sum_{i=1}^{M} log\left\{\sum_{j=1}^{N_{IS}} \exp(-g[\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i^*]/2\sigma^2 + \left\|\mathbf{z}_\mathbf{j}^*\right\|^2/2)\right\}/N_{IS} \end{aligned} \tag{2.114}$$

Proof. Letting N_{IS} denote the number of importance samples to be drawn we can calculate the sample of random effects. An example of such importance sample as given by Pinheiro and Bates (1995) can be generated by selecting a vector $\mathbf{z}^* \sim N(0,1)$, that is, with distribution $N(0,1)$.

From the distribution of the integrand, we can obtain the distribution of $\mathbf{z}^*$ as

$$\mathbf{z}^* = \frac{\mathbf{b}_i^* - \hat{\mathbf{b}}_i}{(\sigma^2[G(\beta,\Delta,\mathbf{y}_i)]^{-1})^{1/2}} \sim N(0,1) \tag{2.115}$$

Hence from (2.115) we can obtain the sample of random effects as

$$\mathbf{b}_i^* = \hat{\mathbf{b}}_i + \sigma[G(\beta,\Delta,\mathbf{y}_i)]^{-1/2}\mathbf{z}^* \tag{2.116}$$

where $[G(\beta,\Delta,\mathbf{y}_i)]^{-1/2}$ denotes the inverse of the Cholesky factor of $G(\beta,\Delta,\mathbf{y}_i)$

Thus the importance sampling approximation to the distribution becomes

$$\begin{aligned} f(\mathbf{y}|\beta,\sigma^2,\Delta) &= L_{IS}(\beta,\sigma^2,\Delta|\mathbf{y}) \\ &= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}}\prod_{i=1}^{M}\int \sigma^q|[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2}\frac{[\exp -g[\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i^*]/2\sigma^2 \times \exp(-\left\|\mathbf{z}^*\right\|^2/2)}{(2\pi\sigma^2)^{q/2}}d\mathbf{z} \\ &= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}}\prod_{i=1}^{M}|[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2}\int \frac{[\exp -g[\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i^*]/2\sigma^2 \times \exp(-\left\|\mathbf{z}\right\|^2/2)}{(2\pi)^{q/2}}d\mathbf{z}^* \\ &= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}}\prod_{i=1}^{M}|[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2} \\ &\sum_{j1=1}^{N_{IS}},\cdots,\sum_{jq=1}^{N_{IS}} \exp\left\{-g[\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i^*]/2\sigma^2 + \left\|\mathbf{z}_\mathbf{j}^*\right\|^2/2\right\}/N_{IS} \end{aligned} \tag{2.117}$$

Hence the importance sampling approximation to the log-likelihood of $\mathbf{y}$ from the distribution above becomes

$$
\begin{aligned}
logL_{IS}(\beta,\sigma^2,\Delta|\mathbf{y}) &= l_{IS}(\beta,\sigma^2,\Delta|\mathbf{y}) \\
&= \frac{-N}{2}log(2\pi\sigma^2) + Mlog|\Delta| - \frac{1}{2}\sum_{i=1}^{M} log|G(\beta,\Delta,\mathbf{y}_i)| \\
&\quad + \sum_{i=1}^{M} log\left\{\sum_{j=1}^{N_{IS}} \exp(-g[\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i^*]/2\sigma^2 + \left\|\mathbf{z_j^*}\right\|^2/2)\right\}/N_{IS}
\end{aligned}
\tag{2.118}
$$

□

Adaptive Gaussian Quadrature

According to Pinheiro and Bates (1995), Gaussian quadrature rules are used to approximate integrals of functions with respect to a given kernel by a weighted average of the integrand evaluated at predetermined abscissas.
Weights and abscissas used in Gaussian quadrature rules, (GQ rules), for the most common kernels can be obtained from the tables of Abramowitz and Stegun (1964) or by using an algorithm proposed by Golub (1973) or Golub and Welsh (1969).
GQ rules for multiple integrals are known to be numerically complex as shown by Davis and Rabinowitz (1984), but by using the structure of the integrand in the nonlinear mixed-effect model by Pinheiro and Bates (1995), we can transform the problem into successive applications of simple one-dimensional GQ rules.
Now considering Pinheiro and Bates (1995), single-level NLME model;
A natural candidate for the kernel function for the quadrature rule in the single-level NLME model is the marginal distribution of the random effects, that is, the $N(0,\psi)$ density.
The Gaussian quadrature rule in this case can be viewed as a deterministic version of a Monte Carlo integration algorithm in which random samples of the random effects, $\mathbf{b}_i$, are generated from the $N(0,\psi)$ distribution.
The samples and weights in the GQ rule are fixed before hand while in Monte Carlo integration are left to random choice.
Importance sampling tends to be much more efficient than simple Monte Carlo integration as shown by Geweke (1989). Now we consider an importance sample version of the GQ rule, which was denoted by **adaptive Gaussian quadrature**
The critical step for the success of importance sampling is the choice of an importance distribution that approximates the integrand.
For the single NLME model the integrand is proportional to

$$\exp[-g(\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i)/2\sigma^2] \tag{2.119}$$

as can be observed from (2.98) which is approximated by a $N(\hat{\mathbf{b}}_i,\sigma^2[G(\beta,\Delta,\mathbf{y}_i)]^{-1})$ density with $\hat{\mathbf{b}}_i$ defined as

$$
\begin{aligned}
\hat{\mathbf{b}}_i &= \hat{\mathbf{b}}_i(\beta,\Delta,\mathbf{y}_i) \\
&= \underset{bi}{\operatorname{argmin}}\, g(\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i)
\end{aligned}
\tag{2.120}
$$

and $G(\beta, \Delta, \mathbf{y}_i)$ defined as

$$\begin{aligned} g''(\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i) &\simeq G(\beta, \Delta, \mathbf{y}_i) \\ &= f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} f_i'(\beta, \mathbf{b}_i)|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} + \Delta^T\Delta \\ &= \frac{\partial f_i(\beta, \mathbf{b}_i)}{\partial \mathbf{b}_i}|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} \frac{\partial f_i(\beta, \mathbf{b}_i)}{\partial \mathbf{b}_i^T}|_{\mathbf{b}_i=\hat{\mathbf{b}}_i} + \Delta^T\Delta \end{aligned} \tag{2.121}$$

as earlier defined in (2.106).

