Mixed Model Methodology, Part I: Linear Mixed Models

Technical Report · January 2015

DOI: 10.13140/2.1.3072.0320

1 author:

jean-louis Foulley
Université de Montpellier
313 PUBLICATIONS **4,486** CITATIONS

SEE PROFILE

Jean-Louis Foulley

Mixed Model Methodology

Jean-Louis Foulley
Institut de Mathématiques et de Modélisation de Montpellier (I3M)
Université de Montpellier 2
Sciences et Techniques
Place Eugène Bataillon
34095 Montpellier Cedex 05

Part I

Linear Mixed Model Models

Citation
Foulley, J.L. (2015). Mixed Model Methodology. Part I: Linear Mixed Models. Technical Report, e-print: DOI: 10.13140/2.1.3072.0320

1

Basics on linear models

1.1 Introduction

Linear models form one of the most widely used tools of statistics both from a theoretical and practical points of view. They encompass regression and analysis of variance and covariance, and presenting them within a unique theoretical framework helps to clarify the basic concepts and assumptions underlying all these techniques and the ones that will be used later on for mixed models.

There is a large amount of highly documented literature especially textbooks in this area from both theoretical (Searle, 1971, 1987; Rao, 1973; Rao et al., 2007) and practical points of view (Littell et al., 2002). Therefore our purpose here is limited to reviewing the main results in statistical theory concerning such models. The first chapter reviews the general formulation of linear models and the main procedures used for estimating parameters and testing hypotheses, first on the premise that data are independent and in a second step that they are correlated. This enables us to formally introduce the concept of linear mixed models. Two ways are presented to that respect either by structuring the residual components of the model or by building the model hierarchically according to several steps of sampling. Finally, we discuss the issue of the distinction

between fixed and random effects both from classical and Bayesian points of view.

1.2 Formulation of the linear model

Classically, a linear model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad (1.1)$$

where $\mathbf{y} = (y_1,..., y_N)'$ is an N-dimensional vector of dependent random variables corresponding to responses, $\mathbf{X} = (\mathbf{X}_1,...,\mathbf{X}_k,...,\mathbf{X}_p)$ is a $(N \text{x} p)$ known matrix of explanatory variables with $p \leq N$, $\boldsymbol{\beta} = (\beta_1,...,\beta_k,...,\beta_p)'$ is a $(p \text{x} 1)$ vector of real-valued parameters pertaining to each $\mathbf{X}_k$ variable, and $\mathbf{e} = (e_1,...,e_N)'$ is an N-dimensional vector of residuals representing the deviations of the data $\mathbf{y}$ from the explained part $\mathbf{X}\boldsymbol{\beta}$.

The variables $\mathbf{X}_1,...,\mathbf{X}_k,...,\mathbf{X}_p$ forming the columns of $\mathbf{X}$ can be either i) continuous such as predictor variables in regression models, or ii) discrete (often coded as 0 and 1) corresponding to levels of factors in analysis of variance (ANOVA) models. In this second case, $\mathbf{X}$ is called a design or an incidence matrix.

The corresponding elements $\beta_1,...,\beta_k,...,\beta_p$ of $\boldsymbol{\beta}$ represent in case i) regression coefficients, and in case ii) what is called “fixed effects”.

Letting $\boldsymbol{\mu} = E(\mathbf{y})$ and $\mathbf{V} = Var(\mathbf{y})$, different assumptions can be made regarding the general model defined in (1.1), that is :

a) $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ stating that the model is correctly specified regarding its systematic (expectation) part;

b) $\mathbf{V} = Diag(V_{ii})$, $i = 1,...,N$ corresponding to independent, or at least uncorrelated, data ;

c) $\mathbf{V} = \sigma^2\mathbf{I}_N$ standing for b) and residuals with homogeneous variances σ^2 ;

d) $\mathbf{y} \sim \mathcal{N}(\mathbf{X\beta}, \mathbf{V})$ stating that vector $\mathbf{y}$ follows a multivariate normal (Gaussian) distribution with expectation $\mathbf{X\beta}$ and variance covariance matrix $\mathbf{V}$ possibly described as in b) or c).

As it will be shown later on in more detail, it is not necessary to make all these assumptions simultaneously, some procedures are very constraintful such as maximum likelihood (ML) estimation or hypothesis testing, while others such as Ordinary Least Squares (OLS) only need a).

Finally, it is important to recall that linearity must be understood with respect to the parameters $\boldsymbol{\beta}$, more precisely such that the partial derivatives $\partial \boldsymbol{\mu}' / \partial \boldsymbol{\beta} = \mathbf{X}'$ of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\beta}$ does not depend on $\boldsymbol{\beta}$. Obviously, this does not preclude incorporating non linear functions of some covariates (e.g. time) in the $\mathbf{X}$ matrix.

1.3 Estimation

1.3.1 Ordinary Least Squares (OLS)

A very simple way of estimating the unknown vector of parameters $\boldsymbol{\beta}$ is by the method called "Least Squares". This method was described by Adrien-Marie Legendre in 1805 for solving overdetermined linear systems of equations in astronomy and geodesy. It was derived on probability grounds by Carl Friederich Gauss in 1809-1823 and by Pierre Simon de Laplace in 1810: see eg Stiegler (1986) about historical aspects of their contributions to this procedure.

Let $S(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X\beta}\|^2$ designate the Euclidian square distance between the data $\mathbf{y}$ and $\boldsymbol{\mu}$ considered as a function of $\boldsymbol{\beta}$. The OLS estimation $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ results from minimizing $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} S(\boldsymbol{\beta}) \qquad (1.2)$$

For this, it can be easily seen that the first and second derivatives can be expressed as :

$$\partial S(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X\beta})$$

$$\partial^2 S(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' = 2\mathbf{X}'\mathbf{X}$$

The condition of convexity being satisfied ($\mathbf{X}'\mathbf{X}$ being definite positive), the minimum of $S(\boldsymbol{\beta})$ is obtained by setting the first derivative to zero which gives

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \qquad (1.3)$$

which is known as the system of OLS or normal equations.

Two situations must be distinguished according to whether $\mathbf{X}'\mathbf{X}$ has full rank p or not. Now $rank(\mathbf{X}'\mathbf{X}) = rank(\mathbf{X})$. We ask the reader to verify this (see exercise 1.1). Thus, the condition reduces to know whether $\mathbf{X}$ is of full column rank (i.e. $p = rank(\mathbf{X})$) or not. In other words, are the columns of $\mathbf{X}$ linearly independent (LIN) or not?

If so, $\mathbf{X}'\mathbf{X}$ is invertible and the system in (1.3) has a unique solution

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \qquad (1.4)$$

This is for instance what happens in linear regression based on continuous real-valued variables provided that none of them is a linear combination of the others.

Example 1.1 *Linear regression*

Let us consider the simple regression model for the response (dependent) variable y_i as a function of a single (independent) covariate x_i measured on a sample of experimental units($i = 1,...,N$)

$$y_i = \beta_0 + \beta_1 x_i + e_i . \qquad (1.5)$$

Here

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & . \\ 1 & x_i \\ 1 & . \\ 1 & x_N \end{bmatrix} \text{ and } \mathbf{y}' = \begin{bmatrix} y_1 & . & y_i & . & y_N \end{bmatrix}$$

so that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}, \ \mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}.$$

Now

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} \sum_{i=1}^{N} x_i^2 & -\sum_{i=1}^{N} x_i \\ -\sum_{i=1}^{N} x_i & N \end{bmatrix} / D$$

with $D = N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2$.

Hence, applying (1.4) gives immediately

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right) / N}{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 / N}, \qquad (1.6)$$

$$\hat{\beta}_0 = \left[\sum_{i=1}^{N} y_i - \hat{\beta}_1 \left(\sum_{i=1}^{N} x_i\right)\right] / N. \qquad (1.7)$$

If $\mathbf{X}$ is not full column rank as in ANOVA, there are several ways for solving (1.3). The first one consists of setting an equivalent model with an incidence matrix whose columns are LIN.

1.3.1.1 Reparameterisation to full rank

This technique can be illustrated in the following example.

Example 1.2 *One-way ANOVA classification*

Let us consider the following data set from a one-way classification trial

Table 1.1. A one-way layout

A	
Level 1	Level 2
6, 6, 8	4, 12

and the corresponding ANOVA model

$$y_i = \mu + \alpha_i + e_{ij} \ , \ i = 1,...,I, j = 1,...,n_i \qquad (1.8)$$

Here $I = 2$, $n_1 = 3$ and $n_1 = 2$

$$\mathbf{y}' = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22})' = (6,6,8,4,12)$$

and,

$$\mathbf{X\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

Letting $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2)$, the model can be written as

$$E(\mathbf{y}) = \mu\mathbf{X}_0 + \alpha_1\mathbf{X}_1 + \alpha_2\mathbf{X}_2 .$$

But we have the following relationship among the columns: $\mathbf{X}_0 = \mathbf{X}_1 + \mathbf{X}_2$ so that the model can be rewritten as $E(\mathbf{y}) = \mu\mathbf{X}_0 + \alpha_1\mathbf{X}_1 + \alpha_2(\mathbf{X}_0 - \mathbf{X}_1)$ that is

$$E(\mathbf{y}) = (\mu + \alpha_2)\mathbf{X}_0 + (\alpha_1 - \alpha_2)\mathbf{X}_1 \qquad (1.9)$$

If we define $\mathbf{X}^* = (\mathbf{X}_0, \mathbf{X}_1)$, $\boldsymbol{\beta}^* = (\mu^*, \alpha^*)'$ with $\mu^* = \mu + \alpha_2$ and $\alpha^* = \alpha_2 - \alpha_1$, the model becomes $E(\mathbf{y}) = \mathbf{X}^*\boldsymbol{\beta}^*$ with $\mathbf{X}^*$ now being full column rank. It involves the first two columns of the original $\mathbf{X}$ matrix and parameters pertaining to the effect $\mu^* = \mu + \alpha_2$ of the last level of the A classification and the difference $\alpha^* = \alpha_2 - \alpha_1$ between the first and last level effects.

More generally, starting from $E(\mathbf{y}) = \mathbf{X\beta}$ with incidence matrix $\mathbf{X}_{(Nxp)}$ and $\boldsymbol{\beta}_{(px1)}$ and $r_X = rank(\mathbf{X})$ being smaller than p, we introduce an alternative equivalent model $E(\mathbf{y}) = \mathbf{X}^*\boldsymbol{\beta}^*$ where the $(N \text{x} r_X)$ $\mathbf{X}^*$ matrix being full column rank and $\boldsymbol{\beta}^*$ is a linear transformation $\boldsymbol{\beta}^* = \mathbf{T\beta}$ of full row rank of the initial parameters $\boldsymbol{\beta}$. The question is then: given $\mathbf{T}$, how can we derive $\mathbf{X}^*$? By identifying the two models, we immediately obtain

$$\mathbf{X}^* = \mathbf{XT}'(\mathbf{TT}')^{-1} \qquad (1.10)$$

We leave up to the reader to prove this result and check it on example (1.2) (see exercise 1.2)

The technique described in the example corresponds to what is called "reference population (or category) coding" which is used in some packages. An alternative would have been to define μ^* as $\mu^* = \mu + ½(\alpha_1 + \alpha_2)$ and $\alpha^* = ½(\alpha_1 - \alpha_2)$, technique that is sometimes called "deviation from the mean model".

Notice also that this reparemeterization is based on a variable transformation implying a reduction in the dimension of the parameter space from p to $r_X < p$, and thus should be distinguished from other procedures keeping this dimension unchanged but adding constraints on the parameters (Searle, 1971).

1.3.1.2 Linear constraints on solutions

For instance, in model (1.8), $E(y_i) = \mu + \alpha_i$, setting constraints on the OLS solutions of the form $\sum_{i=1}^{I} \hat{\alpha}_i = 0$, or $\hat{\alpha}_I = 0$ allows to get unique solutions to the linear system $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$.

Actually, we need to impose $m = p - r_X$ linear constraints on the elements of $\hat{\boldsymbol{\beta}}$ which can be expressed as $\mathbf{L}'\hat{\boldsymbol{\beta}} = \mathbf{0}$ with $\mathbf{L}'$ being an $(m \text{x} p)$ full row rank matrix whose rows are LIN of those of $\mathbf{X}$. Now, we have to mininimize $S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ subject to $\mathbf{L}'\boldsymbol{\beta} = \mathbf{0}$. Using a Lagrange multiplier $(m \text{x} 1)$ vector $2\boldsymbol{\lambda}$, this reduces to minimizing

$$S^*(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}'\mathbf{L}'\boldsymbol{\beta}. \qquad (1.11)$$

By differenciation with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, one has

$$\partial S^*(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\mathbf{L}\boldsymbol{\lambda}$$

$$\partial S^*(\boldsymbol{\beta}) / \partial \boldsymbol{\lambda} = 2\mathbf{L}'\boldsymbol{\beta}$$

Setting these derivatives to zero results in the following system of equations :

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{L} \\ \mathbf{L}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{0} \end{bmatrix}. \qquad (1.12)$$

Some authors (Searle, 1987, page 376) also consider restrictions on parameters. Whereas restriction on parameters imply constraints on the solutions and give

the same system of equations (1.12), the opposite is not true. In particular, making this distinction may matter as far as estimability of parameters is concerned as shown in the next section. Some parameters which are not estimable in the original model, can be estimable in the restricted one. For instance, μ is not estimable in model (1.8), but becomes such in under the restricted model $\sum_{i=1}^{I} \alpha_i = 0$.

1.3.1.3 Generalized inverses and estimable functions

When $\mathbf{X}$ is not full column rank, the OLS system $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ has many solutions, in fact, an infinite number of solutions which can be obtained using generalized inverses of $\mathbf{X}'\mathbf{X}$, denoted by $(\mathbf{X}'\mathbf{X})^-$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{y}. \qquad (1.13)$$

Definition 1.1 *If* $\mathbf{G}$ *stands for* $(\mathbf{X}'\mathbf{X})^-$*, a g-inverse of* $\mathbf{X}'\mathbf{X}$ *is any matrix satisfying* $\mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$.

Since there are many solutions generated by g-inverses of $\mathbf{X}'\mathbf{X}$, a natural idea is to restrict solutions to those which are invariant to the choice of g-inverses. For this, an interesting class of such functions reposes on the so-called "estimable functions". The basic idea underlying this concept is as follows. As the key parameter of linear models are the expectations $\mathbf{X}\boldsymbol{\beta}$ of data (e.g the cell mean μ_{ij} in a two-way classification model), only linear functions of $\mathbf{X}\boldsymbol{\beta}$ have to be considered in the estimation process of $\boldsymbol{\beta}$.

Definition 1.2 *In that context, a linear function* $\mathbf{k}'\boldsymbol{\beta}$ *is an estimable function, if and only if (iff in short), it can be expressed as a linear combination of the expectations of the observations.*

$$\mathbf{k}'\boldsymbol{\beta} \text{ estimable iff } \exists \mathbf{t},\ \mathbf{k}'\boldsymbol{\beta} = \mathbf{t}'\mathbf{X}\boldsymbol{\beta},\ \forall \boldsymbol{\beta} \qquad (1.14)$$

Now, using any g-inverse of $\mathbf{X}'\mathbf{X}$, the OLS estimation $\mathbf{k}'\hat{\boldsymbol{\beta}}$ of $\mathbf{k}'\boldsymbol{\beta}$ can be calculated from

$$\mathbf{k}'\hat{\boldsymbol{\beta}} = \mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{y} \qquad (1.15)$$

Since $\mathbf{k}'\boldsymbol{\beta}$ is an estimable function, $\mathbf{k}' = \mathbf{t}'\mathbf{X}$ and $\mathbf{k}'\hat{\boldsymbol{\beta}} = \mathbf{t}'\mathbf{X}\hat{\boldsymbol{\beta}}$.

Letting

$$\mathbf{P} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}' \qquad (1.16)$$

Substituting $\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{y}$ to $\hat{\boldsymbol{\beta}}$ in $\mathbf{k}'\hat{\boldsymbol{\beta}}$ and using (1.16), $\mathbf{k}'\hat{\boldsymbol{\beta}}$ can be expressed as

$$\mathbf{k}'\hat{\boldsymbol{\beta}} = \mathbf{t}'\mathbf{P}\mathbf{y} \qquad (1.17)$$

The $\mathbf{P}$ (also denoted $\mathbf{H}$ in some textbooks) plays a major role in linear model theory owing its geometric interpretation ($\mathbf{P}$ is a projector as shown later on) and its properties, the main ones being (see appendix):

$$\mathbf{P}\mathbf{X} = \mathbf{X} \qquad (1.18)$$

$$\mathbf{P} \text{ invariant to } \left(\mathbf{X}'\mathbf{X}\right)^{-} \qquad (1.19)$$

$$\mathbf{P} = \mathbf{P}^2 \qquad (1.20)$$

Knowing (1.17), property (1.19) proves that $\mathbf{k}'\hat{\boldsymbol{\beta}}$ is also invariant to $\left(\mathbf{X}'\mathbf{X}\right)^{-}$.

Formulae (1.13) and (1.15) rely on choosing g-inverses $\mathbf{G}$ of $\mathbf{X}'\mathbf{X}$, and a natural question immediately arises on how to calculate such $\mathbf{G}$'s ?

Without entering into technicalities which are beyond our purpose, here are three different ways on how to obtain such matrices in the particular case of $\mathbf{X}'\mathbf{X}$.

The first approach consists in partitioning columns of $\mathbf{X}$ into two sets $\left(\mathbf{X}_1, \mathbf{X}_2\right)$ such that $\mathbf{X}_1$, an $\left(N \text{x} r_X\right)$ matrix is full column rank with $r_X = rank(\mathbf{X})$ and columns of $\mathbf{X}_2$ are linear combinations of those of $\mathbf{X}_1$, that is $\mathbf{X}_2$ can be written as $\mathbf{X}_2 = \mathbf{X}_1\mathbf{T}$ for some known $\mathbf{T}$. Hence, $\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}$ exists and $\mathbf{G}$ can be computed as

$$\mathbf{G}_1 = \begin{bmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad (1.21)$$

A second approach begins by constructing the same matrix as the coefficient matrix of (1.13) with $\mathbf{L}$ being an $(m \times p)$ full row rank matrix such that $m = p - r_{\mathbf{x}}$, and computing its regular inverse

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{L} \\ \mathbf{L}' & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M}^{11} & \mathbf{M}^{12} \\ \mathbf{M}^{21} & \mathbf{M}^{22} \end{bmatrix}. \qquad (1.22)$$

Then, it can be shown (Searle, 1971, pages 21-23) that

$$\mathbf{G}_2 = \mathbf{M}^{11}, \qquad (1.23)$$

and

$$\mathbf{G}_3 = \left(\mathbf{X}'\mathbf{X} + \mathbf{L}\mathbf{L}'\right)^{-1}. \qquad (1.24)$$

are both g-inverses of $\mathbf{X}'\mathbf{X}$. In fact, $\mathbf{G}_2$ is a symmetric reflexive g-inverse (i.e. verifying $\mathbf{AGA} = \mathbf{A}$ and $\mathbf{GAG} = \mathbf{G}$) whereas $\mathbf{G}_3$ is not (see exercise 1.6)

1.3.2 Statistical properties

Under the assumption that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ (i.e that the model is correctly specified for its expectation), the OLS estimator of an estimable function is unbiased

$$E\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right) = \mathbf{k}'\boldsymbol{\beta}. \qquad (1.25)$$

The proof of (1.25) is very easy and worth mentioning. $\mathbf{k}'\boldsymbol{\beta}$ being an estimable function, we know that its OLS estimator can be expressed as $\mathbf{k}'\hat{\boldsymbol{\beta}} = \mathbf{t}'\mathbf{P}\mathbf{y}$ so that , $E\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right) = \mathbf{t}'\mathbf{P}\mathbf{X}\boldsymbol{\beta}$ but, from to (1.18), one has $\mathbf{t}'\mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{t}'\mathbf{X}\boldsymbol{\beta} = \mathbf{k}'\boldsymbol{\beta}$ (QED).

It is important to point out that this proof only relies on the assumption that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, and does not involve any statement about the variance covariance structure $\mathbf{V}$ of the residuals. In addition, since $E\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right) = \mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$, this unbiasedness property provides a simple device for testing estimability of $\mathbf{k}'\boldsymbol{\beta}$ by checking the following relationship:

$$\mathbf{k}' = \mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{X}, \qquad (1.26)$$

for any g-inverse of $\mathbf{X}'\mathbf{X}$.

Moreover, under the additional assumption that $\mathbf{V}=\sigma^2\mathbf{I}_N$ (independence and homoscedasticity of residuals), the OLS estimator of $\mathbf{k}'\boldsymbol{\beta}$ is also the best linear unbiased estimator (BLUE) of $\mathbf{k}'\boldsymbol{\beta}$, that is the one with the smallest sampling variance in the class of all linear unbiased estimators of $\mathbf{k}'\boldsymbol{\beta}$ (exercise 1.4)

$$BLUE\left(\mathbf{k}'\boldsymbol{\beta}\right)=\mathbf{k}'\hat{\boldsymbol{\beta}}, \qquad (1.27)$$

with

$$Var\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right)=\sigma^2\mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{k}. \qquad (1.28)$$

Again, this last expression comes from writing $\mathbf{k}'\hat{\boldsymbol{\beta}}=\mathbf{t}'\mathbf{P}\mathbf{y}$ so that $Var\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right)=\sigma^2\mathbf{t}'\mathbf{P}^2\mathbf{t}$. Now, given that $\mathbf{P}$ is idempotent, and replacing it by its expression (1.16), one has $Var\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right)=\sigma^2\mathbf{t}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{t}$. Then, noting that $\mathbf{t}'\mathbf{X}=\mathbf{k}'$ leads to (1.28). We have seen that $\mathbf{k}'\hat{\boldsymbol{\beta}}$ is a linear, unbiased estimator of $\mathbf{k}'\boldsymbol{\beta}$. We leave up to the reader (exercise) to show that its variance $\sigma^2\mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{k}$ is the smallest possible one.

Finally, assuming that $\mathbf{y}\sim\mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_N\right)$, the OLS estimator of $\mathbf{k}'\boldsymbol{\beta}$ is also normal with expectation and variance given in (1.25) and (1.28) respectively

$$\mathbf{k}'\hat{\boldsymbol{\beta}}\sim\mathcal{N}\left(\mathbf{k}'\boldsymbol{\beta},\sigma^2\mathbf{k}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{k}\right). \qquad (1.29)$$

The results in (1.28) and (1.29) are based on the premise that σ^2 is known. Otherwise, it must be estimated. A simple way to do it, is to refer to the residual sum of squares, denoted classically by SSE

$$SSE=\left\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2. \qquad (1.30)$$

Given $\hat{\mathbf{e}}=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}$ or alternatively $\hat{\mathbf{e}}=\left(\mathbf{I}-\mathbf{P}\right)\mathbf{y}$, SSE can be viewed as a quadratic form $SSE=\mathbf{y}'\left(\mathbf{I}-\mathbf{P}\right)\mathbf{y}$ knowing that $\left(\mathbf{I}-\mathbf{P}\right)$ is idempotent.

This means that it can be easily computed as $SSE = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{P}\mathbf{y}$. But, remember, the data vector can be decomposed as

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}} = \mathbf{P}\mathbf{y} + (\mathbf{I} - \mathbf{P})\mathbf{y} \qquad (1.31)$$

so that the last term $\mathbf{y}'\mathbf{P}\mathbf{y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ and

$$SSE = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} . \qquad (1.32)$$

The last term made of the product of the OLS solution $\hat{\boldsymbol{\beta}}$ by the right hand side of the normal equations $\mathbf{X}'\mathbf{y}$ represents the part of variation of $\mathbf{y}$ accounted for by the model $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. It is often denoted as $R(\boldsymbol{\beta})$ or SSR for « reduction in sums of squares ». Conversely, the total variation in $\mathbf{y}$ can be partitioned into two components, one $R(\boldsymbol{\beta})$ explained by the fitted model, and one due to the unexplained or residual part, SSE

$$\mathbf{y}'\mathbf{y} = R(\boldsymbol{\beta}) + SSE$$

which is the basic principle of the ANOVA procedure.

Since SSE is a quadratic form, its expectation can be easily obtained using the classical result

$$E(\mathbf{y}'\mathbf{Q}\mathbf{y}) = \boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu} + tr(\mathbf{Q}\mathbf{V}) . \qquad (1.33)$$

Here, $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{V} = \sigma^2\mathbf{I}_N$ and $tr(\mathbf{I} - \mathbf{P}) = N - r_X$. The first term $\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P})\mathbf{X}\boldsymbol{\beta}$ cancels out because $\mathbf{P}\mathbf{X} = \mathbf{X}$, and $E(SSE)$ reduces to

$$E(SSE) = \sigma^2 (N - r_X)$$

Therefore, we get an unbiased estimator of σ^2 as

$$\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{y} - R(\boldsymbol{\beta})}{N - r_X} . \qquad (1.34)$$

Furthermore, under the assumption of normality of $\mathbf{y}$, this estimator has a distribution proportional to a chi-square, namely

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{N - r_X} \chi^2_{N - r_X} . \qquad (1.35)$$

The proof of (1.35) is a direct application of the following theorem

Theorem 1.1: If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, then the quadratic $\mathbf{y}'\mathbf{Q}\mathbf{y}$ has a non central chi-square distribution with degrees of freedom $\nu = rank(\mathbf{QV})$ and non centrality parameter $\lambda = ½\boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu}$.

As seen previously in deriving (1.34), here $\mathbf{Q} = (\mathbf{I} - \mathbf{P})/\sigma^2$, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{V} = \sigma^2\mathbf{I}_N$ and $tr(\mathbf{I} - \mathbf{P}) = N - r_X$ so that $\nu = N - r_X$, $\lambda = 0$ and $SSE/\sigma^2 \sim \chi^2_{N-r_X}$ (QED).

Another possibility under the normality assumption of data $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$ would be to recourse to maximum likelihood (ML) estimation of both $\boldsymbol{\beta}$ and σ^2. In that case, letting $f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$ designate the density of $\mathbf{y}$, the log-likelihood $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \log f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$ considered as a function of the parameters is

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -½N(\log 2\pi + \log\sigma^2) - ½(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2 \quad (1.36)$$

Letting $D = -2L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$, differentiation with respect to $\boldsymbol{\beta}$ and σ^2 yields

$$\partial D/\partial\boldsymbol{\beta} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2,$$

$$\partial D/\partial\boldsymbol{\beta} = N/\sigma^2 - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^4,$$

and setting them to zero, leads to

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}, \quad (1.37)$$

$$\hat{\sigma}^2_{ML} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/N. \quad (1.38)$$

The system in $\hat{\boldsymbol{\beta}}$ is exactly the same as the OLS normal equations whereas the estimation of σ^2 differs from the moment estimation in (1.34) with N instead of $N - r_X$ in the denominator making the ML estimator biased downwards.

Example 1.3 *(Continuation of example 1.2)*

We can apply LS theory to the one-way classification data of example 1.2.

For $i = 1$, the data are $y_{11} = 6$, $y_{12} = 6$, $y_{13} = 8$ and for $i = 2$, $y_{21} = 4$, $y_{22} = 12$ so that $\mathbf{y}' = (6,6,8,4,12)$.

The data are fitted to the following model: $y_i = \mu + \alpha_i + e_{ij}$, $i = 1,2$.

Thus, the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ are:

$$\begin{bmatrix} 5 & 3 & 2 \\ 3 & 3 & 0 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \alpha_1 \\ \hat{\alpha}_2 \end{bmatrix} = \begin{bmatrix} 36 \\ 20 \\ 16 \end{bmatrix}.$$

Reparameterization in $\mu^* = \mu + \alpha_2$ and $\alpha^* = \alpha_1 - \alpha_2$ gives the following system

$$\begin{bmatrix} 5 & 3 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} \hat{\mu}^* \\ \hat{\alpha}^* \end{bmatrix} = \begin{bmatrix} 36 \\ 20 \end{bmatrix},$$

with solutions $\hat{\mu}^* = 8$ and $\hat{\alpha}^* = -4/3$.

When setting the constraint $\hat{\alpha}_2 = 0$ i.e $\mathbf{L}' = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$ with a Lagrange multiplier λ, the system becomes

$$\begin{bmatrix} 5 & 3 & 2 & 0 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} 36 \\ 20 \\ 16 \\ 0 \end{bmatrix}$$

which gives $\hat{\mu} = 8$, $\hat{\alpha}_1 = -4/3$, $\hat{\alpha}_2 = 0$ and $\lambda = 0$.

Notice that the inverse of the coefficient matrix provides exactly the g-inverse $\mathbf{G}_2$ defined in (1.23)

$$\mathbf{G}_2 = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 5/6 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This matrix verifies both $\mathbf{A}\mathbf{G}_2\mathbf{A} = \mathbf{A}$ and $\mathbf{G}_2\mathbf{A}\mathbf{G}_2 = \mathbf{G}_2$ where $\mathbf{A}$ is the coefficient matrix $\mathbf{X}'\mathbf{X}$ of the first system. We can also check that the estimability condition $\mathbf{K}' = \mathbf{K}'\mathbf{G}_2\mathbf{X}'\mathbf{X}$ for $\mu + \alpha_2$ and $\alpha_2 - \alpha_1$ holds using $\mathbf{X}'\mathbf{X}$, $\mathbf{G}_2$ and $\mathbf{K}' = \begin{bmatrix} 1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$.

We can also form $\mathbf{X}'\mathbf{X} + \mathbf{L}\mathbf{L}'$ which is

$$\mathbf{X}'\mathbf{X}+\mathbf{L}\mathbf{L}'=\begin{bmatrix}5 & 3 & 2\\ 3 & 3 & 0\\ 2 & 0 & 3\end{bmatrix}$$

and take its regular inverse $\mathbf{G}_3=\left(\mathbf{X}'\mathbf{X}+\mathbf{L}\mathbf{L}'\right)^{-1}$

$$\mathbf{G}_3=\begin{bmatrix}3/2 & -3/2 & -1\\ -3/2 & 11/6 & 1\\ -1 & 1 & 1\end{bmatrix}.$$

As mentioned in (1.24), $\mathbf{G}_3$ is a g-inverse of $\mathbf{A}=\mathbf{X}'\mathbf{X}$ that verifies $\mathbf{A}\mathbf{G}_3\mathbf{A}=\mathbf{A}$ but not $\mathbf{G}_3\mathbf{A}\mathbf{G}_3=\mathbf{G}_3$. However, both $\mathbf{G}_2$ and $\mathbf{G}_3$ provide the same solution to $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{y}$ which is $\hat{\boldsymbol{\beta}}'=\begin{pmatrix}22 & -2 & 2\end{pmatrix}/3$.

Whatever the technique used, the estimable functions have the same values $\hat{\mu}^*=8$ and $\hat{\alpha}^*=-4/3$. The reduction in sums of squares $R(\boldsymbol{\beta})$ and SSE remain invariant and equal to $R(\boldsymbol{\beta})=784/3$ and $SSE=296-784/3=104/3$ giving $\hat{\sigma}^2=SSE/3=104/9$.

1.4 Hypothesis testing

1.4.1 General results

A convenient formulation for testing a linear hypothesis about $\boldsymbol{\beta}$ consists in writing

$$H_0:\mathbf{C}'\boldsymbol{\beta}=\mathbf{0} \text{ vs } H_1=\bar{H}_0:\mathbf{C}'\boldsymbol{\beta}\neq\mathbf{0}, \qquad (1.39)$$

where $\mathbf{C}'$ is a full row rank $\left(r_C \text{ x } p\right)$known matrix of coefficients such that $1\le r_C\le r_X$ and $\mathbf{C}'\boldsymbol{\beta}$ represent r_C estimable functions.

Usual test statistics about (1.30) require the normality asumption $\mathbf{y}\sim\mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_N\right)$. Then according to (1.29) we know that

$$\mathbf{C}'\hat{\boldsymbol{\beta}}-\mathbf{C}'\boldsymbol{\beta}\sim\mathcal{N}\left[0,\sigma^2\mathbf{C}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{C}\right]. \qquad (1.40)$$

Moreover, if $\mathbf{z}_{(nx1)}\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}\right)$, then $\mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z}\sim\chi^2_n$; this is in fact a corollary of theorem 1.1. Applying this result to (1.40) under $H_0:\mathbf{C}'\boldsymbol{\beta}=\mathbf{0}$ gives

$$W = \hat{\boldsymbol{\beta}}'\mathbf{C}\left[\mathbf{C}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{C}\right]^{-1}\mathbf{C}'\hat{\boldsymbol{\beta}}/\sigma^2 \sim \chi^2_{r_c}. \qquad (1.41)$$

However, we cannot take advantage directly of this property since W entails the unknown parameter σ^2 in the denominator. Letting $Q = \sigma^2 W$, this difficulty is overcomed by observing that Q and SSE are both proportional to chi-squares but with the same proportionality constant

$$Q \sim \sigma^2 \chi^2_{r_C}. \qquad (1.42)$$

$$SSE \sim \sigma^2 \chi^2_{N-r_X} \qquad (1.43)$$

In addition, Q and SSE are statistically independent. This property results directly from the following theorem (see exercise)

Theorem 1.2. Let $\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{V}\right)$, the quadratic forms $\mathbf{y}'\mathbf{A}\mathbf{y}$ and $\mathbf{y}'\mathbf{B}\mathbf{y}$ are independently distributed if and only if $\mathbf{AVB} = \mathbf{0}$, or $\mathbf{BVA} = \mathbf{0}$.

Here, for $\mathbf{G} = \left(\mathbf{X}'\mathbf{X}\right)^{-}$, $\mathbf{A} = \mathbf{I} - \mathbf{P}$ and $\mathbf{B} = \mathbf{XGC}\left(\mathbf{CGC}'\right)^{-1}\mathbf{C}'\mathbf{GX}'$, $\mathbf{V} = \sigma^2\mathbf{I}$. Since $\left(\mathbf{I} - \mathbf{P}\right)\mathbf{X} = \mathbf{0}$, then $\mathbf{AVB} = \mathbf{0}$ (QED).

We can now construct the ratio of two independent chi-square variables divided by their respective degrees of freedom. Let

$$F = \frac{Q/\sigma^2 r_C}{SSE/\sigma^2\left(N - r_X\right)}$$

Doing so, σ^2 cancels out and F becomes a computable statistic which can be alternatively expressed as

$$F = \frac{Q/r_C}{SSE/\left(N - r_X\right)}, \qquad (1.44)$$

or

$$F = \frac{Q}{r_C\hat{\sigma}^2}, \qquad (1.45)$$

where $\hat{\sigma}^2$ is the unbiased moment estimator given in (1.34).

Under the null hypothesis, this quantity has a Fisher-Snedecor distribution with r_C and $N - r_X$ degrees of freedom.

$$F \sim_{H_0} \mathcal{F}_{(r_C, N-r_X)} \tag{1.46}$$

and thus provides a statistic for testing H_0 versus H_1.

The case $r_C = 1$ deserves special attention. It reduces to implementing a t-test since a Fisher-Snedecor distribution with 1 and n degrees of freedom is equivalent to the square of a Student's-t distribution with n degrees of freedom. Therefore, testing $H_0 : \mathbf{c}'\boldsymbol{\beta} = \mathbf{0}$ vs $H_1 = \bar{H}_0 : \mathbf{c}'\boldsymbol{\beta} \neq \mathbf{0}$ can be carried out via

$$\frac{\mathbf{c}'\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{\left[\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{c}\right]^{-1}}} \sim T_{N-r_X}. \tag{1.47}$$

The corresponding $100(1-\alpha)\%$ confidence interval of $\mathbf{c}'\hat{\boldsymbol{\beta}}$ is obtained as

$$\mathbf{c}'\hat{\boldsymbol{\beta}} \pm t_{(N-r_X, \alpha/2)} SE, \tag{1.48}$$

where $t_{(N-r_X, \alpha/2)}$ is the $1-\alpha/2$ quantile of T_{N-r_X} and $SE = \hat{\sigma}\sqrt{\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{c}}$ is the standard-error of $\mathbf{c}'\hat{\boldsymbol{\beta}}$.

1.4.2 Alternative computations of Q

Following (1.41) and (1.42), the quantity Q entering the numerator of the F-statistic can be calculated directly as

$$Q = \hat{\boldsymbol{\beta}}'\mathbf{C}\left[\mathbf{C}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{C}\right]^{-1}\mathbf{C}'\hat{\boldsymbol{\beta}}. \tag{1.49}$$

But, there are some shortcut procedures allowing both to simplify the computations and also to interpret Q in terms of a difference between the reduction in sums of squares due to fitting two different models, a complete and a reduced one.

Let us consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with the following partitions of $\mathbf{X} = \left(\mathbf{X}_1, \mathbf{X}_2\right)$ and of $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)$ and assume that we wish to test the null

hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ against its contrary alternative. To this respect, we can apply the general procedure presented in (1.39) and (1.40) with $\mathbf{C}' = \left(\mathbf{0}_{p_1}, \mathbf{I}_{p_2}\right)$, $\mathbf{X}_2$ being full column rank by construction of H_0.

Then $Q = \sigma^{-2}\hat{\boldsymbol{\beta}}_2' \left[Var\left(\hat{\boldsymbol{\beta}}_2\right) \right]^{-1} \hat{\boldsymbol{\beta}}_2$ where $\hat{\boldsymbol{\beta}}_2$ is the solution to the OLS system

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}. \qquad (1.50)$$

After eliminating the equations for $\hat{\boldsymbol{\beta}}_1$, this system reduces to

$$\mathbf{X}_2' \left(\mathbf{I} - \mathbf{P}_{X_1}\right) \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2' \left(\mathbf{I} - \mathbf{P}_{X_1}\right) \mathbf{y}, \qquad (1.51)$$

where $\mathbf{P}_{X_1}$ is the projector defined as $\mathbf{P}_{X_1} = \mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-} \mathbf{X}_1'$.

Expliciting $\hat{\boldsymbol{\beta}}_2$ and $Var\left(\hat{\boldsymbol{\beta}}_2\right)$ from (1.51) and inserting these expressions in that of Q, one obtains

$$Q = \hat{\boldsymbol{\beta}}_2' \mathbf{X}_2' \left(\mathbf{I} - \mathbf{P}_{X_1}\right) \mathbf{y} \qquad (1.52)$$

which appears as the product of the solution of the system (1.51) by its right-hand side.

In addition, the reduction $R\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$ due to fitting the complete (or full) model is by definition

$$R\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right) = \hat{\boldsymbol{\beta}}_1' \mathbf{X}_1' \mathbf{y} + \hat{\boldsymbol{\beta}}_2' \mathbf{X}_2' \mathbf{y}$$

with, from (1.50), $\hat{\boldsymbol{\beta}}_1$ satisfying $\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{y}$ or, alternatively,

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-} \mathbf{X}_1' \left(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2\right).$$

Substituting that expression into that of $R\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$, gives

$$R\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right) = \hat{\boldsymbol{\beta}}_2' \mathbf{X}_2' \mathbf{y} + \left(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2\right)' \mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-} \mathbf{X}_1'\mathbf{y}$$

and, after rearranging,

$$R(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=\hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\left(\mathbf{I}-\mathbf{P}_{X_1}\right)\mathbf{y}+\mathbf{y}'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-}\mathbf{X}_1'\mathbf{y}.$$

In other words, $R(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=Q+R(\boldsymbol{\beta}_1)$. This means that Q, also denoted as $R(\boldsymbol{\beta}_2\,|\,\boldsymbol{\beta}_1)$, can be expressed as the difference between the reductions due to fitting the full model and the reduced model $\mathbf{y}=\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{e}$

$$R(\boldsymbol{\beta}_2\,|\,\boldsymbol{\beta}_1)=R(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)-R(\boldsymbol{\beta}_1). \qquad (1.53)$$

We have seen previously that Q can be written as a quadratic form in $\hat{\boldsymbol{\beta}}_2$, that is

$$R(\boldsymbol{\beta}_2\,|\,\boldsymbol{\beta}_1)=\hat{\boldsymbol{\beta}}_2'\left[\mathbf{X}_2'\left(\mathbf{I}-\mathbf{P}_{X_1}\right)\mathbf{X}_2\right]\hat{\boldsymbol{\beta}}_2.$$

Assuming in addition that $\mathbf{X}_1$ is also full column rank, let

$$\begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix}=\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}. \qquad (1.54)$$

Classical results on inverses of partitioned matrices give

$$\mathbf{C}_{22}^{-1}=\mathbf{X}_2'\left(\mathbf{I}-\mathbf{P}_{X_1}\right)\mathbf{X}_2,$$

so that

$$R(\boldsymbol{\beta}_2\,|\,\boldsymbol{\beta}_1)=\hat{\boldsymbol{\beta}}_2'\mathbf{C}_{22}^{-1}\hat{\boldsymbol{\beta}}_2. \qquad (1.55)$$

This expression is known as “the inverse part of the inverse” (Searle, 1971, page 115; Harvey, 1975, page 7; Searle, 1987, page 268). It may also be applied to the non full rank case provided $\mathbf{C}_{22}$ in (1.54) is taken as a symmetric reflexive g-inverse of $\mathbf{X}_2'\left(\mathbf{I}-\mathbf{P}_{X_1}\right)\mathbf{X}_2$. Then, $R(\boldsymbol{\beta}_2\,|\,\boldsymbol{\beta}_1)=\hat{\boldsymbol{\beta}}_2'\mathbf{C}_{22}^{-}\hat{\boldsymbol{\beta}}_2$.

Finally, we end up with two ways of computing Q when testing $H_0:\boldsymbol{\beta}_2=\mathbf{0}$ either directly as $\hat{\boldsymbol{\beta}}_2'\mathbf{C}_{22}^{-1}\hat{\boldsymbol{\beta}}_2$, or indirectly, from the difference $R(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)-R(\boldsymbol{\beta}_1)$.

This last procedure known after Yates (1934) as the method of fitting constants has played a role in analysis of variance for unbalanced data both for fixed (as here) or mixed (Henderson, 1953) mixed models. It also has the advantage of being very general since it can be extended to any linear testing hypothesis such as $H_0:\mathbf{C}'\boldsymbol{\beta}=\mathbf{m}$ where $\mathbf{m}$ is a $(r_C\,\mathrm{x}\,1)$ known vector, not necessarily nil.

Let $\hat{\boldsymbol{\beta}}_0$ the estimation of $\boldsymbol{\beta}$ under the null hypothesis and $\hat{\boldsymbol{\beta}}$ under the alternative, $\hat{\boldsymbol{\beta}}_0$ can be calculated from the following system

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{C} \\ \mathbf{C}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{m} \end{bmatrix}, \qquad (1.56)$$

and $\hat{\boldsymbol{\beta}}$ from $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$.

Then Q can be simply expressed as

$$Q = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}_0'\mathbf{X}'\mathbf{y} - \boldsymbol{\lambda}'\mathbf{m}, \qquad (1.57)$$

which is equivalent to

$$Q = SSE(Restricted) - SSE(Full), \qquad (1.58)$$

with

$$SSE(Restricted) = \mathbf{y}'\mathbf{y} - \left(\hat{\boldsymbol{\beta}}_0'\mathbf{X}'\mathbf{y} + \boldsymbol{\lambda}'\mathbf{m}\right), \qquad (1.59)$$

$$SSE(Full) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}. \qquad (1.60)$$

We leave up to the reader (see exercise 1.7) to show that the expression of *SSE* under the restricted model is (1.59).

Example 1.4 *Two way crossclassified model*

Let us consider they following two-way cross-classified layout involving gains of barrows sired by 3 different boars and assigned to two feeding regimes (Harvey, 1975, page 42) as shown in table 1.2

Table 1.2: *Data set for the 2 way crossclassified design*

Cell	Regime	Sire	Number	Observations
1	1	1	2	5,6
2	1	2	5	2,3,5,6,7
3	1	3	1	3
4	2	1	2	2,3
5	2	2	3	8,8,9
6	2	3	5	4,4,6,6,7

We suggest analyzing this data set by a two factor linear model with interaction

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \qquad (1.61)$$

where μ is a mean parameter, α_i is the effect due to the ith level of factor A (here regime, $i = 1,2$), β_j is the effect due to the jth level of factor B (here sire, $j = 1,2,3$), γ_{ij} is the interaction effect pertaining to the ijth subclass, and e_{ijk} is the residual for the kth observation $k = 1,...,n_{ij}$ assumed i.i.d. $\mathcal{N}\left(0,\sigma^2\right)$.

Letting $\mu_{ij} = E\left(y_{ij}\right)$, we can reparameterize this model so as to make it full rank with

$$\mu^* = \mu_{23},$$
$$\alpha_1^* = \mu_{13} - \mu_{23},$$
$$\beta_1^* = \mu_{21} - \mu_{23},\ \beta_2^* = \mu_{22} - \mu_{23},$$
$$\gamma_{11}^* = \left(\mu_{11} - \mu_{21}\right) - \left(\mu_{13} - \mu_{23}\right)$$
$$\gamma_{12}^* = \left(\mu_{12} - \mu_{22}\right) - \left(\mu_{13} - \mu_{23}\right).$$

Given that the levels of reference are the last ones i.e. 2 for A and 3 for B, this reparemeterization provides "natural" measures of the main and interaction effects. This is especially true for the interaction effects. Remember that an interaction between two factors (here A and B) ensues from the fact that the effects of the levels of A depends on the modalities of B. For instance, γ_{11}^*, the interaction between level 1 of A and level 1 of B, measures the extent to which the difference between rows 1 and 2 (as reference) varies from column 1 to colum 3 (as reference).

Letting $\mathbf{y} = \left(y_{ijk}\right)$ and $\mathbf{e} = \left(e_{ijk}\right)$ be the vectors of data and residuals respectively, then, the model can be written in matrix notations as

$$\mathbf{y} = \mathbf{1}_N \mu^* + \mathbf{X}_A^* \boldsymbol{\alpha}^* + \mathbf{X}_B^* \boldsymbol{\beta}^* + \mathbf{X}_C^* \boldsymbol{\gamma}^* + \mathbf{e}, \qquad (1.62)$$

where $\boldsymbol{\alpha}^* = \left(\alpha_1^*\right)$, $\boldsymbol{\beta}^* = \left(\beta_1^*, \beta_2^*\right)'$, $\boldsymbol{\gamma}^* = \left(\gamma_{11}^*, \gamma_{12}^*\right)'$; $\mathbf{1}_N$ is an $(N\text{x}1)$ vector of 1's and $\mathbf{e} = \left(e_{ijk}\right)$. $\mathbf{X}_A^*$, $\mathbf{X}_B^*$, $\mathbf{X}_C^*$ are incidence matrices shown below:

$$\mathbf{X}_A^* = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}'$$

$$\mathbf{X}_B^* = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{X}_C^* = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

Notice that in this parameterization $\mathbf{X}_A^*$ ($\mathbf{X}_B^*$ respectively) results from dropping the column pertaining to the last level from the full incidence matrice of factor A (B respectively). In addition, the 2 columns of $\mathbf{X}_C^*$ are simply obtained as the product of the single column of $\mathbf{X}_A^*$ by the two columns of $\mathbf{X}_B^*$ respectively.

Let $\mathbf{X}^* = \left(\mathbf{1}_N, \mathbf{X}_A^*, \mathbf{X}_B^*, \mathbf{X}_C^*\right)$ and $\hat{\mathbf{b}}^* = \left(\mu^*, \boldsymbol{\alpha}^{*'}, \boldsymbol{\beta}^{*'}, \boldsymbol{\gamma}^{*'}\right)'$, the system of OLS equations becomes after reparameterization:

$$\mathbf{X}^{*'}\mathbf{X}^*\hat{\mathbf{b}}^* = \mathbf{X}^{*'}\mathbf{y} \qquad (1.63)$$

that is

$$\begin{bmatrix} 18 & 8 & 4 & 8 & 2 & 5 \\ 8 & 8 & 2 & 5 & 2 & 5 \\ 4 & 2 & 4 & 0 & 2 & 0 \\ 8 & 5 & 0 & 8 & 0 & 5 \\ 2 & 2 & 2 & 0 & 2 & 0 \\ 5 & 5 & 0 & 5 & 0 & 5 \end{bmatrix} \begin{bmatrix} \hat{\mu}^* \\ \hat{\alpha}_1^* \\ \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \hat{c}_{11}^* \\ \hat{c}_{12}^* \end{bmatrix} = \begin{bmatrix} 94 \\ 37 \\ 16 \\ 48 \\ 11 \\ 23 \end{bmatrix}$$

This system being full rank, has unique solutions that are $\hat{\mu}^* = 5.4$, $\hat{\alpha}_1^* = -2.4$, $\hat{\beta}_1^* = -2.9$, $\hat{\beta}_2^* = 2\,{}^{14}\!/_{15}$, $\hat{c}_{11}^* = 5.4$ and $\hat{c}_{12}^* = -1\,{}^{1}\!/_{3}$.

Now using this solution and the right hand side, we can compute the R term for the full model

$$R\left(\mu, \alpha, \beta, \gamma\right) = R\left(\mathbf{b}^*\right) = \hat{\mathbf{b}}^*\mathbf{X}^{*'}\mathbf{y} = 541\,{}^{14}\!/_{15}. \qquad (1.64)$$

An alternative would have been to calculate it from its ANOVA expression in terms of a sum of squares

$$R(\mu,\alpha,\beta,\gamma)=\sum_{i=1}^{I}\sum_{j=1}^{J}y_{ij0}^2/n_{ij}$$

where $y_{ij0}=\sum_{k=1}^{n_{ij}}y_{ijk}$. Here, this gives

$$R(\mu,\alpha,\beta,\gamma)=\frac{11^2}{2}+\frac{23^2}{5}+\frac{3^2}{1}+\frac{5^2}{2}+\frac{25^2}{3}+\frac{27^2}{5}=541\tfrac{14}{15}.$$

Let $\mathbf{X}^{\#}=\left(\mathbf{1}_N,\mathbf{X}_A^{\#},\mathbf{X}_B^{\#}\right)$ and $\mathbf{b}^{\#}=\left(\mu^{\#},\alpha_1^{\#},\beta_1^{\#},\beta_2^{\#}\right)'$, similar computations can be carried out for the additive model $E(\mathbf{y})=\mathbf{X}^{\#}\mathbf{b}^{\#}$ where $\mathbf{X}_A^{\#}=\mathbf{X}_A^{*}$ and $\mathbf{X}_B^{\#}=\mathbf{X}_B^{*}$ as previously, and $\mu^{\#}=\mu+\alpha_2+\beta_3$, $\alpha_1^{\#}=\alpha_1-\alpha_2$, $\beta_1^{\#}=\beta_1-\beta_3$, $\beta_2^{\#}=\beta_2-\beta_3$.