Lemma 9. *The likelihood adaptive Gaussian Quadrature approximation is given by*

$$\begin{aligned} L_{AGQ}(\beta, \sigma^2, \Delta|\mathbf{y}) = \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} |[G(\beta, \Delta, \mathbf{y}_i)]|^{-1/2} \\ \times \sum_{j1=1}^{N_{GQ}}, \cdots, \sum_{jq=1}^{N_{GQ}} \exp\left\{-g[\beta, \Delta, \mathbf{y}_i, \hat{\mathbf{b}}_i + \sigma(G(\beta, \Delta, \mathbf{y}_i))^{-1/2}\mathbf{z_j}]/2\sigma^2 + \|\mathbf{z_j}\|^2/2\right\} \prod_{k=1}^{q} \mathbf{w_{jk}} \end{aligned}$$

where

- $\mathbf{z_j}, j = 1, \cdots, N_{GQ}$ *denote the abscissas*
- $\mathbf{w_j}, j = 1, \cdots, N_{GQ}$ *denote the weights*

for the (one-dimensional) GQ rule with N_{GQ} points based on the $N(0, 1)$ kernel;

Proof. The integrand (2.119) is the importance distribution used in the adaptive GQ, so that the grid of abscissas in the $\mathbf{b}_i$ scale is centered around the conditional modes $\hat{\mathbf{b}}_i$ and $G(\beta, \Delta, \mathbf{y}_i)$ is used for scaling.

And from above we have

$$\begin{aligned} \mathbf{b}_i &\sim N(\hat{\mathbf{b}}_i, \sigma^2[G(\beta, \Delta, \mathbf{y}_i)]^{-1}) \\ &\Rightarrow \mathbf{z} = \frac{\mathbf{b}_i - \hat{\mathbf{b}}_i}{(\sigma^2[G(\beta, \Delta, \mathbf{y}_i)]^{-1})^{1/2}} \sim N(0, 1) \\ &\Rightarrow \mathbf{b}_i = \hat{\mathbf{b}}_i + \sigma[G(\beta, \Delta, \mathbf{y}_i)]^{-1/2}\mathbf{z} \end{aligned} \tag{2.122}$$

Thus from

$$\begin{aligned} \mathbf{b}_i &= \hat{\mathbf{b}}_i + \sigma[G(\beta, \Delta, \mathbf{y}_i)]^{-1/2}\mathbf{z} \\ &\Rightarrow d\mathbf{b}_i = \sigma^q|[G(\beta, \Delta, \mathbf{y}_i)]^{-1/2}|d\mathbf{z} \end{aligned} \tag{2.123}$$

Hence the adaptive Gaussian quadrature (AGQ) rule is given by

$$\begin{aligned}
\int \exp[-g(\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i)/2\sigma^2]d\mathbf{b}_i = &\int \sigma^q|[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2} \\
&\times \exp\left\{-g[\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i+\sigma(G(\beta,\Delta,\mathbf{y}_i))^{-1/2}\mathbf{z}]/2\sigma^2+\|\mathbf{z}\|^2/2\right\} \\
&\times \exp(-\|\mathbf{z}\|^2/2)d\mathbf{z} \\
\simeq\ & \sigma^q|[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2}\times\sum_{j1=1}^{N_{GQ}},\cdots,\sum_{jq=1}^{N_{GQ}} \\
&\times \exp\left\{-g[\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i+\sigma(G(\beta,\Delta,\mathbf{y}_i))^{-1/2}\mathbf{z_j}]/2\sigma^2+\|\mathbf{z_j}\|^2/2\right\} \\
&\times \prod_{k=1}^{q}\mathbf{w_{jk}}
\end{aligned} \tag{2.124}$$

Combining previous work with (2.84) and (2.85) we can define the adaptive Gaussian Quadrature as

$$\begin{aligned}
L_{AGQ}(\beta,\sigma^2,\Delta|\mathbf{y}) &= f(\mathbf{y}|\beta,\sigma^2,\Delta) \\
&= \prod_{i=1}^{M} f(\mathbf{y}_i|\beta,\sigma^2,\Delta) \\
&= \prod_{i=1}^{M}\int f(\mathbf{y}_i|\mathbf{b}_i\beta,\sigma^2,\Delta).f(\mathbf{b}_i|\beta,\sigma^2\Delta)d\mathbf{b}_i \\
&= \prod_{i=1}^{M}\int \frac{1}{(2\pi\sigma^2)^{n_i/2}}\exp\left\{\frac{-\|y_i-f_i(\phi_i,\vartheta_i)\|^2}{2\sigma^2}\right\} \\
&\quad \times \frac{1}{(2\pi)^{q/2}\sigma^q abs|\Delta|^{-1}}\exp\left\{\frac{-\|\Delta\mathbf{b}_i\|^2}{2\sigma^2}\right\} d\mathbf{b}_i \\
&= \prod_{i=1}^{M}\frac{1}{(2\pi\sigma^2)^{n_i/2}}\frac{1}{(2\pi)^{q/2}\sigma^q abs|\Delta|^{-1}}\int \exp\left\{\frac{-\|y_i-f_i(\phi_i,\vartheta_i)\|^2+\|\Delta\mathbf{b}_i\|^2}{2\sigma^2}\right\} d\mathbf{b}_i \\
&= \prod_{i=1}^{M}\int \frac{|\Delta|}{(2\pi\sigma^2)^{n_i/2}}\exp\left\{\frac{-(\|y_i-f_i(\phi_i,\vartheta_i)\|^2+\|\Delta\mathbf{b}_i\|^2)/2\sigma^2}{(2\pi\sigma^2)^{q/2}}\right\} d\mathbf{b}_i \\
&= \prod_{i=1}^{M}\int \frac{|\Delta|}{(2\pi\sigma^2)^{n_i/2}}\exp\left\{\frac{-(\|y_i-f_i(\beta,\mathbf{b}_i)\|^2+\|\Delta\mathbf{b}_i\|^2)/2\sigma^2}{(2\pi\sigma^2)^{q/2}}\right\} d\mathbf{b}_i
\end{aligned} \tag{2.125}$$

Since we noted that

$$\begin{aligned}
f_i(\beta,\mathbf{b}_i) &= f_i(\phi_i,\vartheta_i) \\
&= f_i[\phi_i(\beta,\mathbf{b}_i),\vartheta_i]
\end{aligned}$$

From (2.99) we can re-write (2.125) as

$$L_{AGQ}(\beta,\sigma^2,\Delta|\mathbf{y}) = \prod_{i=1}^{M} \frac{|\Delta|}{(2\pi\sigma^2)^{n_i/2}} \int \exp\left\{ \frac{-g(\beta,\Delta,\mathbf{y}_i,\mathbf{b}_i)/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} d\mathbf{b}_i \tag{2.126}$$

Incorporating (2.122), (2.123) together with (2.124) in (2.126), we end up getting

$$\begin{aligned}
L_{AGQ}(\beta,\sigma^2,\Delta|\mathbf{y}) &= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} \int \sigma^q |[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2} \\
&\quad \times \frac{\exp\left\{-g[\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i + \sigma(G(\beta,\Delta,\mathbf{y}_i))^{-1/2}\mathbf{z}]/2\sigma^2 + \|\mathbf{z}\|^2/2\right\} \times \exp(-\|\mathbf{z}\|^2/2)}{(2\pi\sigma^2)^{q/2}} d\mathbf{z} \\
&= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} |[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2} \\
&\quad \times \int \frac{\exp\left\{-g[\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i + \sigma(G(\beta,\Delta,\mathbf{y}_i))^{-1/2}\mathbf{z}]/2\sigma^2 + \|\mathbf{z}\|^2/2\right\} \times \exp(-\|\mathbf{z}\|^2/2)}{(2\pi)^{q/2}} d\mathbf{z} \\
&= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} |[G(\beta,\Delta,\mathbf{y}_i)]|^{-1/2} \\
&\quad \times \sum_{j1=1}^{N_{GQ}}, \cdots, \sum_{jq=1}^{N_{GQ}} \exp\left\{-g[\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i + \sigma(G(\beta,\Delta,\mathbf{y}_i))^{-1/2}\mathbf{z_j}]/2\sigma^2 + \|\mathbf{z_j}\|^2/2\right\} \prod_{k=1}^{q} \mathbf{w_{jk}}
\end{aligned} \tag{2.127}$$