The normal equations $\mathbf{X}^{\#'}\mathbf{X}^{\#}\hat{\mathbf{b}}^{\#}=\mathbf{X}^{\#'}\mathbf{y}$ are formed by dropping the last two rows and two columns of the previous system, that is

$$\begin{bmatrix}18&8&4&8\\8&8&2&5\\4&2&4&0\\8&5&0&8\end{bmatrix}\begin{bmatrix}\hat{\mu}^{\#}\\\hat{\alpha}_1^{\#}\\\hat{\beta}_1^{\#}\\\hat{\beta}_2^{\#}\end{bmatrix}=\begin{bmatrix}94\\37\\16\\48\end{bmatrix}.$$

They have solutions $\hat{\mu}^{\#}=5.2696$, $\hat{\alpha}_1^{*}=-1.6180$, $\hat{\beta}_1^{*}=-0.4607$, $\hat{\beta}_2^{*}=1.7416$. This leads to the reduction $R(\mu,\alpha,\beta)=R\left(\mathbf{b}^{\#}\right)=\hat{\mathbf{b}}^{\#'}\mathbf{X}^{\#'}\mathbf{y}=511.7079$.

If we want to test the null hypothesis: $H_0:\mathbf{c}^{*}=\mathbf{0}$ that there is no interaction between factors A and B, we construct the statistic

$$F_C=\frac{Q_C/r_C}{SSE/\left(N-r_X\right)}$$

where $Q_C=R(\gamma\,|\,\mu,\alpha,\beta)$, $SSE=\mathbf{y}'\mathbf{y}-R(\mu,\alpha,\beta,\gamma)$, $r_{AB}=rank\left(\mathbf{X}_C^{*}\right)=2$, $N=18$, and $r_X=6$

Q_C can be computed by contrasting the reduction due gto fitting the full model including the interaction with that of the additive model

$$Q_C = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\beta)$$

which gives $Q_C = 541.9333 - 510.7079 = 30.2254$.

Inverting the coefficient matrix of the full model and extracting the block pertaining to interactions leads to

$$C_{22} = \begin{bmatrix} 2.2 & 1.2 \\ 1.2 & 1.7333 \end{bmatrix}$$

Letting $\hat{\mathbf{b}}_2^* = (\hat{c}_{11}^*, \hat{c}_{12}^*)'$, that is here $\hat{\mathbf{b}}_2^* = (5.4, -4/3)'$ and computing $\hat{\mathbf{b}}_2^{*'}\mathbf{C}_{22}^{-1}\hat{\mathbf{b}}_2^*$ returns the same value of Q_C as shown theoretically in (1.55).

In addition, $\mathbf{y}'\mathbf{y} = \sum_{ijk} y_{ijk}^2 = 568$ so that $SSE = \mathbf{y}'\mathbf{y} - R(\mu,\alpha,\beta,\gamma) = 26.0667$.

Hence, $F_C = \dfrac{30.2254/2}{26.0667/12} = 6.96$ and the corresponding P-value equal to $\Pr(F_{2,12} \geq 6.96) = 0.01$ indicates that there is clear evidence against the hypothesis of no interaction.

We might be inclined to go further by using $R(\alpha \mid \mu,\beta)$ (respectively $R(\beta \mid \mu,\alpha)$) and the corresponding F statistics. However, it is only under an additive model that these statistics provide meaningful tests of hypotheses. In such a case, $F(\alpha \mid \mu,\beta)$ allows testing that $I-1$ linearly independent contrasts $\alpha_i - \alpha_I$ are all equal to zero. Otherwise (interaction model), the hypothesis being tested with that F statistic turns out to be very tricky. For instance, Herr (1986) and Searle (1987) showed that using $F(\alpha \mid \mu,\beta)$ is equivalent to testing the equality of

$$\mu_{i**} = \mu_{i'**} \text{ for all } i$$

where $\mu_{i**} = \left(\sum_{j=1}^{J} n_{ij}\mu_{*j}\right) / n_{i0}$ and $\mu_{*j} = \left(\sum_{i=1}^{I} n_{ij}\mu_{ij}\right) / n_{0j}$ with $n_{i0} = \sum_{j=1}^{J} n_{ij}$ and $n_{0j} = \sum_{i=1}^{I} n_{ij}$.

Anyway, the fact remains that there is no legitimate definition of main effects when interaction occurs. For instance, in a treatment (A) by block (B) design, a natural way to characterize the treatment main effect is to consider its arithmetic mean effect over the different blocks

$$\mu_{i.} = \left(\sum_{j=1}^{J} \mu_{ij} \right) / J , \qquad (1.65)$$

so that the hypothesis being tested becomes

$$\alpha_i + \left(\sum_{j=1}^{J} \gamma_{ij} \right) / J - \left[\alpha_I + \left(\sum_{j=1}^{J} \gamma_{Ij} \right) / J \right] = 0 \text{ for } i = 1,...,I$$

In that case, one can get back to the general formulation $H_0 : \mathbf{k}'\mathbf{b}^* = \mathbf{0}$ and the corresponding statistic given in (1.47). Here testing the hypothesis $\mu_{1.} - \mu_{2.} = 0$ is equivalent to testing $\alpha_1^* + \frac{1}{3}\left(\gamma_{11}^* + \gamma_{12}^* \right) = 0$ (see exercise 1.9).

Given $\mathbf{k}' = \begin{pmatrix} 0 & 1 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$, and using the solutions $\hat{\mathbf{b}}^*$ of the full model, one has $\mathbf{k}'\hat{\mathbf{b}}^* = -1.0444$ and $\mathbf{k}'\left(\mathbf{X}^{*\prime}\mathbf{X}^* \right)^{-1}\mathbf{k} = 0.30037$ so that $Q_A = 3.5919$ and $F_A = Q_A / \hat{\sigma}^2 = 1.6535$ with a P-value of $\Pr\left(T_{12} \geq \sqrt{3.5919} \right) = 0.22$.

Computations can alternatively be carried out using the constrained system $\mathbf{k}'\mathbf{b}^* = \mathbf{0}$ presented in (1.56) and (1.57). Letting $\boldsymbol{\theta} = (\mathbf{b}', \lambda)'$, the solutions are

$$\hat{\boldsymbol{\theta}}' = 5.17073, -1.02439, -3.2439, 2.7805, 5.1707, -2.0976, -3.4390$$

and $\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{y} = 538.3414$. Then, Q_A is obtained as the difference between $R(\textit{full}) = 541.9333$ and $R(\textit{constrained}) = 538.3414$ that is 3.5919 as previously.

1.5 Additional features

1.5.1 Geometrical interpretation

Here $\mathbf{y}$ is viewed as an element of an N vectorial space $\mathbb{R}^N$. Let us consider the subspace $\mathcal{C}(\mathbf{X}) = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}\}$ with $\boldsymbol{\beta} \in \mathbb{R}^p$ of dimension p spanned by the

columns of $\mathbf{X}$. The method of OLS consists of finding an element of this subspace such that the Euclidian norm $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2$ is minimized.

Geometrically this condition is equivalent to taking $\mathbf{X}\boldsymbol{\beta}$ as the orthogonal projection of $\mathbf{y}$ on $\mathcal{C}(\mathbf{X})$ (Harville, 1997). This implies that $\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$ is orthogonal to any column vector $\mathbf{x}_j$ of $\mathbf{X}$, or equivalently that the scalar products $\mathbf{x}_j'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ are all zero, in other words that $\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})=\mathbf{0}$. The corresponding system in $\boldsymbol{\beta}$ has a unique solution $\hat{\boldsymbol{\beta}}$ which is the OLS estimator.

An illustration is shown in Figure 1 in the plane Π spanned by $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2)$ with $\overrightarrow{OA}$ being the orthogonal projection of $\overrightarrow{OM}$ where $\overrightarrow{OM}=\overrightarrow{OA}+\overrightarrow{AM}$ is equivalent to $\mathbf{y}=\mathbf{P}\mathbf{y}+(\mathbf{I}-\mathbf{P})\mathbf{y}$. Since $\overrightarrow{OA}=\mathbf{P}\mathbf{y}$, $\mathbf{P}$ is an orthogonal projector of $\mathbf{y}$ on $\mathcal{C}(\mathbf{X})$; similarly, since $\overrightarrow{AM}=(\mathbf{I}-\mathbf{P})\mathbf{y}$, $\mathbf{I}-\mathbf{P}$ is called the orthogonal projector on $\mathcal{C}^{\perp}(\mathbf{X})$, the complementary orthogonal subspace of $\mathcal{C}(\mathbf{X})$.

To see that $\mathbf{P}$ is indeed a projection, consider the scalar (or inner) product of vectors $\overrightarrow{OM}$ and $\overrightarrow{OA}$ noted $<\overrightarrow{OM},\overrightarrow{OA}>$. By definition $<\overrightarrow{OM},\overrightarrow{OA}>=\|\overrightarrow{OA}\|^2$ and has algebric counterpart $\mathbf{y}'\mathbf{P}\mathbf{y}=\mathbf{y}'\mathbf{P}^2\mathbf{y}$ thus implying that $\mathbf{P}$ is idempotent, and consequently a projector.

Following the Pythagorean theorem, $\|\overrightarrow{OM}\|^2=\|\overrightarrow{OA}\|^2+\|\overrightarrow{AM}\|^2$, which is equivalent to $\mathbf{y}'\mathbf{y}=\mathbf{y}'\mathbf{P}\mathbf{y}+\mathbf{y}'(\mathbf{I}-\mathbf{P})\mathbf{y}$ with $\|\overrightarrow{OA}\|^2=R(\boldsymbol{\beta})$ and $\|\overrightarrow{AM}\|^2=SSE$.

As such, testing the null hypothesis $H_0:\mathbf{C}'\boldsymbol{\beta}=\mathbf{0}$ amounts to searching for a vector $\overrightarrow{OB}$ belonging to subspace $\mathcal{D}$ of $\mathcal{C}(\mathbf{X})$ with a lower dimension $r<p$ such that $\overrightarrow{OB}$ is the orthogonal projection of $\mathbf{y}$ onto this subspace. The *OBA* triangle is right-angled in *B* according to the so-called theorem of the three perpendiculars. Hence, $\|\overrightarrow{OA}\|^2=\|\overrightarrow{OB}\|^2+\|\overrightarrow{BA}\|^2$.

In the case of $\mathcal{D}$ corresponding to $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, this is tantamount to writing $R(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = R(\boldsymbol{\beta}_1) + R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)$. The test of H_0 takes advantage of the fact that the statistic $R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = \left\| \overrightarrow{BA} \right\|^2$ is independent of $SSE = \left\| \overrightarrow{AM} \right\|^2$ since $\overrightarrow{BA}$ and $\overrightarrow{AM}$ are orthogonal vectors.

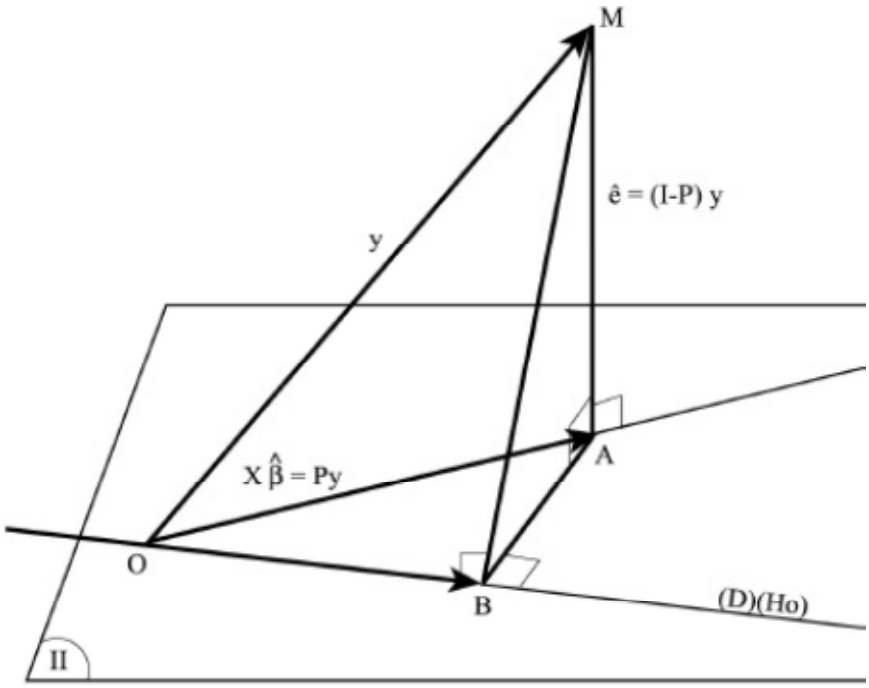

Figure 1. *Geometric interpretation of OLS*

1.5.2 Generalized least squares

We consider the same model as in (1.1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ but now assuming that $\mathbf{V} = Var(\mathbf{e})$ has a general structure not necessarily diagonal nor homogeneous.

$$\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V}). \qquad (1.66)$$

$\mathbf{V}$ being a variance covariance matrix, it is positive definite, and thus can be expressed via a Cholesky decomposition as $\mathbf{V} = \mathbf{U}\mathbf{U}'$ where $\mathbf{U}$ is a lower triangular matrix of full rank.

Let us define the one to one linear transformation

$$\mathbf{y} = \mathbf{U}\mathbf{y}^* \Leftrightarrow \mathbf{y}^* = \mathbf{U}^{-1}\mathbf{y}, \qquad (1.67)$$

the model for $\mathbf{y}^*$ can be written as

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^*, \tag{1.68}$$

with

$$\mathbf{X}^* = \mathbf{U}^{-1}\mathbf{X},\ \mathbf{e}^* = \mathbf{U}^{-1}\mathbf{e}. \tag{1.69}$$

Now, the model (1.68) is linear in $\boldsymbol{\beta}$ and has a residual term $\mathbf{e}^*$ with distribution $\mathbf{e} \sim (\mathbf{0}, \mathbf{I}_N)$. Hence, we can apply to it standard OLS techniques. In particular, the LS estimation of $\boldsymbol{\beta}$ is a solution of the system $\mathbf{X}^{*'}\mathbf{X}^*\hat{\boldsymbol{\beta}} = \mathbf{X}^{*'}\mathbf{y}^*$.

Replacing $\mathbf{X}^*$ and $\mathbf{y}^*$ by their original expressions (see 1.67 and 1.69) and given that $\mathbf{V}^{-1} = (\mathbf{U}')^{-1}\mathbf{U}^{-1}$, this system becomes

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{1.70}$$

This system is usually referred to as the system of generalized least squares and $\hat{\boldsymbol{\beta}}$ as the generalized least squares (GLS) estimator.

In fact, as shown by the formula (1.70), if $\mathbf{V}$ can be known up to a proportionality constant $\mathbf{V} \propto \underline{\mathbf{V}}$, a solution $\hat{\boldsymbol{\beta}}$ can be obtained by replacing $\mathbf{V}^{-1}$ with $\underline{\mathbf{V}}^{-1}$.

The decomposition $\mathbf{y}^* = \mathbf{P}^*\mathbf{y}^* + (\mathbf{I} - \mathbf{P}^*)\mathbf{y}^*$ with $\mathbf{P}^*\mathbf{y}^* = \mathbf{X}^*\hat{\boldsymbol{\beta}}$ and $(\mathbf{I} - \mathbf{P}^*)\mathbf{y}^* = \hat{\mathbf{e}}^*$ still applies. Premultiplying by $\mathbf{U}$ on both sides, and replacing $\mathbf{y}^*$ by $\mathbf{U}^{-1}\mathbf{y}$ gives

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}} = \mathbf{Q}\mathbf{y} + (\mathbf{I} - \mathbf{Q})\mathbf{y}, \tag{1.71}$$

where $\mathbf{Q} = \mathbf{U}\mathbf{P}^*\mathbf{U}^{-1}$, or as a function of $\mathbf{X}$ and $\mathbf{V}$,

$$\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}. \tag{1.72}$$

Since $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, $\mathbf{Q}$ (also denoted as $\mathbf{P}_{\mathbf{X},\mathbf{V}^{-1}}$) is the orthogonal projector on $\mathcal{C}(\mathbf{X})$ but now, according to the $\mathbf{V}^{-1}$ metric . Both $\mathbf{Q}$ and $\mathbf{I} - \mathbf{Q}$ are idempotent (see exercise 1.10) and thus satisfy

$$\mathbf{Q}(\mathbf{I} - \mathbf{Q}) = \mathbf{0}. \tag{1.73}$$

The GLS estimator $\mathbf{k}'\hat{\boldsymbol{\beta}}$ of an estimable function $\mathbf{k}'\boldsymbol{\beta}$ is obtained by

$$\mathbf{k}'\hat{\boldsymbol{\beta}} = \mathbf{k}'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad (1.74)$$

where $\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}$ is any g-inverse of $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$.

An estimable function $\mathbf{k}'\boldsymbol{\beta}$ is defined as previously in reference to invariance with respect to the choice of the g-inverse in (1.74) and must then verify

$$\mathbf{k}' = \mathbf{k}'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \qquad (1.75)$$

However, the matter simplifies since $\mathbf{k}'\hat{\boldsymbol{\beta}}$ is an estimable function under (1.66), if and only if, it is estimable under the simple model $\mathbf{y} \sim \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N\right)$ (see exercise 1.11).

Moreover, the GLS estimator of $\mathbf{k}'\hat{\boldsymbol{\beta}}$ is again, as OLS was in the case of $\mathbf{V} = \sigma^2\mathbf{I}_N$, the BLUE of $\mathbf{k}'\boldsymbol{\beta}$ such that

$$E\left(\mathbf{k}'\hat{\boldsymbol{\beta}}\right) = \mathbf{k}'\boldsymbol{\beta}, \qquad (1.76)$$

$$Var(\mathbf{k}'\hat{\boldsymbol{\beta}}) = \mathbf{k}'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{k}. \qquad (1.77)$$

Let us assume that $\mathbf{V}$ is known up to a coefficient of proportionality σ^2

$$\mathbf{V} = \sigma^2\underline{\mathbf{V}} \qquad (1.78)$$

and premultiplying (1.71) by $\mathbf{y}'\underline{\mathbf{V}}^{-1}$ gives

$$\mathbf{y}'\underline{\mathbf{V}}^{-1}\mathbf{y} = \mathbf{y}'\underline{\mathbf{V}}^{-1}\mathbf{Q}\mathbf{y} + \mathbf{y}'\underline{\mathbf{V}}^{-1}\left(\mathbf{I} - \mathbf{Q}\right)\mathbf{y}. \qquad (1.79)$$

Using similar notations as in the case of OLS, (1.79) can be expressed as

$$\mathbf{y}'\underline{\mathbf{V}}^{-1}\mathbf{y} = R\left(\boldsymbol{\beta}\right) + SSE, \qquad (1.80)$$

with

$$R\left(\boldsymbol{\beta}\right) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\underline{\mathbf{V}}^{-1}\mathbf{y}, \qquad (1.81)$$

$$SSE = \mathbf{y}'\underline{\mathbf{V}}^{-1}\mathbf{y} - R\left(\boldsymbol{\beta}\right), \qquad (1.82)$$

where $\hat{\boldsymbol{\beta}}$ is a GLS solution to $\mathbf{X}'\underline{\mathbf{V}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\underline{\mathbf{V}}^{-1}\mathbf{y}$.

If unknown, the parameter σ^2 can be estimated unbiasedly as

$$\hat{\sigma}^2 = SSE / \left(N - r_X \right) \qquad (1.83)$$

In the same way, the hypothesis $H_0 : \mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$ is tested via the same kind of statistic as in (1.44)

$$F = \frac{Q / r_C}{SSE / \left(N - r_X \right)}, \qquad (1.84)$$

but here, with

$$Q = \hat{\boldsymbol{\beta}}'\mathbf{C}\left[\mathbf{C}'\left(\mathbf{X}'\underline{\mathbf{V}}^{-1}\mathbf{X} \right)^{-} \mathbf{C} \right]^{-1} \mathbf{C}'\hat{\boldsymbol{\beta}}. \qquad (1.85)$$

1.6 Definition of mixed models

The extension of the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathbf{V} = Var\left(\boldsymbol{\varepsilon} \right)$ being any kind of variance covariance matrix allows to relax the assumption of uncorrelated residuals and of homogeneous variances. But, formulated this way, the model is going too far since it generates $N\left(N+1 \right)/2$ possibly distinct elements whose estimation from a sample of N observations might be problematic if not impossible.

Therefore, there is a real need for assigning some structure to $\mathbf{V}$ so as to reduce the number of unknown parameters involved in its expression. There are at least two ways to do that. First, some structure can be imposed directly on the elements of $\mathbf{V}$ on account of the type of data and design. Such situations arise for instance with longitudinal or spatial data for which $\mathbf{V}$ may a priori take some specific forms such as “autoregressive” or “Toeplitz” matrices according to the usual mathematical and software terminology. Alternatively, the vector of residuals $\boldsymbol{\varepsilon}$ can be decomposed into several components attribuable to potentially influencing factors thus resulting indirectly in some structure of $\mathbf{V}$ as proposed originally by Rao and Kleffe (1988).

Another approach consists in viewing mixed models as hierarchical models in which the parameters of the first level (data) describing some characteristics of

the experimental units (e.g individuals) are themselves described at a second level as distributions due to sampling of these experimental units in a larger population. These models have been introduced at the start in sociological research under the name of multilevels (see e.g. Goldstein, 1995), and formalized in a Bayesian context by Lindley and Smith (1972). We will now review these two approaches.

1.6.1 Rao and Kleffe's structural approach

According to these authors (in short RK), a mixed linear model is a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as in (1.66) in which the error random vector $\boldsymbol{\varepsilon}$ is decomposed as a linear combination of unobservable structural random variables $\mathbf{u}_k$ such that

$$\boldsymbol{\varepsilon} = \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k = \mathbf{Z}\mathbf{u}, \qquad (1.86)$$

where $\mathbf{Z} = (\mathbf{Z}_0, \mathbf{Z}_1, ..., \mathbf{Z}_k, ..., \mathbf{Z}_K)$ is a concatenation of $K+1$ incidence matrices $\mathbf{Z}_k$ of dimension $(N \text{ x } q_k)$ and $\mathbf{u} = (\mathbf{u}_0', \mathbf{u}_1', ..., \mathbf{u}_k', ..., \mathbf{u}_K')'$ is the corresponding $q = \sum_{k=0}^{K} q_k$ dimensional vector of structural random variables $\mathbf{u}_k = (u_{kl})$, $l = 1, ..., q_k$ such that

$$\mathbf{u} \sim (\mathbf{0}, \boldsymbol{\Sigma}_u). \qquad (1.87)$$

Although notations in (1.86) originate from symmetry arguments, they should not hide the fact that the true residual term $\mathbf{u}_0 = \mathbf{e}$ with $\mathbf{Z}_0 = \mathbf{I}_N$ is to be distinguished practically from the other K structural random variables $\mathbf{u}_1, ..., \mathbf{u}_K$ conveying real information about variation factors.

In the RK presentation, the variance covariance matrix $\boldsymbol{\Sigma}_u$ is assumed to be a linear function of real-valued parameters θ_m, $m = 1, ..., M$ weighed by square $(q \text{ x } q)$ matrices of known coefficients $\mathbf{F}_m$ such that

$$\boldsymbol{\Sigma}_u = \sum_{m=1}^{M} \theta_m \mathbf{F}_m. \qquad (1.88)$$

No specific restrictions are set on θ_m and $\mathbf{F}_m$ but these quantities should be coherent with $\boldsymbol{\Sigma}_u$ being in the parameter space.

Under this model as $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}'$, the variance covariance matrix $\mathbf{V}$ can be expressed in turn as a linear combination of the θ_m parameters

$$\mathbf{V} = \sum_{m=1}^{M} \theta_m \mathbf{V}_m \qquad (1.89)$$

where $\mathbf{V}_m = \mathbf{Z}\mathbf{F}_m\mathbf{Z}'$.

Therefore, according to RK, a linear mixed model displays linearity properties both at the levels of expectation ($\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$) and variance (1.89). Here, random effects turn out to be a device to assign some structure on $\mathbf{V}$ so as to reduce the number of parameters involved in $\mathbf{V}$.

Notice that we can arrive at (1.89) without setting a linear function of structural variables on $\boldsymbol{\varepsilon}$ since any $(N \mathrm{x} N)$ variance $\mathbf{V} = (\sigma_{ij})$ matrix can be written as $\mathbf{V} = \sum_{i \le j} \sigma_{ij}\mathbf{V}_{ij}$. This indicates that a linear relationship as (1.89) does not necessarily imply parsimony.

Nevertheless, identified structural random variables (in short random effects) may be of special interest *per se* in practice. For instance, in quantitative genetics, the analysis incorporates a vector $\mathbf{u}$ representing genetic or breeding values of individuals, whose prediction serves as a basis for selection of the top ones. Similarly, in repeated data analysis, subject specific random effects allow to predict individual profiles pertaining to e.g. marker information or growth characteristics.

Example 1.5 *Clusters and the intra class structure*

In many situations, data are structured according to clusters such as litters in biology, families in genetics, blocks in agronomy, individuals in clinical trials so that data can no longer be assumed uncorrelated. Let $i = 1,...,I$ be the index for cluster and $j = 1,...,n_i$ the one for an observation j within cluster i, a convenient way to take this structure into account is to describe the observations y_{ij} by the following model

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \varepsilon_{ij}, \qquad (1.90)$$

where $\mathbf{x}_{ij}$ is a $(p\,\mathrm{x}1)$ vector of covariates influencing response with coefficients or fixed effects $\boldsymbol{\beta}$, and ε_{ij} the random residual term decomposed as

$$\varepsilon_{ij} = u_i + e_{ij}, \qquad (1.91)$$

assuming that the cluster effects u_i are i.i.d. $\left(0,\sigma_1^2\right)$ and the within components e_{ij} are i.i.d. $\left(0,\sigma_0^2\right)$ and independent of the u_i's.

Then, $Cov\left(\varepsilon_{ij},\varepsilon_{i'j'}\right)=\sigma_0^2+\sigma_1^2$ if $i=i'$, $j=j'$, and $Cov\left(\varepsilon_{ij},\varepsilon_{i'j'}\right)=\sigma_1^2$ if $i=i'$, $j\neq j'$. Let $\mathbf{V}_i$ be the $(n_i\,\mathrm{x}\,n_i)$ variance covariance matrix of $\mathbf{y}_i=\left(y_{ij}\right)$, $\mathbf{V}_i$ can be written as in (1.89) as a linear function of the variance components σ_0^2 and σ_1^2 as

$$\mathbf{V}_i = \sigma_0^2\mathbf{I}_{n_i} + \sigma_1^2\mathbf{J}_{n_i}, \qquad (1.92)$$

where $\mathbf{I}_{n_i}$ is the $(n_i\,\mathrm{x}\,n_i)$ identity matrix and $\mathbf{J}_{n_i}=\mathbf{1}_{n_i}^{'}\mathbf{1}_{n_i}$ denotes an $(n_i\,\mathrm{x}\,n_i)$ matrix with all elements equal to 1.

Hence, the resulting $\mathbf{V}$ matrix is as follows

$$\mathbf{V}=\begin{bmatrix} \mathbf{V}_1 & \dots & \mathbf{0} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \mathbf{V}_i & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \dots & \dots & \mathbf{V}_I \end{bmatrix} = \bigoplus_{i=1}^{I}\mathbf{V}_i. \qquad (1.93)$$

This structure of $\mathbf{V}$ with $\mathbf{V}_i$ defined in (1.92) is known as «compound symmetry » or « intra-class correlation » since the correlation among any pair $j\neq j'$ of observations in any cluster i is equal to

$$\rho=\frac{\sigma_1^2}{\sigma_1^2+\sigma_0^2}, \qquad (1.94)$$

Oddly enough, the covariance between observations within a cluster in this model is equal to the variance of the true cluster effects

$$\sigma_1^2 = Cov\left(y_{ij},y_{ij'}\right)=Var\left(u_i\right), \qquad (1.95)$$

thus constraining this covariance (or correlation) to be positive (or nil). However, strictly speaking, this condition is not mandatory for $\mathbf{V}_i = \left(\sigma_0^2 + \sigma_1^2\right)\left[(1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i}\right]$ being within the parameter space (see exercise 1.12).

1.6.2 Hierarchical models

The rigorous formulation of hierarchical models dates back to Lindley and Smith (1972) for the analysis of linear models in a Bayesian framework. Now, we will briefly review this approach. Let us consider a Gaussian sampling process according to the following two stages

$$\text{i) } \mathbf{y} \mid \boldsymbol{\theta}, \mathbf{R} \sim \mathcal{N}\left(\mathbf{Z}\boldsymbol{\theta}, \mathbf{R}\right) \qquad (1.96)$$

$$\text{ii) } \boldsymbol{\theta} \mid \boldsymbol{\beta}, \mathbf{G} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\beta}, \mathbf{G}\right). \qquad (1.97)$$

The first stage i) assumes that the data vector satifies a usual linear model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \mathbf{e} \qquad (1.98)$$

where $\mathbf{Z}$ is a $\left(N \text{ x } q\right)$ matrix of known covariates and $\boldsymbol{\theta}$ is the $\left(q \text{ x } 1\right)$ corresponding vector of coefficients, and $\mathbf{e}$ is the vector of residuals assumed $\mathbf{e} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{R}\right)$.

In the second stage ii), the vector $\boldsymbol{\theta}$ of parameters is also assumed to be randomly sampled according to a linear model

$$\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\beta} + \mathbf{u}, \qquad (1.99)$$

with population mean $\mathbf{A}\boldsymbol{\beta}$ and deviation $\mathbf{u}$ from this mean being distributed with mean $\mathbf{0}$ and variance covariance $\mathbf{G}$, viz $\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{G}\right)$.

By combining these two stages, one obtains the marginal distribution of the data given $\boldsymbol{\beta}, \mathbf{G}$ and $\mathbf{R}$ after integrating out $\mathbf{u}$

$$\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{G}, \mathbf{R} \sim \mathcal{N}\left(\mathbf{Z}\mathbf{A}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}\right). \qquad (1.100)$$

In the linear case, the distribution (1.100) can be derived by inserting the model (1.99) for $\boldsymbol{\theta}$ into the model (1.98) for $\mathbf{y}$. Letting $\mathbf{X} = \mathbf{Z}\mathbf{A}$, this reduces to the usual Henderson formula for linear mixed models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \qquad (1.101)$$

where

$$E(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \qquad (1.102)$$

$$Var(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}. \qquad (1.103)$$

Assuming for the sake of simplicity that $\mathbf{R} = \sigma_0^2\mathbf{I}_N$ and $\mathbf{G} = \sigma_1^2\mathbf{I}_q$, one immediately recognizes the typical linear structure of $\mathbf{V}$ shown in (1.89).

$$\mathbf{V} = \sigma_1^2\mathbf{Z}\mathbf{Z}' + \sigma_0^2\mathbf{I}_N. \qquad (1.104)$$

Although very convenient, normality of the distribution of the random effects and residuals is not mandatory; this assumption can be discussed according to its plausibility and the kind of estimators investigated.

Example 1.6 *Random coefficients models for growth data*

The hierarchical model can be illustrated with the small data set due to Pothoff and Roy (1964) about facial growth measurements made on 11 girls and 16 boys at 4 equidistant ages (8,10,12 and 14 years). On account of a graphical visualization of individual profiles, it turns out that this data set can be analyzed by adjusting a straight line per individual.

Let i designate the index for gender ($i = 1,2$ for girls and boys respectively), j the index for period ($j = 1,2,3,4$) with t_j being the age of the child, and k be the individual index within gender ($k = 1,...,11$ for $i = 1$ and $k = 1,...,16$ for $i = 2$), the first stage model can be written as

$$y_{ijk} = A_{ik} + B_{ik}t_j + e_{ijk}, \qquad (1.105)$$

where A_{ik} and B_{ik} represent the intercept and the slope of the regression line, respectively, e_{ijk} being the residual term assumed i.i.d. $\mathcal{N}\left(0,\sigma_e^2\right)$.

Suppose now that the individuals measured are a random sample from a given population of children of both sexes, then the parameters A_{ik} and B_{ik} attached to individual ik are random variables, the expectation and variance of which can be specified as follows

$$\begin{pmatrix} A_{ik} \\ B_{ik} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}\right]. \qquad (1.106)$$

This is equivalent to writing

$$A_{ik} = \alpha_i + a_{ik},\ B_{ik} = \beta_i + b_{ik}, \qquad (1.107)$$

where a_{ik} and b_{ik} represent the deviations in intercept and slope respectively for each subject ik from their gender specific population means α_i and β_i.

Notice that the assumption of randomness of the individuals translates here with the specification of the two first moments of the attached parameters A_{ik} and B_{ik}.

Now, combining (1.107) and (1.105) gives

$$y_{ijk} = \alpha_i + \beta_i t_j + a_{ik} + b_{ik} t_j + e_{ijk}, \qquad (1.108)$$

in which the fixed part $\alpha_i + \beta_i t_j$ describes the gender specific population profile and the random part $a_{ik} + b_{ik} t_j$ its subject specific deviation counterpart.

Let $\mathbf{y}_{ik} = \left(y_{ijk}\right)$ and $\mathbf{e}_{ik} = \left(e_{ijk}\right)$ for $j = 1,2,3,4$, $\boldsymbol{\beta} = \left(\alpha_1, \alpha_2 - \alpha_1, \beta_1, \beta_2 - \beta_1\right)'$, $\mathbf{u}_{ik} = \left(a_{ik}, b_{ik}\right)'$, $\mathbf{X}_{ik} = \left(\mathbf{1}_4, \mathbf{0}_4, \mathbf{t}, \mathbf{0}_4\right)$ if $i = 1$ and $\mathbf{X}_{ik} = \left(\mathbf{1}_4, \mathbf{1}_4, \mathbf{t}, \mathbf{t}\right)$ if $i = 2$ and $\mathbf{Z}_{ik} = \left(\mathbf{1}_4, \mathbf{t}\right)$ with $\mathbf{t} = \left(t_1, t_2, t_3, t_4\right)'$, then the model (1.108) can be written under its typical linear mixed model form

$$\mathbf{y}_{ik} = \mathbf{X}_{ik}\boldsymbol{\beta} + \mathbf{Z}_{ik}\mathbf{u}_{ik} + \mathbf{e}_{ik}, \qquad (1.109)$$

where $\mathbf{u}_{ik} \sim_{iid} \mathcal{N}\left(\mathbf{0}, \mathbf{G}_0\right)$ and $\mathbf{e}_{ik} \sim_{iid} \mathcal{N}\left(\mathbf{0}, \mathbf{R}_0\right)$ with $\mathbf{G}_0$ being the variance covariance matrix in (1.106) and $\mathbf{R}_0 = \sigma_e^2 \mathbf{I}_4$.

1.6.3 Multilevel linear models

In many areas (industry, education, biology, medicine), data naturally have a nested structure. For instance, workers are nested within firms, children within schools, animals within litters, patients within clinics. Models to account for such clustering in data consist of introducing random effects so as to partition the variation into, between and within cluster components. In the examples

given above, there are two levels of variation: elementary response at the lowest level (level 1 by convention) and cluster at the highest level (level 2). But the hierarchy may be more complex with any number of levels. For instance, in crossnational studies, repeated observations are nested within persons, these persons are nested within organizational units such as schools, clinics which themselves might be nested within districts, districts within states, states within countries, etc…

In the sociological sciences, these models are often referred to as multilevel models (Goldstein, 1995). In fact they are from a theoretical point of view a special case of the hierarchical models described previously and they just appear as such in some areas of applications. This is the reason why we are just going to briefly outline their particularity in terms of terminology, formalism and notations.

Although the hierarchy can incorporate many levels, this presentation is restricted to the two level linear model which displays the essential features of such models. A specificity, if not a difficulty, for the non-expert lies in the notations which in some cases (e.g. indices of factors) are reversed as compared to the traditional mixed model literature;

Here, the index i denotes the level-1 units (e.g. measurements) while i refers to level-2 units (e.g. individuals). There are n_2 units at level-2 $\left(j=1,...,n_2\right)$ and n_{1j} for each level-1 unit nested within level 2 $\left(i=1,...,n_{1j}\right)$.

Letting $\mathbf{y}_j=\left(y_{ij}\right)_{1\le i\le n_{1j}}$ for $j=1,...,n_2$, the general model for responses involving q level-2 parameters can be written as

$$\mathbf{y}_j=\mathbf{Z}_j\boldsymbol{\beta}_j+\mathbf{b}_j^{(1)}, \qquad (1.110)$$

where $\mathbf{Z}_j$ is a $\left(n_{1j}\,\mathrm{x}\,m\right)$matrix of explanatory variables with corresponding $\left(m\mathrm{x}1\right)$ vector $\boldsymbol{\beta}_j$ of unknown parameters, and $\mathbf{b}_j^{(1)}$is the $\left(n_{1j}\,\mathrm{x}\,1\right)$vector of the

level-1 random components $b_{ij}^{(1)}$ assumed i.i.d., and usually Gaussian, with mean zero and variance covariance viz. $\mathbf{b}_j^{(1)} \sim_{iid} \mathcal{N}\left(\mathbf{0}, \mathbf{G}_j^{(1)}\right)$.

These components correspond to the usual residual terms $\mathbf{b}_j^{(1)} = \mathbf{e}_j$ of a linear mixed model with $\mathbf{G}_j^{(1)}$ taken often as $\mathbf{G}_j^{(1)} = \sigma_1^2 \mathbf{I}_{n_{ij}}$.

At level-2, the model for $\boldsymbol{\beta}_j$ involves both covariate information such as discrete factors cross-classified with cluster units (e.g. treatments) and random variations among those level-2 units. It can be written as

$$\boldsymbol{\beta}_j = \mathbf{A}_j \boldsymbol{\beta} + \mathbf{b}_j^{(2)} \qquad (1.111)$$

where $\mathbf{A}_j$ is a $(m \mathrm{x} p)$ matrix of covariates with corresponding $(p \mathrm{x} 1)$ vector $\boldsymbol{\beta}$ of unknown coefficients, and $\mathbf{b}_j^{(2)}$ is the $(m \mathrm{x} 1)$ vector of the level-2 random effects (e.g. individuals, litters) assumed $\mathbf{b}_j^{(2)} \sim_{iid} \mathcal{N}\left(\mathbf{0}, \mathbf{G}^{(2)}\right)$.

Several comments are worth mentioning at this stage. First, letting $\mathbf{X}_j = \mathbf{Z}_j \mathbf{A}_j$ and substituting equation (1.111) into (1.110) gives

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{b}_j^{(2)} + \mathbf{b}_j^{(1)}, \qquad (1.112)$$

which clearly indicates that multilevel linear models belong to the class of linear mixed models as defined previously.

Secondly, in certain instances, it may be suitable to assume that some components β_{jk} of $\boldsymbol{\beta}_j$ have no random counterparts (see example 1.7). This may be justified either on theoretical or data-based grounds. For instance, the estimation of the corresponding variance $g_k^{(2)}$ can be very small suggesting that this component does not vary practically across level-2 units. In that case the model for $\boldsymbol{\beta}_j$ becomes

$$\boldsymbol{\beta}_j = \mathbf{A}_j \boldsymbol{\beta} + \mathbf{B}_2 \mathbf{b}_j^{(2)}, \qquad (1.113)$$

where $\mathbf{B}_2$ is a $(m \mathrm{x} m')$ appropriate design matrix having e.g. typical form

$\mathbf{B}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$ if component 3 is not random.

Extensions of such models to higher levels is straightforward. For instance a 3 level model ijk can be expressed as the concatenation of three different models, each of them operating at a given level of the hierarchy

$$y_{ijk} = \mathbf{z}'_{ijk}\boldsymbol{\beta}_{jk} + b^{(1)}_{ijk},$$

$$\boldsymbol{\beta}_{jk} = \mathbf{A}_{jk}\boldsymbol{\beta}_k + \mathbf{B}_2\mathbf{b}^{(2)}_{jk},$$

$$\boldsymbol{\beta}_k = \mathbf{A}_k\boldsymbol{\beta} + \mathbf{B}_3\mathbf{b}^{(3)}_k.$$

Such models can generate complex variance covariance structures of the observations especially in the case of random coefficient models (exercise 1.14).

Example 1.7 *Growth data of children (1.6 continued)*

Here the level-1 units are measurements $(i = 1,...,4)$ and level-2 units are children $(j = 1,...,27)$. The level-1 model can be written as

$$y_{ij} = \beta_{0,j} + \beta_{1,j}t_j + b^{(1)}_{ij} \qquad (1.114)$$

or, in matrix notations,

$$\mathbf{y}_j = \mathbf{Z}_j\boldsymbol{\beta}_j + \mathbf{b}^{(1)}_j$$

as in (1.110), with $\mathbf{Z}_j = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{pmatrix}$ and $\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{0,j} \\ \beta_{1,j} \end{pmatrix}$, $\beta_{0,j}$ and $\beta_{1,j}$ referring to the intercept and slope of subject j.

At the second level (subjects), $\beta_{0,j}$ and $\beta_{1,j}$ are decomposed into

$$\beta_{0,j} = \beta_{0,B} + a_j\Delta\beta_0 + b^{(2)}_{0,j}$$

$$\beta_{1,j} = \beta_{1,B} + a_j\Delta\beta_1 + b^{(2)}_{1,j},$$

or, in vector notations,

$$\boldsymbol{\beta}_j = \begin{pmatrix} \mathbf{I}_2 & a_j \mathbf{I}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_B \\ \boldsymbol{\Delta\beta} \end{pmatrix} + \mathbf{b}_j^{(2)}, \qquad (1.115)$$

where $\boldsymbol{\beta}_j = \left(\beta_{0j}, \beta_{1j}\right)'$, $\boldsymbol{\beta}_B = \left(\beta_{0,B}, \beta_{1,B}\right)'$ refers to intercept and slope in boys, $\boldsymbol{\Delta\beta} = \left(\Delta\beta_0, \Delta\beta_1\right)'$ to the differences between girls and boys in intercept and slope respectively; a_j is an indicator variable equal to 1 if the child is a girl and zero otherwise ; and $\mathbf{b}_j^{(2)} = \left(b_{0,j}^{(2)}, b_{1,j}^{(2)}\right)'$ refers to the subject j specific random deviations from their population means in intercept and slope respectively assumed i.i.d. with mean $\mathbf{0}$ and variance $\mathbf{G}^{(2)} = \begin{pmatrix} g_{11}^{(2)} & g_{12}^{(2)} \\ g_{12}^{(2)} & g_{22}^{(2)} \end{pmatrix}$.

As indicated previously, we may consider the submodel having no random component in the subject specific slope $\beta_{1,j}$.

$$y_{ij} = \beta_{0,B} + \beta_{1,B} t_j + b_{0j}^{(2)} + a_j \left(\Delta\beta_{0,B} + \Delta\beta_{1,B} t_j\right) + b_{ij}^{(1)}. \quad (1.116)$$

Imagine now that children are nested among n_3 districts or other geographical areas (indexed by k), we can readily extend the previous model to take into account this additional level of clustering.

1.6.4 Definition and critical assessment

Definition 1.3 *Under the hierarchical presentation, a linear mixed model can be defined as a linear model in which all or part of the parameters associated to some experimental units are treated as random variables due to random sampling of these units from some predefined larger population.*

It is important to emphasize that this definition relies at least conceptually on some random device to determine what experimental units will be taken out of a given population of them. Consequently, parameters and their inference refer to this population of experimental units. This reference to a random process was clearly outlined in the beginning by Eisenhart (1947) to distinguish what he

called ANOVA Models I (fixed) and Models II (variance components). Features of fixed and random models were also discussed by Wilk and Kempthorne (1955).

In some instances, the distinction between fixed and random effects is relatively easy to do. If an experimenter wants to compare some treatments (e.g. drugs, varieties,…), he will not draw them at random in a population of possible treatments so that treating this factor as fixed and not random makes sense. Conversely, animals, individuals or plants selected for such comparisons can be appropriately sampled at random from some population(s) and viewing their effects as random is generally appropriate.

But, in many cases, data sets do not originate from rigorously planned experiments and the decision about the fixed or random status of some factors remains unclear, if not highly questionable. This is especially true for location and time effects such as field or year effects in agriculture and animal production. For instance, although, the calendar years per se are not drawn at random but successively, one may often infer that their effects on response are more or less unpredictable, and surely not repeatable as such.

This perspective of possible repetitions of the experiment must be kept in mind when deciding whether some effects should be treated as fixed or random. If replicates have to be produced as for instance in a simulation study, would we assume that the effects of say levels 1,2,…, of this factor remain the same, or would we prefer to make these effects vary from one replicate to the other, and thus, draw them from some probability distribution?

This perception of fixed vs random effects on how to generate them over replications of data sets may be very helpful not only in designing the operational models for the statistical analysis of the data, but also for studying the statistical properties of the estimators via simulation.

Another concern is often put forward in order to decide whether effects are fixed or random which consists of what kind of inference the statistician would like to

make. Does he want to restrict inference to the particular effects present in the data set or to the parameters of the distribution from which they are sampled? This is for sure, an important point of debate but this issue should not be brought up before questioning how the experiment was designed and how data were collected. Actually, the goal of the analysis is not *per se* an argument whether effects are to be treated as fixed or random since the objectives should be in agreement with the design. Nevertheless, it is unsure that all these queries may help in reducing the reader's headaches about this question "fixed vs random";

Salvation can come from a completely different point of view as the one proposed by Bayesian statistics. Instead of arguing about randomness or not of effects, Bayesians rather quantify the amount of knowledge (more precisely uncertainty) they have about such effects via probability distributions, the theory of which lies on rationality principles prior to observing data.

In this case, probability does not represent the limiting relative frequency of occurrence of some events in an infinite number of trials but the degree of belief in a proposition. As clearly stated by Malécot (1947) "If one admits determinism in fact, this is not the phenomena that are random, rather, it is the knowledge we have about them". Sometimes this precludes the paradoxal attitude of treating some effects as they were random while they are obviously not, as illustrated below.

Example 1.8 *Why treating bulls as random? (Example 1.5continued)*

Let us again consider the model of example 1.5 but now in the context of genetic evaluation of sires on the phenotypic value of their progeny (e.g. milk yield per lactation of daughters in dairy cattle).,

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + u_i + e_{ij},$$

where y_{ij} is the milk yield of the jth heifer progeny $\left(j = 1,...,n_i\right)$ sired by the ith bull ; u_i the genetic transmitting ability (or breeding value) of that bull and e_{ij},

the usual error term specific to this record assumed i.i.d. with mean zero and variance σ_e^2.

In that model, environmental effects such as those of year, season and herd on yield y_{ij} are accounted for as fixed effects via the term $\mathbf{x}_{ij}^{'}\boldsymbol{\beta}$. Under its simplest form, this term reduces to a common population mean μ and the model becomes

$$y_{ij} = \mu + u_i + e_{ij}. \qquad (1.117)$$

Henderson (1984) proposed to the animal breeding community to rank bulls on the basis of what he called best linear unbiased prediction (BLUP) of $\mu_i = \mu + u_i$, that is , in the case of a known μ,

$$\hat{\mu}_i = \mu + \hat{u}_i = \mu + b_i\left(y_{i.} - \mu\right), \qquad (1.118)$$

where $y_{i.} = \left(\sum_{j=1}^{n_i} y_{ij}\right) / n_i$ and $b_i = Cov\left(u_i, y_{i.}\right) / Var\left(y_{i.}\right)$.

The main reason for that was that $\hat{\mu}_i$ minimizes the mean squared error of prediction $MSE_i = E\left[\left(\hat{\mu}_i - \mu_i\right)^2\right]$ which is also equal to $E\left[\left(\hat{u}_i - u_i\right)^2\right]$. In fact, this means that the u_i's are viewed as random effects $u_i \sim_{iid} \left(0, \sigma_u^2\right)$ so that $b_i = \sigma_u^2 / \left(\sigma_u^2 + \sigma_e^2 / n_i\right)$, this parameter acting as shrinkage factor of the sample sire mean $y_{i.}$ towards the population mean μ.

But, actually, progeny tested sires are far away from being a random sample of bulls since they have been highly selected prior to progeny testing on criteria related to u_i. In addition, on the contrary to what could have been said, there is nothing random in the theory of quantitative genetics regarding the definition of the transmitting ability of a given individual. Moreover, the purpose of animal breeding companies is to evaluate and compare these specific bulls but not to estimate parameters of the population which they come from. Therefore,

logically, sire effects ought to be viewed as fixed while they are treated as random in Henderson's approach.

The Bayesian approach represents a way to reconcile practice and logic by assigning a prior distribution to the u_i's (Lefort, 1980, Blasco, 2001). In addition, this prior information is usually invariant with respect to any permutation of the i indices (exchangeability property) implying that they all have all the same distribution with mean zero and variance σ_u^2. Additional assumptions involve independence and normality but although very convenient, they are not mandatory. As a matter of fact, the Bayesian theory can accommodate prior distributions to real situations encountered in practice such as genetic kinship among bulls, heterogeneity in the recrutement of young bulls with some of them more selected than the others.

1.7 Appendix

1.7.1 Properties of $\mathbf{P}=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$ (see 1.18,1.19,1.20)

Proof $\mathbf{PX}=\mathbf{X}$

This is a corollary of the following two lemmas (see also Searle, 1982)

Lemma 1: For any real $\left(n\,\mathrm{x}\,n\right)$ matrix $=\left(a_{ij}\right)$, the condition $\mathbf{A}'\mathbf{A}=\mathbf{0}$ implies $\mathbf{A}=\mathbf{0}$

As a matter of fact, the jth element of $\mathbf{A}'\mathbf{A}$ is equal to $\sum_{i=1}^{n}a_{ij}^{2}$. Its nullity implies that $a_{ij}=0$ for any i and the same reasoning applies to any element j of the product.

Lemma 2: For any real $\mathbf{R}$, $\mathbf{S}$ and $\mathbf{X}$ matrices, the property $\mathbf{RX}'\mathbf{X}=\mathbf{SX}'\mathbf{X}$ implies $\mathbf{RX}'=\mathbf{SX}'$.