□

Thus the adaptive Gaussian approximation to the log-likelihood function in the single-level NLME model is then

$$\begin{aligned}
logL_{AGQ}(\beta,\sigma^2,\Delta|\mathbf{y}) &= l_{AGQ}(\beta,\sigma^2,\Delta|\mathbf{y}) \\
&= \frac{-N}{2} log(2\pi\sigma^2) + Mlog|\Delta| - \frac{1}{2}\sum_{i=1}^{M} log|G(\beta,\Delta,\mathbf{y}_i)| \\
&\quad + \sum_{i=1}^{M} log\left\{ \sum_{j=1}^{N_{GQ}} \exp(-g[\beta,\Delta,\mathbf{y}_i,\hat{\mathbf{b}}_i + \sigma(G(\beta,\Delta,\mathbf{y}_i))^{-1/2}\mathbf{z_j}]/2\sigma^2 + \|\mathbf{z_j}\|^2/2) \prod_{k=1}^{q} \mathbf{w_{jk}} \right\}
\end{aligned} \tag{2.128}$$

Lemma 10. *The likelihood Gaussian Quadrature version of Monte Carlo integration is given as*

$$L_{GQ}(\beta,\sigma^2,\Delta|\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} \sum_{j1=1}^{N_{GQ}}, \cdots, \sum_{jq=1}^{N_{GQ}} \exp[-\left\|y_i - f_i(\beta,\sigma abs|\Delta|^{-1}\mathbf{z}^*_{\mathbf{j1},\cdots,\mathbf{jq}})\right\|^2]/(2\sigma^2) \prod_{k=1}^{q} \mathbf{w_{jk}}$$

where $\mathbf{z}^*_{\mathbf{j1},\cdots,\mathbf{jq}} = (z^*_{j1},\cdots,z^*_{jq})^T$.

Proof. Now viewing Gaussian Quadrature as a deterministic version of Monte Carlo integration in which random samples of $\mathbf{b}_i$ are generated from the $N(0, \psi)$ distribution and letting $\mathbf{z}_\mathbf{j}^*$ and $\mathbf{w}_\mathbf{j}$ be as previously defined, that is,

- $\mathbf{z}_\mathbf{j}^*, j = 1, \cdots, N_{GQ}$ denote the abscissas
- $\mathbf{w}_\mathbf{j}, j = 1, \cdots, N_{GQ}$ denote the weights

for the one-dimensional GQ rule with N_{GQ} points based on the $N(0, 1)$ kernel; and from (2.125) we get

$$\begin{aligned} f(\mathbf{y}|\beta, \sigma^2, \Delta) &= \prod_{i=1}^{M} \int \frac{|\Delta|}{(2\pi\sigma^2)^{n_i/2}} \exp\left\{ \frac{-(\|y_i - f_i(\beta, \mathbf{b}_i)\|^2 + \|\Delta\mathbf{b}_i\|^2)/2\sigma^2}{(2\pi\sigma^2)^{q/2}} \right\} \, d\mathbf{b}_i \\ &= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} \int \frac{\exp -(\|y_i - f_i(\beta, \mathbf{b}_i)\|^2)/(2\sigma^2) \exp -(\|\Delta\mathbf{b}_i\|^2)/(2\sigma^2)}{(2\pi\sigma^2)^{q/2}} \, d\mathbf{b}_i \end{aligned} \tag{2.129}$$

Since $\mathbf{b}_i \sim N(0, \psi)$, where $\psi = \sigma^2(\Delta^T\Delta)^{-1}$, then $\mathbf{b}_i \sim N(0, \sigma^2(\Delta^T\Delta)^{-1})$

$$\begin{aligned} \mathbf{z}^* &= \frac{\mathbf{b}_i - 0}{[\sigma^2(\Delta^T\Delta)^{-1}]^{1/2}} \sim N(0, 1), \text{(as required)} \\ &\Rightarrow \mathbf{b}_i = [\sigma^2(\Delta^T\Delta)^{-1}]^{1/2}\mathbf{z}^* \\ &\Rightarrow \mathbf{b}_i = \sigma \times abs|\Delta|^{-1}\mathbf{z}^* \end{aligned} \tag{2.130}$$

Applying change of variable technique,then from (2.130), we have

$$d\mathbf{b}_i = \sigma^q |\Delta|^{-1} \, d\mathbf{z}^* \tag{2.131}$$

Hence

$$
\begin{aligned}
L_{GQ}(\beta,\sigma^2,\Delta|\mathbf{y}) &= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}}\prod_{i=1}^{M} \\
&\times \int \frac{\exp-(\|y_i - f_i(\beta,\sigma abs|\Delta|^{-1}\mathbf{z}^*)\|^2)/(2\sigma^2)\exp-(\|\Delta\sigma abs|\Delta|^{-1}\mathbf{z}^*\|^2)/(2\sigma^2)}{(2\pi\sigma^2)^{q/2}} \\
&\times \sigma^q|\Delta|^{-1}\ d\mathbf{z}^* \\
&= \frac{|\Delta|^M}{(2\pi\sigma^2)^{N/2}}\prod_{i=1}^{M} \\
&\times \int \frac{\exp-(\|y_i - f_i(\beta,\sigma abs|\Delta|^{-1}\mathbf{z}^*)\|^2)/(2\sigma^2)\exp-(\|\mathbf{z}^*\|^2)/(2\sigma^2)}{(2\pi\sigma^2)^{q/2}}\sigma^q|\Delta|^{-1}\ d\mathbf{z}^* \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}}\prod_{i=1}^{M} \\
&\times \int \frac{\exp-(\|y_i - f_i(\beta,\sigma abs|\Delta|^{-1}\mathbf{z}^*)\|^2)/(2\sigma^2)\exp-(\|\mathbf{z}^*\|^2)/(2\sigma^2)}{(2\pi)^{q/2}}\ d\mathbf{z}^* \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}}\prod_{i=1}^{M}\sum_{j1=1}^{N_{GQ}},\cdots,\sum_{jq=1}^{N_{GQ}}\exp[-\left\|y_i - f_i(\beta,\sigma abs|\Delta|^{-1}\mathbf{z}^*_{\mathbf{j1},\cdots,\mathbf{jq}})\right\|^2]/(2\sigma^2)\prod_{k=1}^{q}\mathbf{w_{jk}}
\end{aligned}
\tag{2.132}
$$

where $\mathbf{z}^*_{\mathbf{j1},\cdots,\mathbf{jq}} = (z^*_{j1},\cdots,z^*_{jq})^T$. □