This resorts from the following identity

$$\left(\mathbf{RX}'\mathbf{X}-\mathbf{SX}'\mathbf{X}\right)\left(\mathbf{R}-\mathbf{S}\right)'=\left(\mathbf{RX}'-\mathbf{SX}'\right)\left(\mathbf{RX}'-\mathbf{SX}'\right)'$$

If $\mathbf{RX}'\mathbf{X}=\mathbf{SX}'\mathbf{X}$, then the right hand side of the identity is nil and we can apply to it lemma 1. Hence $\mathbf{RX}'=\mathbf{SX}'$.

A g-inverse $\left(\mathbf{X}'\mathbf{X}\right)^{-}$ of $\mathbf{X}'\mathbf{X}$ verifies by definition $\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{X}=\mathbf{X}'\mathbf{X}$. Applying lemma 2 to this identity yields $\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'=\mathbf{X}'$ that is after transposition $\mathbf{PX}=\mathbf{X}$ (QED).

Proof $\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$ *is invariant to the choice of* $\left(\mathbf{X}'\mathbf{X}\right)^{-}$.

Let $\mathbf{G}_1$ and $\mathbf{G}_2$ two different g-inverses of $\mathbf{X}'\mathbf{X}$. Then, knowing that $\mathbf{PX}=\mathbf{X}$, this implies that $\mathbf{XG}_1\mathbf{X}'\mathbf{X}=\mathbf{XG}_2\mathbf{X}'\mathbf{X}$ which following lemma 2 reduces to $\mathbf{XG}_1\mathbf{X}'=\mathbf{XG}_2\mathbf{X}'$ (QED).

Proof $\mathbf{P}=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$ *is idempotent*

By definition, $\mathbf{P}^2=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$ that is $\mathbf{P}^2=\mathbf{PX}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$. Since $\mathbf{PX}=\mathbf{X}$, this yields $\mathbf{P}^2=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'=\mathbf{P}$ (QED).

1.8 Exercises

1.1 Let $\mathbf{X}$ be a $(N \text{x} p)$ matrix of independent variables of a linear model such that $p \leq N$. Show that $rank(\mathbf{X}'\mathbf{X}) = rank(\mathbf{X})$

1.2 Consider the following LS system of equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$

$$\begin{bmatrix} 41 & 64 & 23 \\ 64 & 29 & -35 \\ 23 & -35 & -58 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 119 \\ 112 \\ -7 \end{bmatrix}$$

1) What is the rank of $\mathbf{X}'\mathbf{X}$?

2) Which functions among the following ones are estimable ?

a) $\beta_1 + \beta_2$: b) $\beta_1 - \beta_2 + 2\beta_3$:c) $\beta_1 - \beta_2$:d) $\beta_2 + \beta_3$ e) $2\beta_1 + \beta_2 - \beta_3$

1.3 Show how to derive formula $\mathbf{X}^* = \mathbf{X}\mathbf{T}'(\mathbf{T}\mathbf{T}')^{-1}$ in (1.10) and check numerically on example 1.3.

1.4 Show that the BLUE of the estimable function $\mathbf{k}'\boldsymbol{\beta}$ in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I}_N)$ is $\mathbf{k}'\hat{\boldsymbol{\beta}}_{OLS}$

1.5 Let $\mathbf{H} = \mathbf{G}\mathbf{X}'\mathbf{X}$ where $\mathbf{G}$ is any g-inverse of $\mathbf{X}'\mathbf{X}$. Show that

1) $\mathbf{H}$ is idempotent and 2) $rank(\mathbf{H}) = rank(\mathbf{X})$.

1.6 Let $\mathbf{G}$ be a standard g-inverse of $\mathbf{X}'\mathbf{X}$, show that $\mathbf{G}_R = \mathbf{G}(\mathbf{X}'\mathbf{X})\mathbf{G}'$ is a symmetric, reflexive g-inverse of $\mathbf{X}'\mathbf{X}$.

1.7 Show how to derive the result $SSE(Restricted) = \mathbf{y}'\mathbf{y} - \left(\hat{\boldsymbol{\beta}}_0'\mathbf{X}'\mathbf{y} + \boldsymbol{\lambda}'\mathbf{m}\right)$ given in (1.59)

1.8 Analyze the following data set pertaining to scores given by experts (factor A) to animals out of different sires (factor B) as in Example 1.4, but using sum (Sy) and sums of squares (Sy2) instead of elementary observations.

Cell	Expert	Sire	n	Sy	Sy2
1	1	1	4	344	30242
2	1	2	3	258	22662
3	1	3	4	318	25758
4	1	4	4	274	19036
5	2	1	4	474	61366
6	2	2	2	211	22741
7	2	3	1	115	13225
8	2	4	4	325	27445
9	3	2	2	256	34946
10	3	3	4	509	73773
11	3	4	4	357	33065

n =number of observations Sy : sum of the observations per suclass ; Sy2 : sum of squares

1.9 Consider a two-way ANOVA model $E\left(y_{ijk}\right) = \mu + \alpha_i + \beta_j + \gamma_{ij}$ for $i = 1,2$, and $j = 1,2,3$ with $\mu_{ij} = E\left(y_{ijk}\right)$ being the ij cell mean (Example 1.4) Show that testing the null hypothesis $H_0 : \mu_{1.} = \mu_{2.}$ where $\mu_{i.} = \frac{1}{2}\sum_{j=1}^{2}\mu_{ij}$ is equivalent to testing

$$\alpha_1^* + \frac{1}{3}\left(\gamma_{11}^* + \gamma_{12}^*\right) = 0$$

where $\alpha_1^* = \mu_{13} - \mu_{23}$, $\gamma_{11}^* = (\mu_{11} - \mu_{21}) - (\mu_{13} - \mu_{23})$ $\gamma_{12}^* = (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23})$.

1.10 In the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{V})$, show that $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$ defined in (1.72) and $\mathbf{I}_N - \mathbf{Q}$ are idempotent.

1.11 Show that $\mathbf{k}'\boldsymbol{\beta}$ under the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, $\mathbf{e} \sim (\mathbf{0}, \mathbf{V})$ is an estimable function, if and only if, it is estimable under the same model with $\mathbf{e} \sim \left(\mathbf{0}, \sigma^2\mathbf{I}\right)$

1.12 Let $\mathbf{R}_n = \rho\mathbf{J}_n + (1-\rho)\mathbf{I}_n$ designate an intra class correlation structure among observations within a cluster of size n, with correlation coefficient ρ. Show that ρ is not necessarily positive.

1.13 Consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{V})$ where $\underline{\mathbf{P}}$ is defined as $\underline{\mathbf{P}} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$.

1) Show that $(\mathbf{I}_N - \mathbf{V}\underline{\mathbf{P}})\mathbf{y}$ and $\mathbf{V}\underline{\mathbf{P}}\mathbf{y}$ are uncorrelated random variables.

2) Prove that $\mathrm{Var}(\mathbf{I}_N - \mathbf{V}\underline{\mathbf{P}})\mathbf{y} = \mathbf{V} - \mathbf{V}\underline{\mathbf{P}}\mathbf{V}$.

3) Show that $\mathbf{V}^{-1} - \underline{\mathbf{P}}$ is a g-inverse of $\mathbf{V} - \mathbf{V}\underline{\mathbf{P}}\mathbf{V}$.

4) Use the previous results to derive the GLS estimator of $\mathbf{X}\boldsymbol{\beta}$ from the linear model for $(\mathbf{I}_N - \mathbf{V}\underline{\mathbf{P}})\mathbf{y}$.

1.14 Consider a random coefficient model $y_{it} = \mathbf{x}_i'\boldsymbol{\beta} + a_i + b_i t + e_{it}$ similar to those described in (1.108 and 1.114-115) where $\boldsymbol{\beta}$ refers to fixed effects and a_i and b_i correspond to the individual deviations in intercept and slope respectively such

that the $\mathbf{u}_i = (a_i, b_i)'$ are assumed $\mathbf{u}_i \sim_{iid} (\mathbf{0}, \mathbf{G})$ with $\mathbf{e}_i = (e_{it})$ being $\mathbf{e}_i \sim_{iid} (\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

1) Give expressions of $Var(y_{it})$ and $Cov(y_{is}, y_{it})$ for $s \neq t$ as a function of time points (s,t), σ^2 and elements of $\mathbf{G} = \begin{pmatrix} g_{00} & g_{01} \\ g_{01} & g_{11} \end{pmatrix}$.

2) Extend these expressions to a second degree time adjustment $y_{it} = \mathbf{x}_i'\boldsymbol{\beta} + a_i + b_i t + c_i t^2 + e_{it}$, and next to any degree q.

1.15 Consider the one-way model as in Example 1.8 $y_{ij} = \mu + u_i + e_{ij}$, $i = 1,...,I$ and $j = 1,...,n_i$ with the cluster effects u_i assumed i.i.d. $(0, \sigma_1^2)$ and the within components e_{ij} being i.i.d. $(0, \sigma_0^2)$ and independent of the u_i's.

1) Using the system of equations $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, show that the GLS estimator $\hat{\mu}$ of μ can be expressed as a weighted mean $\hat{\mu} = \left(\sum_{i=1}^{I} w_i y_{i.}\right) / \sum_{i=1}^{I} w_i$ of $y_{i.} = \left(\sum_{j=1}^{n_i} y_{ij}\right) / n_i$ where w_i is an appropriate weighing factor.

2) In the case of $\sigma_1^2 > 0$, give a simple expression of w_i as a function of n_i and $\lambda = \sigma_0^2 / \sigma_1^2$ and discuss the result obtained accordingly.

References

Blasco, A. (2001). The Bayesian controversy in animal breeding. *Journal of Animal Science*, 79, 2023-2046.

Dempfle, L. (1977). Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayésiens. *Annales de Génétique et de Sélection Animale*, **9**, 27–32.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*,**3**, 1-21.

Goldstein, H. (1995). *Multilevel Statistical Models*. Wiley, New York.

Harville, D.A. (1997). *Matrix algebra from a statistician's perspective*. Springer, New York.

Harvey, W. R. (1975). *Least-squares analysis of data with unequal subclass numbers*. Technical Report ARS H-4, Data Systems Application Division, Agricultural Research Service, USDA.

Henderson, C. R. (1953). Estimation of Variance and Covariance Components, *Biometrics*, 9, 226–252.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph.

Herr, D. G., (1986). On the history of ANOVA in unbalanced, factorial designs. *The American Statistician*, **40**, 265-270.

Lefort, G. (1980). Le modèle de base de la sélection: justification et limites. In J. M. Legay et al., (Editor). *Biométrie et Génétique*, pages 1-14. INRA Publications.

Lindley, D.V., & Smith, A.F.M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society* B, **34**, 1-41.

Littell, R.C., Milliken, G. A., Stroup, W.W., & Freund, R. J. (2002). *SAS for Linear Models* (fourth edition). SAS Institute Inc, Cary, NC.

Malécot, G. (1947). Les critères statistiques et la subjectivité de la connaissance scientifique. *Annales de l'université de Lyon*, **10**, 43-74. English translation by D Gianola, *Genetics Selection Evolution*, **31**, 269-298.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, (2nd edition). Wiley, New York.

Rao, C.R., & Kleffe, J. (1988), *Estimation of variance components and applications*. North Holland series in statistics and probability. Elsevier, Amsterdam.

Rao, C.R., Toutenburg, H., Shalabh, & Heumann, C. (2007). *Linear models and generalizations*, (3rd edition). Springer Verlag, Berlin.

Searle, S. R. (1966). *Matrix Algebra for the Biological Sciences*. John Wiley, New York.

Searle, S. R. (1971). *Linear Models.* Wiley. New York

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.

Searle, S. R. (1987). *Linear Models for Unbalanced Data*. Wiley, New York

Stigler, S. M. (1986). *The history of statistics: the measurement of uncertainty before 1900*. The Belknap Press of Harvard University Press, Cambridge MA and London UK.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes, *Journal of the American Statistical Association*, **29**, 51-66.

Wilk, M. B., & Kempthorne, O. (1955). Fixed, mixed and random models. *Journal of the American Statistical Association*, **50**, 1144-1167.

2
Prediction of random effects

2.1 Introduction

In everyday life, the concept of prediction from observations arises naturally when asking questions as:

i) What is the IQ of this person knowing his age, his social environment and his test scores?

ii) How tall will this boy or that girl born 1 year ago be when he/she is 20 given his height records?

iii) How can the genetic merits of bulls be compared given their daughter milk's yield performance?

iv) What is the pattern of the ore grade in a mining block from observed samples taken at different known locations?

Usually, prediction and estimation will be considered as synonyms although they are not. Most people will not make a difference in the issue of predicting the height of a child at 20 versus that of estimating how much taller are boys are than girls at that age, thus requiring a definition of this concept.

This will lead us to review the different methods of prediction such as Best, Best Linear and Best Linear Unbiased Predictions (BP, BLP and BLUP respectively) according to the assumptions made on the joint distribution of the quantity to predict and the observations. In the next section, we will focus on an important indirect approach to BLUP known as Henderson's Mixed Model Equations (HMME) and see also how these equations can be interpreted within a Bayesian framework.

2.2 Direct approach

Definition 2.1 *Making a prediction consists of substituting a non observable random variable W, (the predictand) in the conditions of the experiment, by a new random variable $\hat{W}$, (the predictor), built as a function f of an observable random variable Y i.e. $\hat{W} = f(Y)$ such that the distribution of $\hat{W}$ should be as close as possible to that of W with respect to some criterion.*

Several criteria may be envisioned. This could be a formal measure of distance between two distributions as those of Kullback-Leibler, or more simply a criterion such as the mean square error $E\left[\left(\hat{W} - W\right)^2\right]$.

In summary, while estimation deals with inferring values to population parameters, prediction is concerned with random variables.

Following to the pionnering work of CR Henderson in this field, different methods of prediction are of interest depending on the assumptions made on the joint density of (W, Y), the first one being known as Best Prediction.

2.2.1 Best Prediction (BP)

We will suppose that W is a scalar and Y either a scalar or a vector. The best prediction of W based on Y is defined here with respect to the mean square error (MSE).

Let us write $\hat{W} = f(Y)$ with upper case letters as the predictor, and $\hat{w} = f(y)$ with lower case letters as a realized value of it.The MSE can be decomposed as

$$E\left[(\hat{W} - W)^2\right] = Var(\hat{W} - W) + \left[E(\hat{W}) - E(W)\right]^2. \qquad (2.1)$$

One can apply to the first term on the right side, the theorem of « conditioning-deconditioning" i.e.

$$Var(\hat{W} - W) = E_Y\left\{Var\left[(\hat{W} - W)\middle|Y = y\right]\right\} + Var_Y\left\{E\left[(\hat{W} - W)\middle|Y = y\right]\right\}. \qquad (2.2)$$

Now, conditionally to $Y = y$, the random variable $\hat{W}|Y = y$ is equal to a constant say $\hat{w}$; its expectation is then $\hat{w}$ and its variance is zero so that (2.2) reduces to

$$Var(\hat{W} - W) = E_Y\left[Var(W|Y = y)\right] + Var_Y\left[\hat{w} - \mathrm{E}(W|Y = y)\right]. \quad (2.3)$$

The first quantity in (2.3) does not depend on the choice of the predictor; the second one vanishes if one chooses $\hat{w}$ as the conditional mean (Harville, 1990),

$$\hat{w} = \mathrm{E}(W|Y = y). \quad (2.4)$$

By construction, such a predictor verifies

$$E(\hat{W}) = E_Y\left[E(W|Y = y)\right] = E(W) \quad (2.5)$$

so that the MSE reduces to only the variance term.

It is thus unbiased (in the sense of 2.5) and minimizes (2.1) so that

$$MSE = Var(\hat{W} - W) = E_Y\left[Var(W|Y = y)\right]. \quad (2.6)$$

Similarly, using the same reasoning, we can establish the expression of $Cov(\hat{W}, W)$ from the following identity

$$Cov(\hat{W}, W) = E_Y\left[Cov\left(\hat{W}|Y = y, W|Y = y\right)\right] + Cov_Y\left[E\left(W|Y = y\right), E\left(\hat{W}|Y = y\right)\right].$$

The first term is nil since it involves the covariance between $W|Y = y$ and its expectation, and the second term is the covariance between $\hat{W}$ and itself, so that

$$Cov(\hat{W}, W) = Var(\hat{W}). \quad (2.7)$$

Hence, $\hat{W}$ and $\hat{W} - W$ are uncorrelated so that

$$Var(W) = Var(\hat{W}) + Var(\hat{W} - W) \quad (2.8)$$

Since $Var(\hat{W} - W)$, the variance of prediction errors is always positive or zero, (2.8) indicates that $Var(\hat{W})$, the variance of the predictor in the population is always smaller or equal to $Var(W)$ the variance of the predictand implying some restriction in the variability of the random variable after it has been predicted.

Formula (2.8) can be rewritten as

$$Var(\hat{W} - W) = Var(W) - Var(\hat{W}) \quad (2.9)$$

Moreover, due to (2.7), R^2 the squared coefficient of linear correlation between $\hat{W}$ and W, is simply the ratio of the variance of the predictor to the variance of the predictand

$$R^2 = \frac{\left[Cov\left(\hat{W},W\right)\right]^2}{Var(\hat{W})Var(W)} = \frac{Var(\hat{W})}{Var(W)}. \tag{2.10}$$

Therefore, the variance of the prediction error is simply

$$Var(\hat{W}-W) = Var(W)\left(1-R^2\right). \tag{2.11}$$

Since minimizing this variance is equivalent to maximizing R^2, this formula indicates that the best predictor is also the function of Y which maximizes R^2.

The result that the BP is the conditional mean was first established by Cochran (1951). It is both simple and important but assumes we know the joint distribution of W and Y and we are able to derive the expectation of the conditional distribution of W given Y, a situation that may not be trivial.

Example 2.1 *Best Prediction under normality*

Assume we want to predict W based on vector $\mathbf{Y}$ when their joint density is Gaussian $\mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ with $\boldsymbol{\mu} = \begin{pmatrix}\mu_W \\ \boldsymbol{\mu}_Y\end{pmatrix}$, and $\boldsymbol{\Sigma} = \begin{pmatrix}\Sigma_{WW} & \boldsymbol{\Sigma}_{WY} \\ \boldsymbol{\Sigma}_{YW} & \boldsymbol{\Sigma}_{YY}\end{pmatrix}$.

In this case, the conditional distribution of W given $\mathbf{Y}=\mathbf{y}$ is well known; this is also a Gaussian distribution

$$W \mid \mathbf{Y}=\mathbf{y} \sim \mathcal{N}\left(\mu_{W.Y}, \boldsymbol{\Sigma}_{WW.Y}\right), \tag{2.12}$$

with expectation

$$\mu_{W.Y} = \mu_W + \boldsymbol{\Sigma}_{WY}\boldsymbol{\Sigma}_{YY}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}_Y\right), \tag{2.13}$$

and variance

$$\boldsymbol{\Sigma}_{WW.Y} = \boldsymbol{\Sigma}_{WW} - \boldsymbol{\Sigma}_{WY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YW}. \tag{2.14}$$

Two features of the result in (2.12) deserve attention. First, the conditional mean i.e. BP, is here a linear function of $\mathbf{Y}$; this is actually the regression equation of W in $\mathbf{Y}$. Second, the variance of the conditional distribution does not depend

on the specific value of $\mathbf{Y}=\mathbf{y}$ used in conditioning, and thus represents the expression for the variance of prediction errors.

These results extend to $\mathbf{W}$ being a vector with a general expected squared loss function $E\left[(\hat{\mathbf{W}}-\mathbf{W})'\mathbf{A}^{-1}(\hat{\mathbf{W}}-\mathbf{W})\right]$ ($\mathbf{A}$ being any positive definite symmetric matrix) as the criterion to be minimized (Goffinet, 1983; Searle et al., 1992).

The next example will illustrate the previous result regarding the linear form of BP which arises not only for continuous variables but also, in some cases, for binary ones.

Example 2.2 *Best Prediction when W and Y are binary variables*

We are assuming here that W and Y are binary variables, the joint distribution of which is defined as $p_{ij}=\Pr(W=i;Y=j)$ for $i,j=0,1$, namely

Table 2.1. Notations for the 2x2 contingency table

	$Y=0$	$Y=1$
$W=0$	p_{00}	p_{01}
$W=1$	p_{10}	p_{11}

Here again, the best prediction $\hat{w}=E(W\mid Y=y)$ can be expressed as a linear function of y, $\hat{w}=\alpha+\beta y$. This can be shown as follows. As Y is a binary variable, the conditional expectation of W given $Y=y$ is simply $E(W\mid Y=y)=yE(W\mid Y=1)+(1-y)E(W\mid Y=0)$ i.e.

$$\hat{w}=E(W\mid Y=0)+\left[E(W\mid Y=1)-E(W\mid Y=0)\right]y. \qquad (2.15)$$

Now,

$$E(W\mid Y=0)=\Pr(W=1\mid Y=0)=p_{10}/(p_{00}+p_{10})$$

$$E(W\mid Y=1)=\Pr(W=1\mid Y=1)=p_{11}/(p_{01}+p_{11}),$$

so that

$$\alpha=p_{10}/(p_{00}+p_{10}), \qquad (2.16)$$

$$\beta = \frac{p_{00}p_{11} - p_{10}p_{01}}{(p_{00} + p_{10})(p_{01} + p_{11})}. \qquad (2.17)$$

Notice that $1-\alpha = \Pr(W = 0 \mid Y = 0)$, and $\alpha + \beta = \Pr(W = 1 \mid Y = 1)$ represent respectively what is referred to as negative and positive predictive values of W based on Y.

The next step is to compute the MSE of this predictor i.e., $MSE = E_Y\left[Var(W \mid Y = y)\right]$. Remember, W being binary, $E(W) = E(W^2)$ and the variance can be obtained as $Var(W) = E(W)\left[1 - E(W)\right]$. Here, this gives

$$Var(W \mid Y = 0) = \frac{p_{00}p_{10}}{(p_{00} + p_{10})^2},$$

$$Var(W \mid Y = 1) = \frac{p_{01}p_{11}}{(p_{01} + p_{11})^2}.$$

Weighing these two terms by $\Pr(Y = 0) = p_{00} + p_{10}$, and $\Pr(Y = 1) = p_{01} + p_{11}$ respectively, the expression for the MSE of the predictor can be derived as

$$E\left[(\hat{W} - W)^2\right] = \frac{p_{00}p_{10}}{p_{00} + p_{10}} + \frac{p_{01}p_{11}}{p_{01} + p_{11}}. \qquad (2.18)$$

One can also derive the formula for the R^2. Here, since the predictor is linear, it can be established directly as the square of the correlation between W and Y (see exercise 2.1) which yields the following expression

$$R^2 = \frac{(p_{00}p_{11} - p_{10}p_{01})^2}{(p_{00} + p_{01})(p_{10} + p_{11})(p_{00} + p_{10})(p_{01} + p_{11})}, \qquad (2.19)$$

Actually, there is a connection between R^2 and the chi-square since (2.19) applied to observed frequencies turns out to be equal to N (total number of observations) times the Pearson chi-square statistics (Fienberg, 1985, page 12).

These formulae can be applied to the diagnosis of the *Clamydia trachomatis* infection of the cercix based on ligase chain reaction for which the following data have been observed (Schachter et al., 1994, table 2),

Table 2.2. Absolute frequencies observed in *C. Trachomatis*

	$Y=0$	$Y=1$
$W=0$	1896	84
$W=1$	13	139

Assuming relative frequencies equal to probabilities, one obtains $\alpha=0.0068$, $\beta=0.6165$ and $R^2=0.54$. The relatively low value of R^2 is due mainly to a poor positive predictive value of the test $\alpha+\beta=0.6233$ resulting in a high proportion of false positives whereas its negative predictive value $1-\alpha$ is very close to one.

2.2.2 Best Linear Prediction (BLP)

As pointed out previously, deriving the BP requires the knowledge of the joint distribution of the predictand and observations. We will relax this restrictive assumption but we still assume that the first two moments are known.

Actually we will restrict our attention a priori to a particular class of predictors, namely for convenience the linear one of the form $\hat{W}=a_0+\mathbf{a}'\left(\mathbf{Y}-\boldsymbol{\mu}_Y\right)$ where a_0 and $\mathbf{a}=\left(a_i\right)$ for $1\le i\le N$ are the coefficients to be determined.

In such conditions,

$$E\left(\hat{W}-W\right)=a_0-\mu_W,\ Var\left(\hat{W}-W\right)=\mathbf{a}'\boldsymbol{\Sigma}_{YY}\mathbf{a}-2\mathbf{a}'\boldsymbol{\Sigma}_{YW}+\Sigma_{WW}.$$

Letting $Q\left(a_0,\mathbf{a}\right)$ designate the MSE, minimizing it with respect to the unknowns a_0 and $\mathbf{a}$ involves solving

$$\frac{\partial Q\left(a_0,\mathbf{a}\right)}{\partial a_0}=2\left(a_0-\mu_W\right)=0, \qquad (2.20)$$

$$\frac{\partial Q\left(a_0,\mathbf{a}\right)}{\partial \mathbf{a}}=2\left(\boldsymbol{\Sigma}_{YY}\mathbf{a}-\boldsymbol{\Sigma}_{YW}\right)=0. \qquad (2.21)$$

Equation (2.20) leads to $a_0=\mu_W$ indicating that the predictor we obtain is unbiased. Notice that this is a property of the BLP not an imposed condition.