Thus, the corresponding approximation to the log-likelihood function is

$$
l_{GQ}(\beta,\sigma^2,\Delta|\mathbf{y}) = \frac{-N}{2}log(2\pi\sigma^2) + \sum_{i=1}^{M} log\left\{\sum_{j}^{N_{GQ}}\exp[-\left\|y_i - f_i(\beta,\sigma abs|\Delta|^{-1}\right\|^2]/(2\sigma^2)\prod_{k=1}^{q}\mathbf{w_{jk}}\right\}
\tag{2.133}
$$

Having gone through the theory on determining these estimates, we find that, there is no closed form solution for these estimates by maximum likelihood or restricted maximum likelihood methods. Thus, they are determined by iterative algorithms such as EM iterations or general nonlinear optimization which actually need a computer software. In our data analysis to estimate these parameters we would use lme4 package in R statistical software. Inference on the parameters of a LME model usually relies on approximate distributions for the MLE and REML estimates derived from asymptotic results. Pinheiro (1994) showed that under certain regularity conditions generally satisfied in practice, the maximum likelihood estimates in the general LME model (2.1) are consistent and asymptotically normal.

Chapter 3

Data Analysis

3.1 Introduction

Helminth (parasites that reside in an animal's intestines) constitute one of the most important constraints to small ruminant livestock production in the tropics resulting in widespread infection in grazing animals, associated production losses, high costs of treatment and death. Current control methods in the tropics focus on reducing contamination of pastures through anthelmintic treatment of animals and/or controlled grazing. But there are problems with increasing frequencies of drug resistance. An attractive, alternative and sustainable solution is the breeding for disease resistance. Indeed, anecdotal evidence suggests that among the large and diverse range of indigenous breeds of sheep and goats in the tropics there are some that appear to have the genetic ability to resist or tolerate helminthiasis. One of these is the Red Maasai breed found in East Africa and perceived to be resistant to the disease. The Red Maasai is a fat-tailed sheep associated with the Maasai tribe found in northern Tanzania and south-central Kenya.

ILRI decided in 1990 to investigate the degree of resistance exhibited by this Red Maasai breed and initiated a study at Diani Estate of the Baobab Farms, 20 km south of Mombasa in the sub-humid coastal region of Kenya. To do so, a susceptible breed, the Dorper, originally from South Africa, was chosen to provide a direct comparison with the Red Maasai. The Dorper breed was developed in South Africa in the 1940s by inter breeding the Dorset Horn and Black Head Persian breeds. The Dorper is particularly well adapted to harsh, arid conditions and was imported into Kenya in the 1960s. This breed is also larger than the Red Maasai, and this makes these sheep attractive to farmers.

As well as comparing the performance of the different genotypes when exposed to helminthiasis, it is also of interest to examine genetic variation among rams and ewes within genotypes. To do this we need to use what are known as restricted or residual maximum likelihood (REML) procedures which are able to simultaneously estimate random and fixed effects.

Once the random estimates are known these can then be used to obtain heritability estimates which determine the proportion of the variation among offspring that has been handed down from parents.

Our major objective is examining incorporation of random effects to study variations among rams (sires) and ewes (dams) and their influences on lamb weaning weight.

We have only two breeds of ram and so it would not be sensible to infer that these two breeds are a random sample from a much larger population of ram breeds. This is not only because they were specifically chosen for this study, but also because a sample of two would not be considered large enough to generalize to "all breeds". Here the possibility of year being random might also be considered. Six levels, as here, are probably about the minimum number that could be considered as adequate for estimating random components. Thus, for a study carried out over only three or four years, the sample would be hardly large or random enough to be representative of a wider population of years.

3.2 Description of contents of the Data

The data set contains information on 882 lambs born and raised at Diani Farm on Kenya coast between 1991 and 1996. Records for weaning weights are missing in 182 of the lambs, mostly because of earlier death or because recording was missed. Missing data are indicated by blanks. A! at the end of the variable name implies that the variable is being used as a factor.
This data could be found in ILRI 2006, Biometrics and Research Methods Teaching Resource Version 1 edited by John Rowland, Case Study 4.
The fields contained in the data are in Table 3.1

Table 3.1: Description of contents of the data

Field	Description
LAMB	Individual lamb identification
EWE-ID	Identification of lamb's dam
EWE-BRD	Breed of ewe (D = Dorper and R = Red Maasai)
RAM-ID	Identification of lamb's sire
RAM-BRD	Breed of ram (D = Dorper and R = Red Maasai)
BREED!	Breed of the lamb (DD = pure bred Dorper, DR = Dorper sire × Red Maasai dam, RD = Red Maasai sire × Dorper ewe, RR = pure bred Red Maasai)
YEAR!	The year of birth of the lamb (1991-1996)
SEX!	The sex of the lamb (M = male and F = female)
BIRTHWT	Weight (kg) of lamb at birth
AGEWEAN	Age (day) of lamb at weaning
DAMAGE!	Age (year) of dam
WEANWT	Weight (kg) of lamb at weaning
DAMAGE7!	Calculated from DAMAGE in order to represent DAMAGE in 7 categories ($\leq$ 2,3,4,5,6,7,$\geq$ 8)
DL	Duplicate of DAMAGE7 but considered as a covariate, not a factor
DQ	Calculated as DL*DL
DAMAGE4!	Calculated from DAMAGE7 but collapsed into four categories ($\leq$ 2,3-4,5-6,$\geq$ 7)

3.3 Data Exploration

Sex of a lamb is an example of a fixed effect, can only have one of two values: male and female.
Assumption made:
The sample of rams (or ewes) used in the study is a random selection of rams (or ewes) from the particular genotype at large hence influence of the ram (or ewe) on the growth of its offspring is now considered to be a random effect.
In mixed model analysis we have different types of units occurring at different layers namely in this example: lambs, ewes, rams. The investigational or observational units defined within layers are assumed to be chosen independently of one another; usually they are chosen at 'random'. They will therefore be random effects in our mixed model.
We have two breeds. From within each of the two breeds a number of rams is selected. These are the observational units (ram chosen as a random effect) against which the two breeds of rams should be compared.
We do exactly the same for ewes.
Rams and ewes are mated both within and across breeds to produce their offspring. These offsprings(lambs) are the investigational units at the next layer. Fixed effects or attributes that might be considered for each lamb are: breed, age of ewe, ram breed × ewe breed, year of birth, sex and age at weaning.
Before undertaking our statistical analysis it is useful to first explore the relationships between weaning weight and various covariates of interest to see how best these relationships might be included in the statistical model.
The weight at weaning appears normally distributed, as indicated by the relative position of the median within the box that contain half the data though there are some few 'outliers' as shown in the Figure 3.1. A plot of weaning weight against ewe breed reveals that offsprings