The solutions to (2.21) is $\mathbf{a} = \boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YW}$ so that eventually the predictor can be written as

$$\hat{W} = \mu_W + \boldsymbol{\Sigma}_{WY}\boldsymbol{\Sigma}_{YY}^{-1}\left(\mathbf{Y} - \boldsymbol{\mu}_Y\right). \qquad (2.22)$$

This predictor is the same as BP under normality, but whereas linearity was a property of the predictor in the Gaussian case, here this is a pre-assigned feature. As for BP, it is easily seen from (2.22) that $\mathrm{Cov}(\hat{W}, W) = \mathbf{a}'\boldsymbol{\Sigma}_{YW}$ is equal to $Var\left(\hat{W}\right) = \mathbf{a}'\boldsymbol{\Sigma}_{YY}\mathbf{a}$, so that properties (2.9), (2.10) and (2.11) still hold for BLP in (2.22).

2.2.3 Best Linear Unbiased Prediction (BLUP)

This type of predictor is now universally known under the acronym of BLUP; it was originally proposed by Goldberger (1962) but its main features were derived and applications implemented by Charles R. Henderson, his students and disciples (DA Harville, RL Quaas, LR Schaeffer) at Cornell University.

This predictor was developed to extend the prediction of W based on $\mathbf{Y}$ when their first moments are unknown. In fact, Henderson restricted the problem to the case of expectations expressed as linear functions of some p-dimensional vector $\boldsymbol{\beta}$ of unknown coefficients.

More specifically, it is assumed that $\mu_W = \mathbf{k}'\boldsymbol{\beta}$ and $\boldsymbol{\mu}_Y = \mathbf{X}\boldsymbol{\beta}$ where $\mathbf{k}'\boldsymbol{\beta}$ is any linear estimable function of the parameter vector $\boldsymbol{\beta}$. Actually, this is tantamount to the problem of prediction within a linear mixed model framework. Data are supposed to be generated according to such a mixed model structure $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ with $\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$, $\mathbf{u} \perp \mathbf{e}$, and the predictand is formulated as a linear combination of fixed and random effects, say $W = \mathbf{k}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{u}$.

Classical derivation

We are looking for a predictor $\hat{W}$ which

i) belongs a priori to the class of linear predictors,

ii) is unbiased in the sense $\mathrm{E}(\hat{W}) = \mathrm{E}(W)$,

iii) minimises the MSE.

These three conditions can be expressed as follows

i) $\hat{W} = \mathbf{a}'\mathbf{Y}$,

ii) $\mathbf{a}'\mathbf{X} - \mathbf{k}' = \mathbf{0}$,

iii) $\mathrm{Var}\left(\hat{W} - W\right) = \mathbf{a}'\mathbf{V}\mathbf{a} + \mathbf{m}'\mathbf{G}\mathbf{m} - 2\mathbf{a}'\mathbf{C}\mathbf{m}$ minimum,

where $\mathbf{C}=Cov(\mathbf{Y},\mathbf{u}')=\mathbf{ZG}$.

Identity ii) comes from setting the condition $E(\hat{W}) = E(W)$ which should be true for any $\boldsymbol{\beta}$, and iii) from the expression of MSE under the condition of unbiasedness.

Minimizing iii) with respect to $\mathbf{a}$ under the condition ii) is equivalent to minimizing the following function

$$Q(\mathbf{a},\boldsymbol{\theta}) = \mathbf{a}'\mathbf{V}\mathbf{a} - 2\mathbf{a}'\mathbf{C}\mathbf{m} + 2\boldsymbol{\theta}'(\mathbf{X}'\mathbf{a} - \mathbf{k}), \qquad (2.23)$$

where $\boldsymbol{\theta}$ is a $(p \times 1)$ vector of Lagrange multipliers.

Differentiation of $Q(\mathbf{a},\boldsymbol{\theta})$ with respect to $\mathbf{a}$ and $\boldsymbol{\theta}$ gives

$$\frac{\partial Q(\mathbf{a},\boldsymbol{\theta})}{\partial \mathbf{a}} = 2\mathbf{V}\mathbf{a} - 2\mathbf{C}\mathbf{m} + 2\mathbf{X}\boldsymbol{\theta}, \qquad (2.24)$$

$$\frac{\partial Q(\mathbf{a},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\left(\mathbf{X}'\mathbf{a} - \mathbf{k}\right). \qquad (2.25)$$

Setting (2.25) to zero leads to $\mathbf{a} = \mathbf{V}^{-1}(\mathbf{Cm} - \mathbf{X\theta})$, and substituting this value in (2.25) gives $\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Cm} - \mathbf{X\theta}) = \mathbf{k}$. Solving this system in $\boldsymbol{\theta}$ and substituting the solution into the expression of $\mathbf{a}$, we obtain $\mathbf{a}' = (\mathbf{m}'\mathbf{C}' - \boldsymbol{\theta}'\mathbf{X}')\mathbf{V}^{-1}$ so that

$$\hat{W} = \mathbf{a}'\mathbf{Y} = \mathbf{m}'\mathbf{C}'\mathbf{V}^{-1}\mathbf{Y} + \left(\mathbf{k}' - \mathbf{m}'\mathbf{C}'\mathbf{V}^{-1}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},$$

which, after some rearrangement becomes

$$\hat{W} = \mathbf{k}'\hat{\boldsymbol{\beta}} + \mathbf{m}'\hat{\mathbf{u}}. \qquad (2.26)$$

In (2.26) $\hat{\boldsymbol{\beta}}$ is a GLS solution to the system

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \qquad (2.27)$$

and, $\hat{\mathbf{u}}$ is the BLUP of $\mathbf{u}$ based on $\mathbf{Y}$ given by

$$\hat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right). \qquad (2.28)$$

Notice that the BLUP of $\hat{\mathbf{u}}$ can be obtained in practice as the BLP of $\mathbf{u}$ i.e. $\hat{\mathbf{u}}_{BLP} = E(\mathbf{u}) + Cov(\mathbf{u}, \mathbf{Y}')\left[Var(\mathbf{Y})\right]^{-1}\left[\mathbf{Y} - \mathrm{E}(\mathbf{Y})\right]$ in which $\mathrm{E}(\mathbf{Y})$ is replaced by its GLS estimator $\mathbf{X}\hat{\boldsymbol{\beta}}$. This result was established independently by Goldberger (1962) (page 371, eq 3.13) and Henderson (1963) (page 161, equations 19 and 20).

Example 2.3 *BLP and BLUP of a random intercept*

These two procedures can be illustrated with the simple random intercept model:

$$y_{ij} = \mu + a_i + e_{ij}, \qquad (2.29)$$

where y_{ij} is the j^{th} record $(j = 1,...,n_i)$ on the i^{th} experimental unit $(i = 1,...,I)$ such as a cluster (e.g. a family) made of several subunits; μ is the population mean; a_i is the true effect of the i^{th} experimental unit , and e_{ij} is the random error.

Since experimental units $i = 1,...,I$ are supposed to be randomly sampled from a population, they are modelled as random variables with expectation zero, variance σ_a^2 and zero correlation among them i.e. $a_i \sim_{iid} \left(0, \sigma_a^2\right)$; similarly, it is assumed that $e_{ij} \sim_{iid} \left(0, \sigma_e^2\right)$, and in addition, that any a_i is uncorrelated to any error term, $Cov\left(a_i, e_{i'j'}\right) = 0$, $\forall i, i', j'$.

We are interested in predicting the a_i's, and shall first use BP to that respect. Notice that the couples $\left(a_i, y_{i.}\right)$ are uncorrelated among them, so that all the information about a_i is in the data sample mean $y_{i.} = \left(\sum_{j=1}^{n_i} y_{ij}\right) / n_i$ pertaining to the same experimental unit i. Then, the BP of a_i based on $y_{i.}$ is given by the equation of linear regression of a_i in $y_{i.}$ that is,

$$\hat{a}_i = E(a_i) + \frac{Cov(a_i, y_{i.})}{Var\left(y_{i.}\right)}\left(y_{i.} - \mu\right)$$

Here $E\left(a_i\right) = 0$, $Cov\left(a_i, y_{i.}\right) = \sigma_a^2$ and $Var(y_{i.}) = \sigma_a^2 + \left(\sigma_e^2 / n_i\right)$, so that the BLP $\hat{a}_i$ of a_i is

$$\hat{a}_i = b_i\left(y_{i.} - \mu\right). \qquad (2.30)$$

where $b_i = n_i\sigma_a^2 / \left(n_i\sigma_a^2 + \sigma_e^2\right)$ acts as a skrinkage factor $(0 \le b_i \le 1)$ of $y_{i.} - \mu$ towards zero.

If $n_i \to \infty$ or $\sigma_e^2 \to 0$, then $b_i \to 1$ whereas $b_i \to 0$ when $n_i \to 0$ or $\sigma_a^2 \to 0$.

The result in (2.30) can also be derived from the general formula of BP given in (2.22) as illustrated by exercise (2.2).

Next, one can obtain the BLUP of a_i from (2.30) by replacing μ by its GLS estimator $\hat{\mu}$ (see exercise 1.13). If one wants to predict a future observation $y_{ij'}$ which is not recorded, its BLUP is then $\hat{y}_{ij'} = \hat{\mu} + b_i\left(y_{i.} - \hat{\mu}\right)$ which clearly

illustrates that this predictor is "shrinking" the least square mean $y_{i.}$ of $\mu + a_i$ towards the GLS estimation $\hat{\mu}$ of the population mean μ.

Bulmer's derivation

Bulmer (1980) tackled the prediction of $\mathbf{u}$ not from the data vector $\mathbf{Y}$ which is causing the difficulty, but from an counterpart of it $\mathbf{Y}_c$ adjusted for fixed effects. Technically, one proceeds into two steps:

i) adjust $\mathbf{Y}$ for fixed effects via GLS i.e. compute $\mathbf{Y}_c = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$

ii) predict $\mathbf{u}$ by BLP based on $\mathbf{Y}_c$ which is legitimate since both $\mathrm{E}(\mathbf{u})$ and $\mathrm{E}(\mathbf{Y}_c)$ are equal to zero, and so exactly known.

By definition, such a predictor, say $\tilde{\mathbf{u}}$ should be of the form

$$\tilde{\mathbf{u}} = \mathrm{Cov}(\mathbf{u}, \mathbf{Y}_c^{'})\left[\mathrm{Var}(\mathbf{Y}_c)\right]^{-} \mathbf{Y}_c . \qquad (2.31)$$

Notice that this expression includes a generalized inverse of $\mathrm{Var}(\mathbf{Y}_c)$ instead of a standard inverse since $\mathrm{Var}(\mathbf{Y}_c)$ no longer has rank N but only $N - \mathrm{rank}(\mathbf{X})$. In fact, $\mathbf{Y}_c$ can be expressed as

$$\mathbf{Y}_c = \mathbf{V}\underline{\mathbf{P}}\mathbf{Y} , \qquad (2.32)$$

where $\underline{\mathbf{P}}$ is defined as $\underline{\mathbf{P}} = \mathbf{V}^{-1}(\mathbf{I} - \mathbf{Q})$ and $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ is the $\mathbf{V}^{-1}$ orthogonal projector defined in (1.72). Then, $\mathbf{V}$ being a g-inverse of $\underline{\mathbf{P}}$ (see exercise 2.3) ,

$$\mathrm{Var}(\mathbf{Y}_c) = \mathbf{V}\underline{\mathbf{P}}\mathbf{V}\underline{\mathbf{P}}\mathbf{V} = \mathbf{V}\underline{\mathbf{P}}\mathbf{V} ,$$

and

$$\mathrm{Cov}(\mathbf{u}, \mathbf{Y}_c^{'}) = \mathbf{C}'\underline{\mathbf{P}}\mathbf{V} , \qquad (2.33)$$

$$(\mathbf{V}\underline{\mathbf{P}}\mathbf{V})^{-} = \mathbf{V}^{-1} \qquad (2.34)$$

After substituting, (2.32), (2.33), and (2.34) into (2.31),

$$\tilde{\mathbf{u}} = \mathbf{C}'\underline{\mathbf{P}}\mathbf{Y}_c = \mathbf{C}'\underline{\mathbf{P}}\mathbf{V}\underline{\mathbf{P}}\mathbf{Y} = \mathbf{C}'\underline{\mathbf{P}}\mathbf{Y}, \qquad (2.35)$$

Since $\underline{\mathbf{P}}\mathbf{Y} = \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$, it becomes obvious that this predictor is identical to the BLUP of $\mathbf{u}$ based on $\mathbf{Y}$ (Gianola and Goffinet, 1982) contrarily to what Bulmer (1980) initially thought. In addition, formula (2.35) illustrates the property of translation invariance of BLUP (see exercise 2.3).

This formula is also convenient for deriving the expression of the variance of the predictor and that of the error of prediction. Indeed

$$Var(\hat{\mathbf{u}}) = Cov(\hat{\mathbf{u}}, \mathbf{u}') = \mathbf{C}'\underline{\mathbf{P}}\mathbf{C}, \qquad (2.36)$$

and, as $\hat{\mathbf{u}}$ and $\hat{\mathbf{u}} - \mathbf{u}$ are thus uncorrelated

$$Var(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{G} - \mathbf{C}'\underline{\mathbf{P}}\mathbf{C}. \qquad (2.37)$$

It is worthwhile mentioning the difference between this expression of the variance of prediction error using BLUP and that would apply to BLP

$$Var(\hat{\mathbf{u}}_{BLP} - \mathbf{u}) = \mathbf{G} - \mathbf{C}'\mathbf{V}^{-1}\mathbf{C}. \qquad (2.38)$$

In (2.37), $\underline{\mathbf{P}}$ replaces $\mathbf{V}^{-1}$ so as to account for the uncertainty due to estimating $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ via GLS instead of assuming $\boldsymbol{\beta}$ known. We will find the same substitution when passing from maximum likelihood to restricted maximum likelihood.

Formulae (2.27), (2.28) and (2.37) apply directly to the case of the analysis of repeated measurements, $\mathbf{y} = \left(\mathbf{y}_1', ..., \mathbf{y}_i', ..., \mathbf{y}_I'\right)'$ recorded on independent individuals $i = 1, ..., I$ according to the model $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i$, where $\mathbf{y}_i = \left(y_{ij}\right)$, $1 \le j \le n_i$ represents the data vector for subject i; $\mathbf{X}_i\boldsymbol{\beta}$ refers to the fixed part of the model and $\mathbf{Z}_i\mathbf{u}_i$ to its subject specific random component such that

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i, \qquad (2.39)$$

where $\mathbf{u}_i \sim_{iid} (\mathbf{0}, \mathbf{G}_0)$, $\mathbf{e}_i \sim_{iid} (\mathbf{0}, \mathbf{R}_{0,i})$ and $\mathbf{u}_i \perp \mathbf{e}_i$. For instance, $\mathbf{u}_i$ can include an intercept and a slope, $\mathbf{u}_i = (a_i, b_i)'$ as in Example 1.6.

Letting $\mathbf{V}_i = \mathbf{R}_{0,i} + \mathbf{Z}_i \mathbf{G}_0 \mathbf{Z}_i'$, then the BLUP of $\mathbf{u}_i$ is

$$\hat{\mathbf{u}}_i = \mathbf{G}_0 \mathbf{Z}_i' \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} \right), \tag{2.40}$$

where $\hat{\boldsymbol{\beta}}$ is solution of

$$\left(\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right) \hat{\boldsymbol{\beta}} = \sum_{i=1}^{I} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i . \tag{2.41}$$

The corresponding variance of prediction errors (VPE) is

$$Var\left(\hat{\mathbf{u}}_i - \mathbf{u}_i \right) = \mathbf{G}_0 - \mathbf{G}_0 \mathbf{Z}_i' \underline{\mathbf{P}}_i \mathbf{Z}_i \mathbf{G}_0 , \tag{2.42}$$

where

$$\underline{\mathbf{P}}_i = \mathbf{V}_i^{-1} - \mathbf{V}_i^{-1} \mathbf{X}_i \left(\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-} \mathbf{X}_i' \mathbf{V}_i^{-1} . \tag{2.43}$$

Finally, the BLUP $\hat{\mathbf{y}}_i^* = (y_{ik})$ of a vector of future (new) observations $\mathbf{y}_i^* = \mathbf{X}_i^* \boldsymbol{\beta} + \mathbf{Z}_i^* \mathbf{u}_i + \mathbf{e}_i^*$ for subject i with $k \notin \{1, ..., j, ..., n_i\}$, can be written as

$$\hat{\mathbf{y}}_i^* = \mathbf{X}_i^* \hat{\boldsymbol{\beta}} + \mathbf{Z}_i^* \mathbf{G}_0 \mathbf{Z}_i' \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} \right), \tag{2.44}$$

In the special case where $\mathbf{X}_i^* = \mathbf{X}_i$ and $\mathbf{Z}_i^* = \mathbf{Z}_i$ (e.g. no time dependent covariates), formula (2.44) can be alternatively expressed as

$$\hat{\mathbf{y}}_i^* = \mathbf{R}_{0,i} \mathbf{V}_i^{-1} \mathbf{X}_i \hat{\boldsymbol{\beta}} + \left(\mathbf{I}_{n_i} - \mathbf{R}_{0,i} \mathbf{V}_i^{-1} \right) \mathbf{y}_i . \tag{2.45}$$

Formulae (2.45) clearly illustrates the James-Stein interpretation of BLUP as a simple weighted mean of $\mathbf{X}_i \hat{\boldsymbol{\beta}}$ and the data vector $\mathbf{y}_i$.

We leave up to the reader to see how to derive (2.45) and the corresponding variance of prediction errors $Var\left(\hat{\mathbf{y}}_i^* - \mathbf{y}_i^* \right)$ (exercise 2.5)

2.3 Mixed model equations

2.3.1 Henderson's approach

One major difficulty in using (2.40) to obtain BLUP lies in computing the inverse of the variance covariance matrix of the observations that cannot be simplified in complex data structures with $\mathbf{V}$ being large and non block diagonal. This led Charles Henderson (1948, 1950, 1952) to propose as an alternative a set of equations, known now as Henderson's mixed model equations (HMME), providing both BLUP of random effects and GLS of fixed effects, and which are

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \qquad (2.46).$$

Here $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ are solutions of the equations for fixed $\boldsymbol{\beta}$ and random $\mathbf{u}$ effects arising in a linear mixed model $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}+\mathbf{e}$ such that $E(\mathbf{y})=\mathbf{X}\boldsymbol{\beta}$, $Var(\mathbf{y})=\mathbf{V}=\mathbf{Z}\mathbf{G}\mathbf{Z}'+\mathbf{R}$ with $\mathbf{u}\sim(\mathbf{0},\mathbf{G})$, $\mathbf{e}\sim(\mathbf{0},\mathbf{R})$ and $Cov(\mathbf{u},\mathbf{e}')=\mathbf{0}$.

This system does not require inversion of $\mathbf{V}$; it only requires inversion of $\mathbf{R}$ and $\mathbf{G}$ which are often diagonal or block diagonal, or if not have special structure making inversion much more easier than that of $\mathbf{V}$. Actually, this system mimics the normal equations of least squares except that $\mathbf{G}^{-1}$ is added to the block pertaining to random effects.

2.3.2 Justification

Henderson derived these equations by maximizing the joint density $f(\mathbf{y},\mathbf{u})$ of the data and random effects with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ under the normality assumption (Henderson et al., 1959; Henderson, 1973),

$$\tilde{\boldsymbol{\beta}},\tilde{\mathbf{u}} = \arg\max_{\boldsymbol{\beta},\mathbf{u}} \log f(\mathbf{y},\mathbf{u}). \qquad (2.47)$$

As $f(\mathbf{y},\mathbf{u})=f(\mathbf{y}|\mathbf{u})f(\mathbf{u})$ where $\mathbf{y}|\mathbf{u}\sim\mathcal{N}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u},\mathbf{R})$ and $\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{G})$, one has

$$-2\log f(\mathbf{y}|\mathbf{u}) = N\log 2\pi + \log|\mathbf{R}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})$$

$$-2\log f(\mathbf{u}) = q\log 2\pi + \log\left|\mathbf{G}\right| + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u}\,.$$

Maximizing $l(\boldsymbol{\beta},\mathbf{u};y) = \log f(\mathbf{y},\mathbf{u})$ is equivalent to minimizing the sum of these two terms. Differentiating this sum with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$,

$$\frac{\partial\left[-2l(\boldsymbol{\beta},\mathbf{u};y\right]}{\partial\boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})$$

$$\frac{\partial\left[-2l(\boldsymbol{\beta},\mathbf{u};y\right]}{\partial\mathbf{u}} = -2\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) + 2\mathbf{G}^{-1}\mathbf{u}\,,$$

and equating them to zero gives the following system

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\tilde{\mathbf{u}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}\,, \qquad (2.48)$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + \left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)\tilde{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}\,, \qquad (2.49)$$

that coincides precisely with HMME.

However, one will observe that neither $l(\boldsymbol{\beta},\mathbf{u};y)$ is the formal expression of the likelihood function of the data, nor $\mathbf{u}$ is a parameter so that Henderson's derivation of the MME has been perceived for a long time as strange if not dubious.

Happily, as will be seen later on, this maximization can be perfectly justified within a Bayesian framework (see section 2.3.4 and exercise 2.10). But before that, what can be done is to show that the solutions to HMME in $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ coincide with their $\hat{\boldsymbol{\beta}}$ GLS and $\hat{\mathbf{u}}$ BLUP counterparts respectively.

The first equation can be rewritten as $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{Z}\tilde{\mathbf{u}}\right)$. Similarly for the second: $\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\right)\tilde{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)$. Substituting the expression of $\tilde{\mathbf{u}}$ from this last equation into the first one leads to

$$\mathbf{X}'\mathbf{W}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{W}\mathbf{y}\,, \qquad (2.50)$$

where

$$\mathbf{W} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}. \qquad (2.51)$$

Now, it can be shown (see appendix 2.5) that:

$$\mathbf{W} = \left(\mathbf{Z}'\mathbf{G}\mathbf{Z} + \mathbf{R}\right)^{-1} = \mathbf{V}^{-1}, \qquad (2.52)$$

so that $\tilde{\boldsymbol{\beta}}$ is a GLS solution for $\boldsymbol{\beta}$.

Rewriting equation (2.49) as

$$\tilde{\mathbf{u}} = \left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \qquad (2.53)$$

and, using in (2.53) the identity derived in the appendix i.e.

$$\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}, \qquad (2.54)$$

one proves that $\tilde{\mathbf{u}}$ of (2.57) is eventually the BLUP of $\mathbf{u}$.

2.3.3 Sampling and prediction error variances

We are concerned here with the sampling variances of estimable $\mathbf{k}'\boldsymbol{\beta}$ of fixed effects as well as variances of prediction errors $\hat{\mathbf{u}} - \mathbf{u}$ of random effects.
Let $\mathbf{C}$ be an inverse of the coefficient matrix of HMME,

$$\begin{pmatrix} \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u} \\ \mathbf{C}_{u\beta} & \mathbf{C}_{uu} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}^{-1} \qquad (2.55)$$

For the sake of simplicity, the coefficient matrix is supposed here to be full rank so that $\mathbf{C}$ is a regular inverse, but results obtained apply as well to any generalized inverse.

It will be shown that all expressions for accuracy of fixed effects and BLUP can be obtained directly from the HMME as follows (Henderson, 1973)

$$Var(\mathbf{k}'\hat{\boldsymbol{\beta}}) = \mathbf{k}'\mathbf{C}_{\beta\beta}\mathbf{k}\,, \tag{2.56}$$

$$Cov(\mathbf{k}'\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}') = \mathbf{0}\,, \tag{2.57}$$

$$Cov\left[\mathbf{k}'\hat{\boldsymbol{\beta}}, (\hat{\mathbf{u}} - \mathbf{u})'\right] = \mathbf{k}'\mathbf{C}_{\beta u}\,, \tag{2.58}$$

$$Var(\hat{\mathbf{u}}) = \mathbf{G} - \mathbf{C}_{uu}\,. \tag{2.59}$$

$$Var(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{uu}\,. \tag{2.60}$$

Proof

Formula (2.56) comes directly from the expression of the inverse of a partitioned matrix. In fact, from (2.55),

$$\mathbf{C}_{\beta\beta} = \left[\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\right]^{-1}$$

which, according to (2.51) and (2.52) reduces to $\mathbf{C}_{\beta\beta} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$, QED.

Formula (2.57) is a consequence of the orthogonality of $\mathbf{Q}$ and $\mathbf{I} - \mathbf{Q}$. $\mathbf{k}'\boldsymbol{\beta}$ being an estimable function, it can be expressed as a linear combination of the data expectation i.e., $\mathbf{k}'\boldsymbol{\beta} = \boldsymbol{\lambda}'\mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\lambda}$, so that $\mathbf{k}'\hat{\boldsymbol{\beta}} = \boldsymbol{\lambda}'\mathbf{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\lambda}'\mathbf{Q}\mathbf{y}$. In addition, as shown in (2.35), $\hat{\mathbf{u}} = \mathbf{C}'\underline{\mathbf{P}}\mathbf{y}$; then $Cov(\mathbf{k}'\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}') = \boldsymbol{\lambda}'\mathbf{Q}\mathbf{V}\underline{\mathbf{P}}\mathbf{C}$. Now $\mathbf{Q}\mathbf{V}\underline{\mathbf{P}} = \mathbf{Q}\mathbf{V}\left[\mathbf{V}^{-1}(\mathbf{I} - \mathbf{Q})\right]$ which is equal to zero since $\mathbf{Q}$ and $\mathbf{I} - \mathbf{Q}$ are orthogonal, and $Cov(\mathbf{k}'\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}') = \mathbf{0}$, QED.