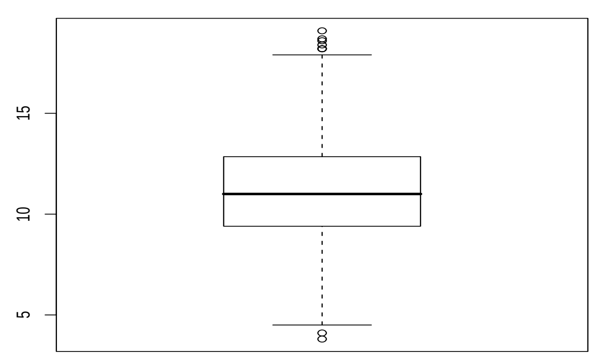

Figure 3.1: A box plot of weight at weaning

from Dorper ewe breed are found to have high weaning weight than those from Red Maasai ewe breed (see Figure 3.2) but rams from both the two breeds have almost an equal effect on the offsprings' weaning weight as Figure 3.3 shows. Figure 3.4 shows that sex of the lamb affects the weaning weight slightly. Mean weaning weight of the male lambs is slightly higher than that of female lambs. Mean weaning weight of lambs decreased gradually as the year

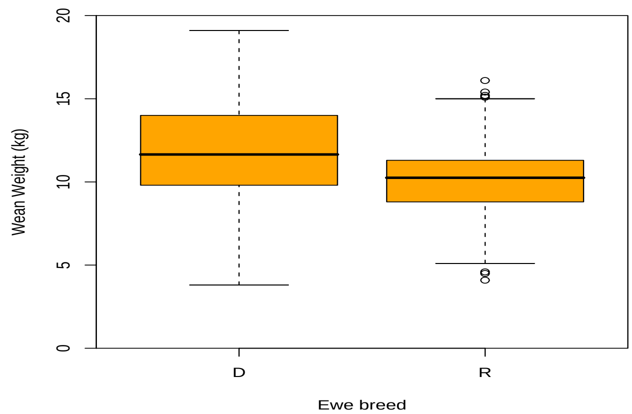

Figure 3.2: Effect of ewe breed on the weight at weaning

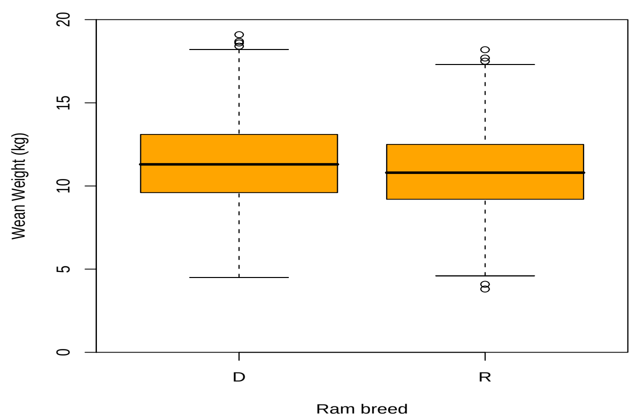

Figure 3.3: Effect of ram breed on the weight at weaning

of birth moved from 1991 to 1996 [see Figure 3.5]. The box plot by age of dam (DL)[see Figure 3.6] illustrates the association between weaning weight and the age of a lamb's dam. There are more 'outliers' shown in this diagram. This is probably because the variation among genotypes is not accounted for in this series of box plots. The plot shows that an offspring's weaning weight appears to increase as a dam increase in age from 2 to 5 years and to decrease from 6 years onwards. We can fit DL as a factor with seven levels. Figure 3.7 demonstrates a linear relationship between age of lamb at weaning and the weaning weight. Hence we can include the age at weaning as a continuous covariate in order to correct for its effect on weaning weight.

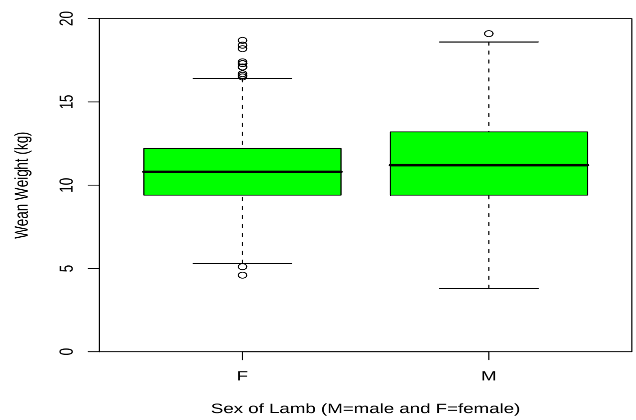

Figure 3.4: Effect of sex of lamb on the weight at weaning

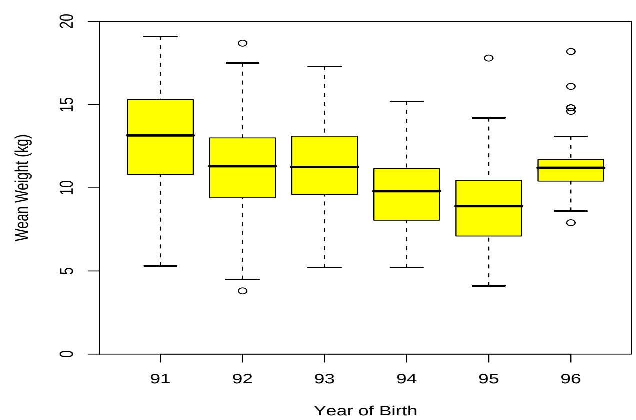

Figure 3.5: Effect of year of birth on the weight at weaning

3.4 Data Analysis by REML

3.4.1 Model selection

We now undertake a full mixed model analysis for lamb weaning weight with ram and ewe defined as random effects for a combined model to investigate the influence of each of the fixed effects on weaning weight.
We first fitted a generalized linear model to check the significance of each of our fixed effects, that is: Year; Sex; Agewean; DL-linear term for dam age; DQ-quadratic term for dam age; Ewe breed; and Ram breed.
An R - Output of our fit is given in the Appendix A.3.1. Calculated lower level (LL) and upper level (UL) confidence intervals of our fixed effects were also given in Table 3.2.

- Response variable:WEANWT
- Fixed effects:YEAR,SEX,AGEWEAN,DL,DQ,EWE-BRD,RAM-BRD