Given (2.57), formula (2.58) is equivalent to $Cov\left(\mathbf{k}'\hat{\boldsymbol{\beta}}, \mathbf{u}'\right) = -\mathbf{k}'\mathbf{C}_{\beta u}$. By definition of HMME, $\hat{\mathbf{u}}$ can be formulated as $\mathbf{k}'\begin{pmatrix}\mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u}\end{pmatrix}\begin{pmatrix}\mathbf{X}' \\ \mathbf{Z}'\end{pmatrix}\mathbf{R}^{-1}\mathbf{y}$ so that

$$Cov\left(\mathbf{k}'\hat{\boldsymbol{\beta}}, \mathbf{u}'\right) = \mathbf{k}'\begin{pmatrix}\mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta u}\end{pmatrix}\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\end{pmatrix}\mathbf{G}\,.$$

By definition of $\mathbf{C}$, one has $\left(\mathbf{C}_{\beta\beta} \quad \mathbf{C}_{\beta u}\right)\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{pmatrix}=\mathbf{0}$. Hence $Cov\left(\mathbf{k}'\hat{\boldsymbol{\beta}},\mathbf{u}'\right)=-\mathbf{k}'\mathbf{C}_{\beta u}\mathbf{G}^{-1}\mathbf{G}$ which completes the proof.

If $\hat{\mathbf{u}}$ is BLUP of $\mathbf{u}$, it follows that $Var(\hat{\mathbf{u}})=Cov(\hat{\mathbf{u}},\mathbf{u}')$ (see (2.36)). One can write $\hat{\mathbf{u}}$ as previously from HMME, $\hat{\mathbf{u}}=\left(\mathbf{C}_{u\beta} \quad \mathbf{C}_{uu}\right)\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\end{pmatrix}\mathbf{R}^{-1}\mathbf{y}$, and consequently $Cov\left(\hat{\mathbf{u}},\mathbf{u}'\right)=\left(\mathbf{C}_{u\beta} \quad \mathbf{C}_{uu}\right)\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\end{pmatrix}\mathbf{G}$, which, on account of the property of $\mathbf{C}$ being an inverse of the coefficient matrix, is also equal to $\left(\mathbf{I}-\mathbf{C}_{uu}\mathbf{G}^{-1}\right)\mathbf{G}$, QED. Formula (2.60) results from (2.59) and the fact that $Var(\mathbf{u})$ is the sum of $Var(\hat{\mathbf{u}})$ and $Var(\hat{\mathbf{u}}-\mathbf{u})$.

Finally, if interest is on a linear combination of fixed and random effects, $W=\mathbf{k}'\boldsymbol{\beta}+\mathbf{m}'\mathbf{u}$, its variance of prediction errors is obtained as

$$Var\left(\hat{W}-W\right)=\mathbf{k}'\mathbf{C}_{\beta\beta}\mathbf{k}+\mathbf{m}'\mathbf{C}_{uu}\mathbf{m}-2\mathbf{k}'\mathbf{C}_{\beta u}\mathbf{m}. \qquad (2.61)$$

Example 2.4. *Prediction in a two-way additive mixed model*

We are concerned here by cross-classified designs involving two factors, say A and B, and response data y_{ijk} which can be presented as a tabular layout with the levels of one factor (say A) being rows, and the levels of the other (say B) being columns, each elementary combination ij having n_{ij} observations. Examples of such designs are:

-in agriculture: yield per ha according to variety of plant (A) in different fields (B);

-in breeding and genetics: milk production per lactation of cows raised in different herds and/or environmental conditions (A) and sired by different bulls (B);

-in management: mileage per gallon or litres per 100km according to car type (A) with different drivers (B);

-in clinical trials: health status of patients according to treatment or medication (A) in different hospitals (B);

Such designs can be analyzed under three different situations: i) both factors as fixed (see exercise 1.4) , ii) both factors as random, and iii) one factor fixed and the other random.

Here we will concentrate on the last situation. This means that effects of the random factor (B in our convention) are assumed to be sampled from a conceptual population of effects according to some randomization process. For instance, in the agriculture example, specific varieties are compared when grown on fields randomly drawn from a collection of fields having some specified characteristics (size, soil, etc…). Similarly, we will consider in breeding a random sample of bulls out of a given population (breed, country, age,…) with daughters raised in specific herds. The same reasoning will apply to a random sample of workers, drivers and hospitals.

Finally, for the sake of simplicity, we will assume that the effects of factors A and B are additive so that the model is written as:

$$y_{ijk} = \mu + a_i + b_j + e_{ijk}, \qquad (2.62)$$

where y_{ijk} is the k^{th} response ($k = 1,...,n_{ij}$) obtained in the ij combination of the levels of factors A and B ; μ is the overall mean; a_i is the fixed effect of the i^{th} level of factor A ($i = 1,...,I$); b_j is the random effect of the j^{th} level of factor B ($j = 1,...,J$); e_{ijk} is the residual term.

Classical assumptions are made about the distribution of random effects of this model, viz. $b_j \sim_{iid} \mathcal{N}\left(0, \sigma_b^2\right)$, $e_{ijk} \sim_{iid} \mathcal{N}\left(0, \sigma_e^2\right)$, and $b_j \perp e_{ijk}$, $\forall i$, $\forall j$, $\forall k$.

Suppose that at this stage, the interest is predicting the b_j values by BLUP. Here, $\mathbf{R}=\sigma_e^2\mathbf{I}_N$ with $N=\sum_{ij}n_{ij}$, and $\mathbf{G}=\sigma_b^2\mathbf{I}_q$ so that we can multiply both sides of HMME by σ_e^2 to form the following system

$$\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\lambda\mathbf{I}_q\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\end{pmatrix}, \tag{2.63}$$

where $\lambda=\sigma_e^2/\sigma_b^2$.

In the case of model (2.62), this gives

$$\begin{pmatrix}n_{10} & 0 & n_{11} & n_{11}\\ 0 & n_{i0} & n_{i1} & n_{ij}\\ n_{11} & n_{i1} & n_{01}+\lambda & 0\\ n_{11} & n_{ij} & 0 & n_{0j}+\lambda\end{pmatrix}\begin{pmatrix}\hat{\mu}+\hat{a}_1\\ \hat{\mu}+\hat{a}_i\\ \hat{b}_1\\ \hat{b}_j\end{pmatrix}=\begin{pmatrix}y_{100}\\ y_{i00}\\ y_{010}\\ y_{0j0}\end{pmatrix} \tag{2.64}$$

where $n_{i0}=\sum_{j=1}^{J}n_{ij}$, $n_{0j}=\sum_{i=1}^{I}n_{ij}$, $y_{i00}=\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}y_{ijk}$, $y_{0j0}=\sum_{i=1}^{I}\sum_{k=1}^{n_{ij}}y_{ijk}$.

In some instances (I very large), one can calculate the $\hat{b}_j$'s from a reduced system after eliminating the $\hat{\mu}+\hat{a}_i$'s unknowns (process sometimes known as absorption). Then, starting from (2.63), the system becomes

$$\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}+\lambda\mathbf{I}_q\right)\hat{\mathbf{u}}=\mathbf{Z}'\mathbf{M}\mathbf{y}, \tag{2.65}$$

where $\mathbf{M}=\mathbf{I}_N-\mathbf{P}$ with $\mathbf{P}=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}$ being the projector of $\mathbf{y}$ onto the subspace generated by the columns of $\mathbf{X}$ (see 1.16). In the case of (2.64), one gets

$$\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}\right)_{jj}=n_{0j}-\sum_{i=1}^{I}n_{ij}^2/n_{i0}, \tag{2.66}$$

$$\left(\mathbf{Z}'\mathbf{M}\mathbf{Z}\right)_{jj'}=-\sum_{i=1}^{I}n_{ij}n_{ij'}/n_{i0}\ \text{ for }\ j\neq j', \tag{2.67}$$

$$\left(\mathbf{Z}'\mathbf{M}\mathbf{y}\right)_{j}=y_{0j0}-\sum_{i=1}^{I}n_{ij}y_{i00}/n_{i0}. \tag{2.68}$$

This system can be solved either by direct inversion of the coefficient matrix, or if J is very large, by specialized algorithms e.g., Gauss-Seidel or Jacobi. It is worth noticing that the coefficient matrix $\mathbf{Z}'\mathbf{M}\mathbf{Z}+\lambda\mathbf{I}_q$ is always full rank provided $\lambda>0$, In practice, this means that HMME always have some solution in $\hat{\mathbf{u}}$ regardless of the structure of data. In addition, these solutions verify the identity $\mathbf{1}'\hat{\mathbf{u}}=0$ (see exercise 2.7)

This can be illustrated numerically by the data set shown in table 2.3 involving birth weight of the progeny of three sires (1, 2, 3) born out of heifers (A1) or mature cows (A2).

Table 2.3. *Distribution of progeny according to sire and age of cows*

Sire	A1		A2	
	n	$\sum y$	n	$\sum y$
1	10	440	80	3880
2	4	175	16	720
3	40	1730		

Assuming that $\lambda=15$, Henderson's mixed model equations are

$$\begin{pmatrix} 54 & 0 & 10 & 4 & 40 \\ 0 & 96 & 80 & 16 & 0 \\ 10 & 80 & 105 & 0 & 0 \\ 4 & 16 & 0 & 35 & 0 \\ 40 & 0 & 0 & 0 & 55 \end{pmatrix} \begin{pmatrix} \hat{\mu}+\hat{a}_1 \\ \hat{\mu}+\hat{a}_2 \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{pmatrix} = \begin{pmatrix} 2345 \\ 4600 \\ 4320 \\ 895 \\ 1730 \end{pmatrix},$$

where μ is the general mean, a_1 (a_2) is the effect of heifer (mature) calving, and b_1, b_2 and b_3 are the sire effects. This system has the following solutions: $\hat{\mu}+\hat{a}_1=43.625$, $\hat{\mu}+\hat{a}_2=47.204$; $\hat{b}_1=1.048$, $\hat{b}_2=-0.964$ and $\hat{b}_3=-0.084$. Under its reduced form, the system is

$$\begin{pmatrix} 36.4815 & -14.0741 & -7.4074 \\ -14.0741 & 32.0370 & -2.9630 \\ -7.4074 & -2.9630 & 25.3704 \end{pmatrix} \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{pmatrix} = \begin{pmatrix} 52.4074 \\ -45.3704 \\ -7.0370 \end{pmatrix}$$

with the same sire solutions the sum of which is zero.

One might be intererested in what is going on if the data structure is slightly modified (see exercise 2.7).

2.3.4 Bayesian interpretation

Numerous studies have shown the links between BLUP theory and Bayesian statistics (Dempfle, 1977; Lefort, 1980, Gianola & Fernando, 1986). Bayesian analysis of linear models dates back to the seminal paper by Lindley and Smith (1972) based on a hierarchical Gaussian model with two levels and briefly summarized below (see sections 1.6.2 and 1.6.3 for more details)

$$\text{i) } \mathbf{y} \mid \boldsymbol{\theta}, \mathbf{R} \sim \mathcal{N}(\mathbf{T}\boldsymbol{\theta}, \mathbf{R}), \qquad (2.69)$$

$$\text{ii) } \boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\Omega} \sim \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Omega}). \qquad (2.70)$$

The first level (2.69) describes the sampling distribution of data given the location $\boldsymbol{\theta}$ and dispersion $\mathbf{R}$ parameters, whereas the second one (2.70) specifies the prior distribution of $\boldsymbol{\theta}$.

According to Bayes' theorem, the posterior density is proportional to the product of the conditional density of the data in i) and of the prior in ii). Since these two are assumed to be normal, the posterior belongs to the same family as the prior (known as "conjugacy" property), so that

$$f(\boldsymbol{\theta} \mid \mathbf{y}) \propto \exp(-Q/2) \qquad (2.71)$$

the kernel Q being simply the sum of the kernels of i) and ii), namely

$$Q = (\mathbf{y} - \mathbf{T}\boldsymbol{\theta})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{T}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{\Omega}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \qquad (2.72)$$

Let $\hat{\boldsymbol{\theta}}$ be the solution to the following system of equations

$$\left(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T}+\boldsymbol{\Omega}^{-1}\right)\hat{\boldsymbol{\theta}}=\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}+\boldsymbol{\Omega}^{-1}\boldsymbol{\theta}_0, \qquad (2.73)$$

the expression of Q can be rearranged (see exercise 2.8) as to contain a quadratic in $\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}$ viz.

$$\begin{aligned} Q=&\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)'\left(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T}+\boldsymbol{\Omega}^{-1}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)+\mathbf{y}'\mathbf{R}^{-1}\mathbf{y}\\ &-\hat{\boldsymbol{\theta}}'\left(\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}+\boldsymbol{\Omega}^{-1}\boldsymbol{\theta}_0\right)+\boldsymbol{\theta}_0'\boldsymbol{\Omega}^{-1}\boldsymbol{\theta}_0 \end{aligned}. \qquad (2.74)$$

The first term in (2.74) only is needed for the expression of the density of the posterior distribution, the rest contributing to the integration constant. This first term can be easily identified as a Gaussian kernel so that the posterior distribution is a normal one

$$\boldsymbol{\theta}\,|\,\mathbf{y}\sim\mathcal{N}\left(\hat{\boldsymbol{\theta}},\mathbf{C}\right), \qquad (2.75)$$

with expectation $\hat{\boldsymbol{\theta}}$ solution to the system (2.73) and variance covariance matrix $\mathbf{C}=\left(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T}+\boldsymbol{\Omega}^{-1}\right)^{-1}$ being the inverse of the coefficient matrix of the same system (see exercise 2.9 for another interpretation of $\hat{\boldsymbol{\theta}}$).

At this point, we can make the connection between this hierarchical Bayes approach and mixed model methodology by defining $\boldsymbol{\theta}=(\boldsymbol{\beta}',\mathbf{u}')'$ and $\mathbf{T}=(\mathbf{X},\mathbf{Z})$. In hierarchical modelling, no distinction is made between fixed and random effects so that we shall assume that each component vector has a Gaussian prior distribution i.e., $\boldsymbol{\beta}\sim\mathcal{N}(\boldsymbol{\beta}_0,\mathbf{B})$ and $\mathbf{u}\sim\mathcal{N}(\mathbf{u}_0,\mathbf{G})$ with the two components being independent. Then, $\boldsymbol{\Omega}^{-1}=\begin{pmatrix}\mathbf{B}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{pmatrix}$, $\boldsymbol{\theta}_0=\begin{pmatrix}\boldsymbol{\beta}_0\\ \mathbf{u}_0\end{pmatrix}$, and the system in (2.76) becomes

$$\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\mathbf{B}^{-1} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}+\mathbf{B}^{-1}\boldsymbol{\beta}_0\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}+\mathbf{G}^{-1}\mathbf{u}_0\end{pmatrix}. \qquad (2.76)$$

Moreover, on account of (2.78), we can exactly specify the posterior densities of the marginal distributions of $\boldsymbol{\beta}$ and $\mathbf{u}$ as

$$\boldsymbol{\beta} \mid \mathbf{y}, else \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \underline{\mathbf{C}}_{\beta\beta}\right), \tag{2.77}$$

$$\mathbf{u} \mid \mathbf{y}, else \sim \mathcal{N}\left(\hat{\mathbf{u}}, \underline{\mathbf{C}}_{uu}\right), \tag{2.78}$$

where $\underline{\mathbf{C}}_{\beta\beta}$ and $\underline{\mathbf{C}}_{uu}$ are the diagonal blocks pertaining to $\boldsymbol{\beta}$ and $\mathbf{u}$ in the inverse of the coefficient matrix in (2.79). "else" standing for all parameters ($\boldsymbol{\beta}_0, \mathbf{u}_0, \mathbf{B}$, $\mathbf{G}, \mathbf{R}$) involved in the conditional distributions.

In a Bayesian setting of mixed models, the distinction between fixed and random effects will be accomplished by specifying $\mathbf{B}^{-1} \to \mathbf{0}$ and $\mathbf{u}_0 = \mathbf{0}$. The last condition conveys the property of random effects being usually viewed in mixed linear models as centered random variables. Equating $\mathbf{B}^{-1}$ to zero is equivalent to setting a prior distribution for $\boldsymbol{\beta}$ with infinite variance so as to make this prior close to an uniform one and thus, non informative in some sense. By doing this in (2.76) and assuming in addition that $\mathbf{X}$ is full column rank, $\hat{\boldsymbol{\beta}}$ becomes the GLS estimator of $\boldsymbol{\beta}$ which can be alternatively interpreted as the expectation of the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{G}$ and $\mathbf{R}$. This allows to parallel the presentation of the GLS estimator of fixed effects in classical statistics $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \mathbf{C}_{\beta\beta}\right)$ versus its Bayesian counterpart $\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{G}, \mathbf{R} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \mathbf{C}_{\beta\beta}\right)$ where $\mathbf{C}_{\beta\beta}$ corresponds to $\underline{\mathbf{C}}_{\beta\beta}$ for $\mathbf{B}^{-1} \to \mathbf{0}$ and $\mathbf{u}_0 = \mathbf{0}$.

Similarly, BLUP of $\mathbf{u}$ turns out to be the expectation of the posterior distribution of $\mathbf{u}$ given $\mathbf{G}$ and $\mathbf{R}$, and the variance of this distribution $\mathbf{C}_{uu}$ is equivalent to the variance of prediction errors under the normality assumption

$$\hat{\mathbf{u}}_{BLUP} = E\left(\mathbf{u} \mid \mathbf{y}, \mathbf{G}, \mathbf{R}\right), \tag{2.79}$$

$$Var(\hat{\mathbf{u}}_{BLUP} - \mathbf{u}) = Var(\mathbf{u} \mid \mathbf{y}, \mathbf{G}, \mathbf{R}). \tag{2.80}$$

In addition, the Bayesian hierarchical modelling helps to understand Henderson's original justification of the mixed model equations in terms of a Bayesian-type approach (see exercise 2.10). Actually, maximizing the logarithm of $f(\mathbf{y},\mathbf{u}|\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ as done by Henderson is equivalent under normality to seeking the first moments of the posterior distribution $f(\boldsymbol{\beta},\mathbf{u}|\mathbf{y})$.

2.4 Discussion-Conclusion

This presentation has provided us with a rigorous conceptual framework for dealing with the problem of prediction. There are different types of prediction according to assumptions made on the joint distribution of the predictand (W) and the observations ($\mathbf{Y}$).

The expectation of the conditional distribution of W given $\mathbf{Y}=\mathbf{y}$ turns out to be the best predictor with respect to minimizing the mean square error of predictions, the expression of which reduces to the equation of linear regression of W into $\mathbf{y}$ under normality.

When the first moment of W and $\mathbf{Y}$ depend on unknown parameters as happens in mixed linear models, the best predictor among the class of linear and unbiased predictors is BLUP. This predictor can be obtained via Henderson's mixed model equations. The ease of implementation of such equations has allowed the application of BLUP to a vast domain of models, disciplines and situations, especially to large data sets with unbalanced structures (Robinson, 1991). On the other hand, the Bayesian interpretation of HMME gives a broader scope to the theory of BLUP so that this system of equations has become the cornerstone of today mixed model methodology as we shall see it again later on.

In particular, one is naturally interested in predictions of linear combinations of fixed and random effects, say $W=\mathbf{k}'\boldsymbol{\beta}+\mathbf{m}'\mathbf{u}$ taking into account that the variance covariance matrices $\mathbf{G}$ and $\mathbf{R}$ are unknown. The classical theory answers this problem in two steps: i) derive BLUP of W as if $\mathbf{G}$ and $\mathbf{R}$ (or the

parameters $\boldsymbol{\gamma}$ they are depending of) were known, and ii) estimate $\boldsymbol{\gamma}$ according to some relevant statistical procedures (e.g. Maximum Likelihood) from the data and replace $\mathbf{G}$ and $\mathbf{R}$ in the BLUP or HMME equations by their estimates $\mathbf{G}(\hat{\boldsymbol{\gamma}})$ and $\mathbf{R}(\hat{\boldsymbol{\gamma}})$. The methods for estimating the $\boldsymbol{\gamma}$ parameters will be presented in the next two chapters. This procedure involving a plugin stage with parameters replaced by their estimates typically reflects what it is called an Empirical Bayes (EB) approach. We will see later on in the section on stochastic tools how to solve the problem of predicting W via a completely Bayesian setting with prior information on $\boldsymbol{\gamma}$.

2.5 Appendix

Inversion of **V**

Consider the following partition into blocks of a square matrix and its inverse

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix} \tag{2.81}$$

where $\mathbf{A}_{11}$ et $\mathbf{A}_{22}$ are also square and non singular.

The proof proposed is based on the well known following results on the inverse of partitioned matrices: see e.g., Searle, (1966) pages 210-211

$$\mathbf{A}^{11} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}, \tag{2.82}$$

$$\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22} = -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}. \tag{2.83}$$

Similarly for $\mathbf{A}^{22}$ and $\mathbf{A}^{21}$.

Now let us define:

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{R}^{-1} & \mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}. \tag{2.84}$$

By applying the first part of (2.82) to $\mathbf{A}^{22}$, one has

$$\mathbf{A}^{22} = \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} = \left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\right)^{-1} = \mathbf{G}$$

Similarly for $\mathbf{A}^{11}$

$$\mathbf{A}^{11} = \left[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\right]^{-1} = \mathbf{W}^{-1}. \tag{2.85}$$

Then, the application of the second result of (2.82) gives:

$$\mathbf{A}^{11} = \mathbf{R} + \mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{R} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}' = \mathbf{V}, \tag{2.86}$$

QED.

We can do the same for $\mathbf{A}^{12}$ using (2.83)

$$\mathbf{A}^{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}^{22} = -\mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\mathbf{G} = -\mathbf{Z}\mathbf{G}$$

$$\mathbf{A}^{12} = -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} = -\mathbf{V}\mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}$$

Equating these two terms yields:

$$\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} = \left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}, \qquad (2.87)$$

proving the equivalence between the BLUP $\hat{\mathbf{u}}$ of $\mathbf{u}$ and HMME.

2.6 Exercises

2.1 In Example 2.1,

1) derive the analytical expressions of $Var(W)$, $Var(Y)$ and $Cov(W,Y)$ and then R^2 the square of the correlation coefficient between W and Y;

2) apply these formulae to the data in table 2.2 for computing α, β and R^2.

2.2 Derive formula $\hat{a}_i = b_i(y_{i.} - \mu)$ in (2.30) using the general BLP theory of vector $\mathbf{a} = (a_i)$, $1 \le i \le I$ based on the data vector $\mathbf{y} = \left(\mathbf{y}_1', ..., \mathbf{y}_i', ..., \mathbf{y}_I'\right)'$ where $\mathbf{y}_i = (y_{ij})$, $1 \le j \le n_i$.

Hint: write $\mathbf{V}_i$*, the variance of* $\mathbf{y}_i$ *under the form* $\mathbf{V}_i = a\mathbf{I}_{n_i} + b\mathbf{J}_{n_i}$.

2.3. Show that $\mathbf{V}$ is a g-inverse of $\underline{\mathbf{P}} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$. Deduce from this that the BLUP of $w = \mathbf{k}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{u}$ based on $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ such that

$E(\mathbf{y}) = \mathbf{X\beta}$, $Var(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ is translation invariant with respect to fixed effects $\boldsymbol{\beta}$

2.4 Consider longitudinal binary data $y_{ij} = 0$, 1 made of repeated measurements $(j = 1,...,n_i)$ recorded on different individuals $(i = 1,...,I)$. The statistical analysis is carried out according to the following hierarchical model:

a) $y_{ij} \mid \pi_i \sim_{id} B(1,\pi_i)$ are conditionally independent rv's having Bernouilli distributions with parameter π_i ;

b) $\pi_i \sim_{iid} \mathcal{L}(\pi,\rho)$ are independent continuous rv defined on $(0,1)$ with mean π and variance $\rho\pi(1-\pi)$.

1) Show that $E(y_{ij}) = \pi$. In the same way, express $Var(y_{ij})$ and $Cov(y_{ij}, y_{ij'})$ for $j \neq j'$ as a function of π and ρ.

2) From these, derive the expression of $Var(y_{i.})$ where $y_{i.} = \left(\sum_{j=1}^{n_i} y_{ij}\right) / n_i$ as a function of π, ρ and n_i .

3) Show that $Cov(y_{i.}, \pi_i) = \rho\pi(1-\pi)$.

4) Assuming that π, ρ are known, derive the BLP $\hat{\pi}_i$ of π_i based on $y_{i.}$. Show that this predictor can be written as $\hat{\pi}_i = \pi + b_i(y_{i.} - \pi)$ where b_i is a function of ρ and n_i.