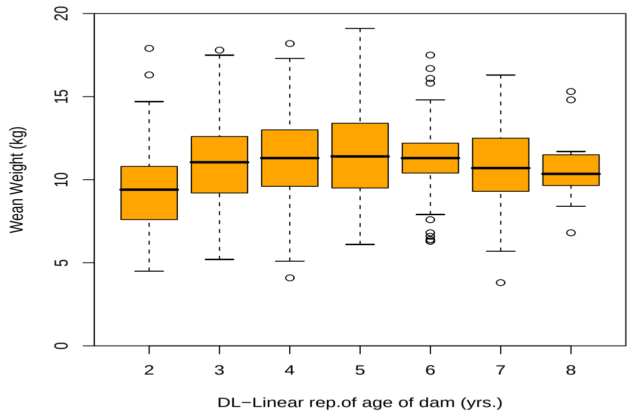

Figure 3.6: Effect of age of dam on the weight at weaning

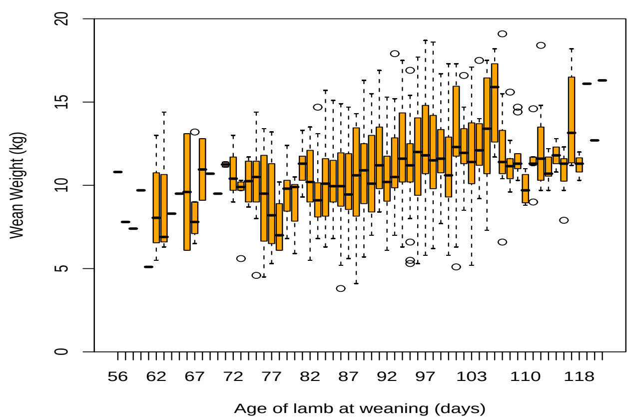

Figure 3.7: Effect of age of lamb at weaning on the weight at weaning

From the Table 3.2, we found

- Lambs born in the later years had lower weaning weights compared with those born in the earlier years. All the years were very significant relative to 1991.
- Male lambs had an average weaning weight slightly higher by 0.48($\pm$0.17)kg than females. Male lambs were significant relatively to female lambs.
- The age at which weaning was done was very significant from the confidence interval and the p-value.
- Age of ewe (DL and DQ) and the main effects (ewe breed and ram breed) were also very significant.

  DL and DQ are different representation of the effect of DAMAGE(age of dam) on weaning weight. DL represents a linear fit while DQ represents a quadratic fit. The two effects were highly significant. DAMAGE is divided into two categories: DAMAGE7-with 7 (categories) levels, DAMAGE4-with 4 levels.

Table 3.2: Table of Model giving significance of Fixed Effects

		Estimate	Std. Error	95% CI		p-value
				LL	UL	
YEAR	91	Reference				
	92	-1.57	0.29	-2.14	-0.99	$< .001$
	93	-1.10	0.28	-1.64	-0.56	$< .001$
	94	-2.83	0.36	-3.53	-2.13	$< .001$
	95	-3.23	0.34	-3.90	-2.55	$< .001$
	96	-2.35	0.39	-3.12	-1.59	$< .001$
SEX	F	Reference				
	M	0.48	0.17	0.15	0.81	0.005
AGEWEAN		0.07	0.01	0.05	0.09	$< .001$
DL		2.73	0.32	2.11	3.34	$< .001$
DQ		-0.27	0.03	-0.34	-0.20	$< .001$
EWE BREED	Dorper	Reference				
	Red Maasai	-0.59	0.24	-1.05	-0.12	0.014
RAM BREED	Dorper					
	Red Maasai	-0.44	0.17	-0.78	-0.10	0.011

A linear regression done before showed that reduced number of categories (DAMAGE4) was not a good representation of the association with age as with 7 categories (DAMAGE7). A linear regression done of the comparison between DL, DQ and DAMAGE7 gave an insignificant variation showing the quadratic fit was a good one.

To introduce our random effects, we compared three linear mixed models,

1. Model 2: with ewe and ram as random components.

$$Y_{ijk} = \mu + y_{ijk} + s_{ijk} + A_{ijk} + d_j + d_j^2 + e_j + r_i + E_j + R_i + \epsilon_{ijk} \tag{3.1}$$

   where Y_{ijk} is the weaning weight of the $k-$th lamb born to the $j-$th dam and the $i-$th ram, μ is the overall mean, y_{ijk} is the year of birth of the $k-$th lamb born to the $j-$th dam and the $i-$th ram, s_{ijk} is the sex of the $k-$th lamb born to the $j-$th dam and the $i-$th ram, A_{ijk} is the age at weaning of the $k-$th lamb born to the $j-$th dam and the $i-$th ram, d_j and d_j^2 is the linear and quadratic representation of age of the $j-$th dam respectively, e_j and r_i is the main effects of the $j-$th ewe breed and the $i-$th ram breed respectively, E_j is the random effect of the $j-$th ewe, R_i is the random effect due of the $i-th$ ram and ϵ_{ijk} is the random error of the $k-$th lamb born to the $j-$th dam and the $i-$th ram.

   It could as well be shortened to be in the form

$$Y_{ijk} = \mathbf{X}\beta + E_j + R_i + \epsilon_{ijk} \tag{3.2}$$

   where Y_{ijk} is the weaning weight of the $k-$th lamb born to the $j-$th dam and the $i-$th ram, $\mathbf{X}$ is incidence matrix for the fixed effects, β is vector of associated parameters, E_j is the random effect due to the $j-$th ewe ,R_i is the random effect due to the $i-th$ ram and ϵ_{ijk} is the random error of the $k-$th lamb born to the $j-$th dam and the $i-$th ram.

2. Model 3: with ewe as a random component, where only R_k is removed in the above model.

3. Model 4: with ram as a random component, where only E_j is removed in the above model.

In all the models the following assumptions on the random terms hold E_j are i.i.d $N(0, \sigma_e^2)$ R_k are i.i.d $N(0, \sigma_r^2)$ ϵ_{ijk} are i.i.d $N(0, \sigma^2)$ and E_j, R_k and ϵ_{ijk} are assumed to be independent. The number of fixed effects in the three models were the same.
The first two models have all their log-likelihood, maximum-likelihood deviance and restricted maximum-likelihood deviance equal, but differ in their Akaike Information Criterion (AIC) and Bayesian or Schwarz Information Criterion (BIC). Thus we can choose model 3 with the smaller values of AIC and BIC.
These two comparison criteria could be evaluated as

$$AIC = -2l(\hat{\theta}|\mathbf{y}) + 2n_{par} \tag{3.3}$$

and

$$BIC = -2l(\hat{\theta}|\mathbf{y}) + 2n_{par}log(N) \tag{3.4}$$

where n_{par} denotes number of parameters in the model.
Though ram contributes genetically to the variation in the lambs slightly, we may chose model 3 on the basis of law of parsimony and its low values of AIC and BIC. The table below gives their log-likelihood and deviance.

Table 3.3: Model selection measures

Model	AIC	BIC	logLik	Ml deviance	REMLdeviance
2	3110	3173	-1541	3053	3082
3	3108	3167	-1541	3053	3082
4	3146	3205	-1560	3092	3120

Our model 3 gives a linear mixed effect model fit by restricted maximum-likelihood for lamb weaning weight with ewe as random effects.

Table 3.4: Table of Effects for the model relating

		Estimate	Std. Error	95% CI	
				LL	UL
YEAR	91	Reference			
	92	-1.60	0.26	-2.11	-1.08
	93	-1.09	0.26	-1.60	-0.59
	94	-3.01	0.34	-3.67	-2.34
	95	-3.30	0.34	-3.96	-2.64
	96	-2.44	0.39	-3.20	-1.68
SEX	F	Reference			
	M	0.41	0.16	0.09	0.72
AGEWEAN		0.07	0.01	0.05	0.08
DL		2.92	0.30	2.34	3.50
DQ		-0.29	0.03	-0.35	-0.23
EWE BREED	Dorper	Reference			
	Red Maasai	-0.46	0.27	-0.99	0.06
RAM BREED	Dorper				
	Red Maasai	-0.42	0.16	-0.74	-0.10

But our main objective was to examine incorporation of random effects to study variations among rams (sires) and ewes (dams) and their influences on lamb weaning weight. Thus to achieve our goal we must choose Model 2 since it contains both rams and ewes.
Comparison of the RAM-ID and EWE-ID variance components indicates that the variance component for ewes (1.456488) is highly significant but that for ram (0.066577) is not.
With the random terms (ewes and rams) specified in the model the estimate of the residual among lamb variance is reduced from 4.9 to 3.427208 kg.
This is due to taking into account the variations among rams and ewes within breeds.
With the ewe random term alone specified in the model,the estimate of the residual among lamb variance is 3.4968kg.
This output is due to taking into account the variations among ewes within breeds reduced from an approximate 4.9kg output assuming all variation to be at the lamb level.
Ewe variance component indicates the variance component for ewe (1.4459kg)is highly significant.
The mixed model with the ewe component alone included utilizes almost an equivalent information as the mixed model with both ewe and ram components included.

Chapter 4

Conclusion

We had set out to review methods of estimating both linear mixed effects and nonlinear mixed effects models, investigate the computational efficiency and accuracy of these and other computational methods, like the b-splines, that could be used to approximate the log-likelihood function in non-linear mixed effects models.
We have critically reviewed methods of estimating both linear mixed effects and nonlinear mixed effects models.We have gone through the review of LME approximations, Laplacian approximations, importance sampling and Gaussian quadratures in approximating the log-likelihood. We also wanted to investigate the possibility of developing REML versions for the last three approximation methods.
Computational and numeric methods of approximating the likelihood in nonlinear mixed effects models needs quite some longer time to be reviewed. Investigation of the computational efficiency and accuracy of the b-splines, that could be used as an alternative to approximate the log-likelihood function in non-linear mixed effects models was also not achieved in our study. This could make quite an interesting area for some further research.
Mixed effects models constitute a powerful tool for modeling dependence within clustered data. They give an intuitive interpretation for the source and the structure of the dependence and can easily handle the unbalanced data that are frequently encountered in many areas of scientific investigation.We have given a case study using R on a data set from livestock with linear mixed-effects.
Despite their usefulness, mixed effects models remain a difficult area for many researchers that could benefit from their application.

4.1 References

1. Bates, D.M.(2007). Theory and Computational methods for mixed models. Department of statistics U. of Wisconsin-Madison

2. Bates, D.M.(2007). Linear mixed-effects models using S4 classes. U. of Wisconsin-Madison

3. Bates, D.M.(2007). Linear mixed model implementation in lme4. Department of statistics U. of Wisconsin-Madison

4. Bates, D.M.; R Development Core Team (2006). Statistical Inferences in Linear Mixed-Effects Models

5. Bates, D.M.(2004). Multilevel Models in R: Present and Future. U. of Wisconsin-Madison, U.S.A.

6. Bates, D.M. and DebRoy, S.(2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*,

7. Bates, D.M.; R Development Core Team (2004). Sparse Matrix Representation of Linear mixed models.

8. Bates, D.M. and Pinheiro,J.C.(). Computational methods for Multilevel Modeling.

9. Case Study 4. Mixed model analysis for the estimation of components of genetic variation in lamb weaning weight. *Biometrics and Research Methods Teaching Resource Version 1 (ed. John Rowlands).* ILRI (International Livestock Research Institute), Nairobi, Kenya.

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

11. Lindstrom, M.J. and Bates, D.M.(1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673 – 687.

12. Lindstrom, M.J. and Bates, D.M.(1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects models for Repeated Measures Data. *Journal of the American Statistical Association*, **83**, 1014 – 1022.

13. Pinheiro, J.C. and Bates, D.M.(2000). Mixed-Effects models in S and S-PLUS. Springer, 2000. ISBN 0-387-98957-9.

14. Pinheiro, J.C. and Bates, D.M.(1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-Effects model. *Journal of Computational and Graphical Statistics*, **4**, 12 – 35.

15. Pinheiro, J.C. and Bates, D.M.(1995). Mixed-Effects models methods and classes for S and Splus. U. of Wisconsin-Madison.

16. Pinheiro, J.C.(1994). "Topics in Mixed-Effects models." PhD thesis, U. of Wisconsin-Madison, Department of Statistics.

17. Wolfinger, R.(1993). Laplace's Approximation for Nonlinear Mixed models. *Biometrika*, **80**, 791 – 795.

Appendix A

The Appendix

A.1 Orthogonal Triangular Decomposition

Orthogonal Triangular Decomposition of rectangular matrices are a prefered numerical method for solving least squares problems. Also called QR decomposition. The decomposition can be written as

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix} \tag{A.1}$$

In this case X is an $n \times p$ matrix $(n \geq p)$ of rank p, Q is a $n \times n$ and orthogonal matrix, R is a $p \times p$ and upper triangular matrix and 0 is a $(n-p) \times p$ matrix of zeroes. The equation in (A.1) can also be written as

$$X = Q_t R$$

where Q_t (Q-truncated) consists of the first p columns of Q. It is important to note that Q is orthogonal, that is

$$\begin{aligned} Q^T Q &= Q Q^T \\ &= I \\ &\Rightarrow Q_t^T Q_t = I \end{aligned} \tag{A.2}$$

Orthogonal matrices preserve norms of vectors under multiplication either by Q or by Q^T. The transformation represented by Q is a generalization of a rotation or a reflection in the plane. We have

$$\begin{aligned} \left\|Q^T y\right\|^2 &= (Q^T y)^T Q^T y \\ &= y^T Q Q^T y \\ &= y^T y \\ &= \|y\|^2 \end{aligned} \tag{A.3}$$

Applying this transformation to the residual vector in a least squares problem we get

$$\begin{aligned} \|y - X\beta\|^2 &= \left\|Q^T(y - X\beta)\right\|^2 \\ &= \left\|Q^T y - Q^T X\beta\right\|^2 \\ &= \left\|c - Q^T Q \begin{bmatrix} R \\ 0 \end{bmatrix} \beta\right\|^2 \\ &= \left\|c - \begin{bmatrix} R \\ 0 \end{bmatrix} \beta\right\|^2 \\ &= \|c_1 - R\beta\|^2 + \|c_2\|^2 \end{aligned} \tag{A.4}$$

where

$$\begin{aligned} c &= (c_1^T c_2^T)^T \\ &= Q^T y \end{aligned} \tag{A.5}$$

is the rotated response vector in which case the components c_1 and c_2 are of lengths p and $n - p$ respectively. The least-squares solution $\hat{\beta}$ is easily evaluated as the solution to

$$R\hat{\beta} = c_1 \tag{A.6}$$

and the residual sum of squares is $\|c_2\|^2$.

A.2 Definitions

A.2.1 Definition of a Cholesky Decomposition

Given a symmetric positive definite square matrix $\mathbf{X}$, the Cholesky decomposition of $\mathbf{X}$ is the factorization

$$\mathbf{X} = \mathbf{U}^T\mathbf{U}, \tag{A.7}$$

where $\mathbf{U}$ is the square root of $\mathbf{X}$ and satisfies:

1. $$\mathbf{U}^T\mathbf{U} = \mathbf{X} \tag{A.8}$$

2. $\mathbf{U}$ is upper triangular (it has all $\mathbf{0}$'s below the major diagonal)

One can calculate the inverse of $\mathbf{X}$ more easily after computing $\mathbf{U}$ since

$$\mathbf{X}^{-1} = \mathbf{U}^{-1}\mathbf{U}^{T^{-1}}, \tag{A.9}$$

whereby inverses of $\mathbf{U}$ and $\mathbf{U}^T$ are easier to compute.

A.3 Data analysis

A.3.1 General Linear Modeling

```
Call:
lm(formula = WEANWT ~ YEAR. + SEX. + AGEWEAN + DL + DQ + EWE_BRD. +
    RAM_BRD., data = data4)

Residuals:
     Min       1Q   Median       3Q      Max
-7.40371 -1.32744 -0.01093  1.44031  7.70632

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.274005   1.065133   0.257  0.79706
YEAR.92     -1.565831   0.292949  -5.345 1.23e-07 ***
YEAR.93     -1.095781   0.275268  -3.981 7.60e-05 ***
YEAR.94     -2.832501   0.357504  -7.923 9.34e-15 ***
YEAR.95     -3.228367   0.343630  -9.395  < 2e-16 ***
YEAR.96     -2.351101   0.389751  -6.032 2.64e-09 ***
SEX.M        0.477910   0.169498   2.820  0.00495 **
AGEWEAN      0.070217   0.008856   7.928 8.97e-15 ***
DL           2.726355   0.315012   8.655  < 2e-16 ***
DQ          -0.268882   0.034007  -7.907 1.05e-14 ***
EWE_BRD.R   -0.585536   0.236554  -2.475  0.01355 *
RAM_BRD.R   -0.442866   0.172768  -2.563  0.01058 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.221 on 688 degrees of freedom
  (182 observations deleted due to missingness)
Multiple R-Squared: 0.3835,     Adjusted R-squared: 0.3736
F-statistic:  38.9 on 11 and 688 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: WEANWT
           Df Sum Sq Mean Sq F value    Pr(>F)
YEAR.       5 1208.1   241.6 48.9853 < 2.2e-16 ***
```

```
SEX.        1   56.0    56.0 11.3494 0.0007968 ***
AGEWEAN     1  344.2   344.2 69.7804 3.651e-16 ***
DL          1  151.5   151.5 30.7160 4.258e-08 ***
DQ          1  275.8   275.8 55.9115 2.316e-13 ***
EWE_BRD.    1   42.7    42.7  8.6548 0.0033717 **
RAM_BRD.    1   32.4    32.4  6.5708 0.0105780 *
Residuals 688 3393.7     4.9
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

A.3.2 Mixed model

```
Linear mixed-effects model fit by REML
Formula: WEANWT ~ YEAR. + SEX. + AGEWEAN + DL + DQ + EWE_BRD. + RAM_BRD. +      (1 | EWE_ID.)
   Data: data4
  AIC  BIC logLik MLdeviance REMLdeviance
 3108 3167  -1541       3053         3082
Random effects:
 Groups   Name        Variance Std.Dev.
 EWE_ID.  (Intercept) 1.4459   1.2025
 Residual             3.4968   1.8700
Number of obs: 700, groups: EWE_ID., 358

Fixed effects:
             Estimate Std. Error t value
(Intercept)  0.218577   1.025299   0.213
YEAR.92     -1.595738   0.263857  -6.048
YEAR.93     -1.090956   0.257263  -4.241
YEAR.94     -3.006466   0.339047  -8.867
YEAR.95     -3.299145   0.337383  -9.779
YEAR.96     -2.439967   0.387341  -6.299
SEX.M        0.405743   0.162486   2.497
AGEWEAN      0.065870   0.008594   7.665
DL           2.921073   0.295619   9.881
DQ          -0.290233   0.031904  -9.097
EWE_BRD.R   -0.464975   0.266548  -1.744
RAM_BRD.R   -0.420091   0.163578  -2.568

> anova(fit2,fit3)
```

```
Data: data4
Models:
fit3: WEANWT ~ YEAR. + SEX. + AGEWEAN + DL + DQ + EWE_BRD. + RAM_BRD. +
fit2:     (1 | EWE_ID.)
fit3: WEANWT ~ YEAR. + SEX. + AGEWEAN + DL + DQ + EWE_BRD. + RAM_BRD. +
fit2:     (1 | RAM_ID.) + (1 | EWE_ID.)
       Df     AIC     BIC  logLik Chisq Chi Df Pr(>Chisq)
fit3.p 13  3079.0  3138.1 -1526.5
fit2.p 14  3080.7  3144.4 -1526.4  0.24      1     0.6242
> anova(fit2,fit4)
Data: data4
Models:
fit4: WEANWT ~ YEAR. + SEX. + AGEWEAN + DL + DQ + EWE_BRD. + RAM_BRD. +
fit2:     (1 | RAM_ID.)
fit4: WEANWT ~ YEAR. + SEX. + AGEWEAN + DL + DQ + EWE_BRD. + RAM_BRD. +
fit2:     (1 | RAM_ID.) + (1 | EWE_ID.)
       Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fit4.p 13  3117.5  3176.7 -1545.8
fit2.p 14  3080.7  3144.4 -1526.4 38.814      1  4.661e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Linear mixed-effects model fit by REML
Formula: WEANWT ~ YEAR. + SEX. + AGEWEAN + DL + DQ + EWE_BRD. + RAM_BRD. +      (1 | RAM_ID.) + (1 | E
   Data: data4
  AIC  BIC logLik MLdeviance REMLdeviance
 3110 3173  -1541       3053         3082
Random effects:
 Groups   Name        Variance Std.Dev.
 EWE_ID.  (Intercept) 1.456488 1.20685
 RAM_ID.  (Intercept) 0.066577 0.25803
 Residual             3.427208 1.85127
Number of obs: 700, groups: EWE_ID., 358; RAM_ID., 74

Fixed effects:
             Estimate Std. Error t value
(Intercept)  0.185787   1.026340   0.181
YEAR.92     -1.570905   0.267785  -5.866
YEAR.93     -1.076631   0.264297  -4.074
```

YEAR.94	-3.002506	0.344568	-8.714
YEAR.95	-3.288317	0.345214	-9.525
YEAR.96	-2.450082	0.394632	-6.209
SEX.M	0.403811	0.162311	2.488
AGEWEAN	0.065929	0.008613	7.655
DL	2.922318	0.294546	9.921
DQ	-0.289973	0.031784	-9.123
EWE_BRD.R	-0.454294	0.266448	-1.705
RAM_BRD.R	-0.413038	0.175535	-2.353