5) Compute the values of this predictor for the following data assuming $\pi = 0.15$ and $\rho = 1/11$.

i	n_i	$\sum_{j=1}^{n_i} y_{ij}$
1	90	15
2	20	4
3	5	1

Hint for 1): *One may take advantage of the following identities:*

$E(Y) = E_X(Y \mid X = x)$,

$Var(Y) = Var_X\left[\mathrm{E}(Y \mid X = x)\right] + E_X\left[\mathrm{Var}(Y \mid X = x)\right]$,

$Cov(Y,Z) = Cov_X\left[E(Y \mid X = x), E(Z \mid X = x)\right] + E_X\left[Cov(Y,Z \mid X = x)\right]$.

2.5 Show that the BLUP $\hat{\mathbf{y}}_i^* = \left(y_{ik}^*\right)$ of new observations $\mathbf{y}_i^* = \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{Z}_i^*\mathbf{u}_i + \mathbf{e}_i^*$ based on already observed data $\mathbf{y}_i = \left(y_{ij}\right)$, $1 \le j \le n_i$ can be expressed as $\hat{\mathbf{y}}_i^* = \mathbf{R}_{0,i}\mathbf{V}_i^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}} + \left(\mathbf{I}_{n_i} - \mathbf{R}_{0,i}\mathbf{V}_i^{-1}\right)\mathbf{y}_i$ if $\mathbf{X}_i^* = \mathbf{X}_i$ and $\mathbf{Z}_i^* = \mathbf{Z}_i$ (see Formula (2.45)) and express the corresponding variance of prediction errors $Var\left(\hat{\mathbf{y}}_i^* - \mathbf{y}_i^*\right)$.

2.6 Consider the linear mixed model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ such that

$$E\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} \text{ and } Var\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{Z}\mathbf{G}_{uu}\mathbf{Z}' & \mathbf{Z}\mathbf{G}_{uu} & \mathbf{R}\\ \mathbf{G}_{uu}\mathbf{Z}' & \mathbf{G}_{uu} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix}.$$

1) Let $\hat{\mathbf{u}}$ be the BLUP of $\mathbf{u}$ based on $\mathbf{y}$. Show that $\hat{\mathbf{u}}$ is translation invariant with respect to fixed effects $\boldsymbol{\beta}$.

2) We want to predict the vector $\mathbf{v}$ that does not occur directly in the model for $\mathbf{y}$ but that is correlated to $\mathbf{y}$ via $\mathbf{u}$ as shown below

$$E(\mathbf{v}) = \mathbf{0} \text{ and } Var\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} = \begin{pmatrix}\mathbf{G}_{11} & \mathbf{G}_{12}\\ \mathbf{G}_{21} & \mathbf{G}_{22}\end{pmatrix}.$$

Show that the BLUP $\hat{\mathbf{v}}$ of $\mathbf{v}$ based on $\mathbf{y}$ can be expressed as $\hat{\mathbf{v}} = \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\hat{\mathbf{u}}$. It is assumed that $\mathrm{Cov}(\mathbf{v},\mathbf{e}') = \mathbf{0}$.

3) Write the following system

$$\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{0}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{11} & \mathbf{G}^{12}\\ \mathbf{0} & \mathbf{G}^{21} & \mathbf{G}^{22}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{0}\end{pmatrix}$$

where $\begin{pmatrix} \mathbf{G}^{11} & \mathbf{G}^{12} \\ \mathbf{G}^{21} & \mathbf{G}^{22} \end{pmatrix} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix}^{-1}$.

Show that its solution in $\hat{\mathbf{v}}$ gives the same results as in (2).

2.7 Consider the linear mixed model $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$ such that $E(\mathbf{y}) = \mathbf{X\beta}$ and $Var(\mathbf{y}) = \sigma_u^2 \mathbf{ZZ}' + \sigma_e^2$

1) Show that, in the case of a two way cross-classification analyzed by an additive mixed model (2.62), the reduced system $(\mathbf{Z}'\mathbf{MZ} + \lambda \mathbf{I}_q)\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{My}$ in (2.65) has elements defined in (2.66) (2.67) and (2.68)

2) Prove that $\mathbf{1}'\hat{\mathbf{u}} = 0$.

3) Compute $\hat{\mathbf{u}}$ from the data set of table 2.3 assuming that heifer (A2) progeny out of sire 1 are missing (take $\lambda = 15$).

Sires	A1		A2	
	n	$\sum y$	n	$\sum y$
1	10	440		
2	4	175	16	720
3	40	1730		

4) Same question as in 3) but assuming that sires 1, 2 and 3 have progeny only out of heifers (A1).

5) Compare the results obtained in 3) and 4). How would you explain that?

2.8 Show that $Q(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{T}\boldsymbol{\theta})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{T}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ in (2.72) can be written as $Q(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Omega}^{-1})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + Cst$ where $\hat{\boldsymbol{\theta}}$ is the solution to the system $(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Omega}^{-1})\hat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{R}^{-1}\mathbf{y} + \boldsymbol{\Omega}^{-1}\boldsymbol{\theta}_0$.

2.9. Show that the solution to (2.73) i.e. $(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Omega}^{-1})\hat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{R}^{-1}\mathbf{y} + \boldsymbol{\Omega}^{-1}\boldsymbol{\theta}_0$ can be derived as a weighted mean of two « natural » estimators of $\boldsymbol{\theta}$ viz.

i) the GLS estimator $\hat{\boldsymbol{\theta}}_1 = (\mathbf{T}'\mathbf{R}^{-1}\mathbf{T})^{-1}\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}$

ii) the « prior » estimator $\hat{\boldsymbol{\theta}}_2 = \boldsymbol{\theta}_0$

where weights are equal to their accuracies (inverse variances).

2.10 Show that Henderson's derivation of the Mixed Model Equations via maximizing the joint density $f(\mathbf{y}, \mathbf{u})$ with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ under the normality assumption can be interpreted as a Bayesian procedure?

References

Bulmer, M. G. (1980).*The Mathematical Theory of Quantitative Genetics*. Oxford University Press.

Dempfle, L. (1977). Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayésiens. *Annales de Génétique et de Sélection Animale*, **9**, 27–32.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Discrete Data*. Springer, New-York.

Goldberger, A. S. (1962). Best Linear Unbiased Prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57, 369-375.

Harville, D. A. (1990). BLUP (Best Linear Unbiased Prediction) and beyond. In D. Gianola and K. Hammond, eds., *Advances in Statistical Methods for Genetic Improvement of Livestock*, volume 18 of *Advanced Series in Agricultural Sciences*, 239–276, Springer-Verlag, Berlin.

Henderson, C. R. (1948). *Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine*. Unpublished Ph.D. thesis, Iowa State College Library, Ames, Iowa.

Henderson, C. R. (1950). Estimation of genetic parameters. *Biometrics*, **6**, 186-187.

Henderson, C. R. (1952). Specific and general combining ability. In J. W. Gowen (Ed), *Heterosis*, 352-370. Iowa State College Press, Ames, IA.

Henderson, C.R., Kempthorne, O., Searle, S. R., & von Krosigk, C.N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **13**, 192-218

Henderson, C. R. (1963). Selection index and expected genetic advance. In W. D. Hanson & H. F. Robinson (Eds), *Statistical Genetics and Plant Breeding*. 141-163. National Academy of Sciences, NRC, vol 982, Washington, D C.

Henderson, C.R. (1973). Sire evaluation and genetic trends, In *Proceedings of the animal breeding and genetics symposium in honor of Dr J Lush*. American Society Animal Science-American Dairy Science Association, 10-41, Champaign, IL.

Gianola, D., & Goffinet, B. (1982). Sire evaluation with best linear unbiased predictors. *Biometrics*, **38**, 1085-1088.

Gianola, D., & Fernando, R. F. (1986). Bayesian methods in animal breeding theory. *Journal of Animal Science*, **63**, 217-244

Goffinet, B. (1983). *Risque quadratique et sélection : quelques résultats appliqués à la sélection animale et végétale*. Thèse de Docteur Ingénieur, Université Paul Sabatier, Toulouse.

Lindley, D.V., & Smith, A.F.M. (1972). Bayes Estimates for the Linear Model, *Journal of the Royal Statistical Society* B, **34**, 1-41.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6**,15-32.

Schachter, J., Stamm, W. E., Quinn, T. C., Andrews, W. W., Burczak, J. D., & Lee, H. H. (1994). Ligase chain reaction to detect Chlamydia trachomatis infection of the cervix. *Journal of Clinical Microbiology*, **32**, 2540–2543.

Searle, S. R. (1966). *Matrix Algebra for the Biological Sciences*. Wiley, New York.

Searle, S.R., Casella, G., & Mc Culloch, C.E. (1992). *Variance components*, Wiley, New-York

3
Maximum likelihood procedures

3.1 Introduction

Maximum likelihood (later on abbreviated as ML) is a general procedure due to Fisher (1922, 1925) which has interesting statistical properties chiefly in asymptotic conditions (Cox and Hinkley, 1974). As far as variances are concerned, this procedure was used by Crump (1947) in simple cases (one-way classification, balanced designs). But, Hartley and Rao (1967) were the first who formalized the general ML approach for the estimation of variance components for the linear mixed model. This publication marks a break with the quadratic estimator era which began with the work of Fisher on the intra-class correlation and reached its summit with Henderson's I, II and III methods (1953). These methods were conceived as an extension to unbalanced data of ANOVA procedures designed either for balanced or unbalanced data. They are based on quadratic forms obtained mostly under fixed models and are derived on the sole property of being unbiased under the true mixed model without any optimality consideration in the choice of these quadratic forms. Therefore, for most statisticians, they just have a historical interest although they are relatively easy to compute and they have been shown to be quite efficient in simulation studies. In addition, they return minimum variance (MINVAR) and REML estimators for balanced data under the normality assumption. This is why a brief but

general description of the most sophisticated version of them (Method III) is provided in Appendix 3.1 with an application to the mixed model for the two-way crossed classification with interaction.

Similarly, Rao's (1971ab) and Lamotte's (1973) methods just appear today as a transition between Henderson's and ML procedures since MINQUE ("minimum norm quadratic unbiased estimation) when iterated, results in a maximum likelihood estimator (REML under normality as defined below).

To that respect, two approaches of maximum likelihood must be distinguished. The first one, known as ML, relies on the standard concept of likelihood as a function of all parameters involved in the data distribution. The second one was introduced and developed by Patterson and Thompson (1971) for the Gaussian linear mixed model applied to interblock variation. It utilizes linear combinations of data, the so-called error contrasts in Harville's (1977) terminology; these are free of fixed effects and the likelihood function so obtained yields after maximization what is called "residual or restricted maximum likelihood" (acronym REML). In addition, this residual likelihood has a Bayesian interpretation (Harville, 1974) as a marginalized likelihood after integration of fixed effects assuming a uniform prior distribution on them.

Both ML and REML have received considerable practical interest over the last twenty years. This is due primarily to their growing numerical feasibility through more and more efficient computers, software and algorithms. The objectives of this chapter are to provide the reader with the basic theoretical foundations and derivations underlying these two procedures. We will first review (section 3.2) the basic theory of maximum likelihood including models, derivation of ML estimators, variants, numerical aspects and hypothesis testing of fixed effects. In a second section (3.3), we will deal with REML both from classical and Bayesian points of view and discuss the repercussions of using REML instead of ML on procedures for testing fixed and random effects.

3.2 Basic theory of maximum likelihood

3.2.1 Models and notations

Here, for convenience reasons, we will follow Henderson's notations as they are now adopted in major textbooks (see e.g., Searle et al., 1992) and software (SAS-Proc Mixed, ASREML). The linear mixed model we are dealing with, is written under its generic form as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \qquad (3.1)$$

where $\mathbf{y}$ is the $(N \text{ x } 1)$ vector of observations ; $\mathbf{X}$ is the $(N \text{ x } p)$ matrix of explanatory variables with the corresponding vector $\boldsymbol{\beta} \in R^p$ of coefficients or « fixed effects » ; $\mathbf{u}$ is the $(q \text{ x } 1)$ vector of structural random variables or « random effects » with their corresponding incidence $(N \text{ x } q)$ matrix $\mathbf{Z}$,and $\mathbf{e}$ is the $(N \text{ x } 1)$ vector of residual random variables.

This model is characterized by its mean and variance

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X\beta} \qquad (3.2)$$

$$\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} \qquad (3.3)$$

where $\mathbf{u}$~$(\mathbf{0},\mathbf{G})$, $\mathbf{e}$~$(\mathbf{0},\mathbf{R})$, and $\mathrm{Cov}(\mathbf{u},\mathbf{e}')=\mathbf{0}$.

This general expression can encompass most of the particular situations encountered in practice, notably that of an ANOVA type model with several independent random factors of variation $k = 1,...,K$ as considered e.g. in Henderson's methods.

$$\mathbf{y} = \mathbf{X\beta} + \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{u}_k + \mathbf{e}. \qquad (3.4)$$

Assuming $\mathbf{u}_k \sim (\mathbf{0}, \sigma_k^2 \mathbf{I}_{q_k})$, $\mathbf{e} \sim (\mathbf{0}, \sigma_0^2 \mathbf{I}_N)$, $\mathbf{u}_k \perp \mathbf{e}, \forall k$, and in the case of uncorrelated random effects $\mathbf{u}_k \perp \mathbf{u}_l$ for $k \neq l$, the variance covariance matrix $\mathbf{V}$ takes the linear form

$$\mathbf{V} = \sum_{k=1}^{K} \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k^{'} + \sigma_0^2 \mathbf{I}_N , \qquad (3.5)$$

where the parameters $\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2, \ \ldots, \sigma_K^2$ are the so called variance components.

3.2.2 Derivation of ML

Likelihood function

Let us first consider the case of a general Gaussian linear model written as

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X\beta}, \mathbf{V}) \qquad (3.6)$$

where $\mathrm{E}(\mathbf{y}) = \mathbf{X\beta}$ as in (4.1) and $\mathbf{V} = \mathbf{V}(\boldsymbol{\gamma})$ is a $(N \times N)$ symmetric positive definite matrix depending on a parameter $\boldsymbol{\gamma} \in \Gamma$.

Under (3.6), the density of observations is

$$\mathrm{p}_Y(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = (2\pi)^{-N/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y} - \mathbf{X\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X\beta}) \right]. \qquad (3.7)$$

Its logarithm $\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = \log \mathrm{p}_Y(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$ viewed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for a given data sample $\mathbf{y}$ is called the log-likelihood which under its $-2L$ form is expressed as

$$-2\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = N \log(2\pi) + \ln|\mathbf{V}| + (\mathbf{y} - \mathbf{X\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X\beta}). \qquad (3.8)$$

Maximization

i) first derivatives

Searching for points $\boldsymbol{\alpha} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ which maximize $\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$ (or alternatively minimize $-2\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$) i.e.

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathrm{A}} \mathrm{L}(\boldsymbol{\alpha}; \mathbf{y})$$

is usually carried out by setting the first derivatives to zero. Such a procedure has to be applied with much care. First, one has to check that the points so obtained are inside the parameter space $A = \mathbb{R}^p \times \Gamma$. Second, one has to verify that at those points, the matrix of second derivatives $\partial^2 \mathrm{L}(\boldsymbol{\alpha}) / \partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'$ is negative definite. Regarding the parameter space, the condition on $\boldsymbol{\beta}$ does not raise any

difficulty. On the contrary, as far as γ is concerned, its parameter space Γ must be specified in detail for each model. For instance, in a linear mixed model with K independent random factors as in (4.4), one will impose $\sigma_0^2 > 0$ and $\sigma_k^2 \geq 0$, $\forall k \in \{1,...,K\}$. Obviously, these restrictions are more severe than that requiring $\mathbf{V}$ positive-definite.

The property of negativity of the matrix of second derivatives at the points zeroing the first derivatives and that are at the interior points of the parameter space provides a necessary condition for the existence of a maximum, but this one may not be global. It might be difficult to identify all such local maxima and then to evaluate the value of the log-likelihood function at these points as well as on the border of the parameter space. In the latter case, this may require special procedures of maximization under constraints. Things become much easier when L is a concave function of the parameters since then first-order conditions guarantee the existence of a global maximum.

The first derivatives of $-2\mathrm{L}$ with respect to the parameters are

$$\frac{\partial(-2\mathrm{L})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.9}$$

$$\frac{\partial(-2\mathrm{L})}{\partial \gamma_k} = \frac{\partial \log|\mathbf{V}|}{\partial \gamma_k} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.10}$$

Now, from standard results in differentiation of matrix expressions (see e.g., Searle, 1982, pages 335-337; Harville, 1997, pages 305-308),

$$\frac{\partial \ln|\mathbf{V}|}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right), \tag{3.11}$$

$$\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k} = -\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}. \tag{3.12}$$

Hence,

$$\frac{\partial(-2\mathrm{L})}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.13}$$

Equating (4.9) and (4.10) to zero gives the following system

$$\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{3.14}$$

$$\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right)_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} - (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}\left.\frac{\partial \mathbf{V}}{\partial \gamma_k}\right|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}}\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = 0, \tag{3.15}$$

where $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ are the solutions to this system (if they exist) and $\hat{\mathbf{V}}$ stands for $\mathbf{V}(\hat{\boldsymbol{\gamma}})$.

Some simplifications can be made. First, $\hat{\boldsymbol{\beta}}$ can be eliminated from (3.15) by substituting its expression $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ from (3.14) into (3.15), and by observing that $\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \underline{\hat{\mathbf{P}}}\mathbf{y}$ where

$$\underline{\mathbf{P}} = \mathbf{V}^{-1}(\mathbf{I}-\mathbf{Q}) = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}\mathbf{V}^{-1}. \tag{3.16}$$

where $\mathbf{Q}$ is the GLS projector defined in (1.72).

Then (3.15) becomes

$$\operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right)_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} - \mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{P}\mathbf{y}\Big|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} = 0. \tag{3.17}$$

ii) general case

The system in (3.17) cannot generally be solved analytically and one has to recourse to numerical analysis such as the Newton-Raphson or the Fisher scoring algorithms. This involves computing the Hessian $\ddot{\mathbf{L}}(\boldsymbol{\alpha};\mathbf{y}) = \partial^2 \mathrm{L}(\boldsymbol{\alpha};\mathbf{y})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'$ or the Fisher information matrix $\mathbf{J}(\boldsymbol{\alpha}) = \mathrm{E}_{Y|\boldsymbol{\alpha}}\left[-\ddot{\mathbf{L}}(\boldsymbol{\alpha};\mathbf{y})\right]$. In the latter case, this gives (see appendix 3.6)

$$\mathbf{J}(\boldsymbol{\alpha}) = \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}/2 \end{pmatrix}, \tag{3.18}$$

where

$$(\mathbf{F})_{kl} = \operatorname{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \tag{3.19}$$

Hence, the method of scoring applied to $\boldsymbol{\gamma}$ requires iterating with

$$\mathbf{J}(\boldsymbol{\gamma}^{[n]})\boldsymbol{\Delta}^{[n+1]} = \dot{\mathbf{L}}(\boldsymbol{\gamma}^{[n]}), \tag{3.20}$$

where

$$\boldsymbol{\Delta}^{[n+1]} = \boldsymbol{\gamma}^{[n+1]} - \boldsymbol{\gamma}^{[n]}$$

$$\dot{\mathbf{L}}(\boldsymbol{\gamma}^{[n]}) = \left\{ -\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k} \right) + \frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\mathbf{y} \right\}_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{[n]}} \tag{3.21}$$

$$\mathbf{J}(\boldsymbol{\gamma}^{[n]}) = 1/2\mathbf{F}\left(\boldsymbol{\gamma}^{[n]}\right). \tag{3.22}$$

Once the solution in $\boldsymbol{\gamma}$ is found, one obtains $\hat{\boldsymbol{\beta}}$ by plugging $\hat{\boldsymbol{\gamma}}$ into $\mathbf{V}$, and solving (3.14). Notice that this system is similar to the one yielding the GLS estimator but here $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\gamma}})$ replaces $\mathbf{V}$.

iii) case of the linear mixed model

In this situation, $\mathbf{V}$ takes the form $\mathbf{V} = \sum_{k=0}^{K}\mathbf{V}_k\gamma_k$ with $\partial\mathbf{V}/\partial\gamma_k = \mathbf{V}_k$, $\mathbf{V}_k$ being a $(N \text{ x } N)$ matrix of known coefficients e.g., $\mathbf{V}_k = \mathbf{Z}_k\mathbf{Z}_k^{'}$ for models defined in (3.4) and (3.5). Then equation (3.17) becomes

$$\mathrm{tr}\left(\hat{\mathbf{V}}^{-1}\mathbf{V}_k\right) - \mathbf{y}'\hat{\underline{\mathbf{P}}}\mathbf{V}_k\hat{\underline{\mathbf{P}}}\mathbf{y} = 0. \tag{3.23}$$

On account of the linearity property of $\mathbf{V}$, the first term in (3.23) can be decomposed as the following sum

$$\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{V}_k\right) = \sum_{l=0}^{K}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{V}_k\mathbf{V}^{-1}\mathbf{V}_l\right)\gamma_l.$$

Therefore, the system (3.23) of ML equations can be written as

$$\sum_{l=0}^{K}\mathrm{tr}\left(\hat{\mathbf{V}}^{-1}\mathbf{V}_k\hat{\mathbf{V}}^{-1}\mathbf{V}_l\right)\hat{\gamma}_l = \mathbf{y}'\hat{\underline{\mathbf{P}}}\mathbf{V}_k\hat{\underline{\mathbf{P}}}\mathbf{y} \;;\; (k = 0,1,...,K). \tag{3.24}$$

Under a matrix form, this is tantamount to

$$\hat{\mathbf{F}}\hat{\boldsymbol{\gamma}} = \hat{\mathbf{g}}, \tag{3.25}$$

where $\mathbf{F}$ is a symmetric $\left((K+1)\mathrm{x}(K+1)\right)$ matrix, and $\mathbf{g}$, a $(K+1)$ vector defined respectively as

$$\mathbf{F} = \left({}_m f_{kl} \right) = \left({}_m \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{V}_k \mathbf{V}^{-1}\mathbf{V}_l \right) \right), \tag{3.26}$$

$$\mathbf{g} = \left({}_c g_k \right) = \left({}_c \mathbf{y}'\underline{\mathbf{P}}\mathbf{V}_k\underline{\mathbf{P}}\mathbf{y} \right), \tag{3.27}$$

$\hat{\mathbf{F}}$ and $\hat{\mathbf{g}}$ being $\mathbf{F}$ and $\mathbf{g}$ respectively evaluated at $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$.

Equation (3.25) forms a non linear system which generally has no analytical solution. This system can be solved numerically by an iterative algorithm, each iteration of which, however, having a linear form

$$\mathbf{F}\left(\boldsymbol{\gamma}^{[n]} \right) \boldsymbol{\gamma}^{[n+1]} = \mathbf{g}\left(\boldsymbol{\gamma}^{[n]} \right), \tag{3.28}$$

where $\boldsymbol{\gamma}^{[n]}$ is the current value of the parameter at iteration n, and $\boldsymbol{\gamma}^{[n+1]}$ is the updated value.

It can be shown that (3.28) is equivalent to solving Fisher's scoring algorithm in (3.20).

Example 3.1 *ML equations for a single random factor model.*

Let us consider the same model as in (3.4) but with just a single random factor i.e., $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ where $\mathbf{u} \sim \left(\mathbf{0}, \sigma_1^2 \mathbf{I}_q \right)$, $\mathbf{e} \sim \left(\mathbf{0}, \sigma_0^2 \mathbf{I}_N \right)$, and $\operatorname{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$. We can explicit the elements of the system $\mathbf{F}\boldsymbol{\gamma} = \mathbf{g}$ as follows. The two elements of $\mathbf{g} = \left(g_0, g_1 \right)'$ are the quadratic forms: $g_0 = \mathbf{y}'\underline{\mathbf{P}}^2\mathbf{y}$ and $g_1 = \mathbf{y}'\underline{\mathbf{P}}\mathbf{Z}\mathbf{Z}'\underline{\mathbf{P}}\mathbf{y}$. The coefficients of $\mathbf{F}$ are

$$f_{00} = \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{V}_0 \mathbf{V}^{-1}\mathbf{V}_0 \right) = \operatorname{tr}\left(\mathbf{V}^{-2} \right),$$

$$f_{01} = \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{V}_0 \mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}' \right) = \operatorname{tr}\left(\mathbf{Z}'\mathbf{V}^{-2}\mathbf{Z} \right),$$

$$f_{11} = \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}' \right) = \operatorname{tr}\left[\left(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z} \right)^2 \right].$$

Notice that the calculations of g_0 and g_1 can take advantage of their sum of squares structure i.e., $g_0 = \sum_{i=1}^{N} \left(\underline{\mathbf{P}}\mathbf{y} \right)_i^2$, and $g_1 = \sum_{j=1}^{q} \left(\mathbf{Z}'\underline{\mathbf{P}}\mathbf{y} \right)_j^2$. Similarly, the calculations of f_{00} and f_{11} can be simplified knowing that the trace of the

product of a matrix $\mathbf{A}=\left({}_{m}a_{ij}\right)$ and its transpose is equal to the sum of its squared elements, $\mathrm{tr}\left(\mathbf{AA}'\right)=\sum_{ij}a_{ij}^{2}$. Finally, the system to be solved iteratively to get the ML estimations of σ_0^2 and σ_1^2 is

$$\begin{pmatrix} f_{00}^{(n)} & f_{01}^{(n)} \\ f_{01}^{(n)} & f_{11}^{(n)} \end{pmatrix}\begin{pmatrix} \sigma_0^{2(n+1)} \\ \sigma_1^{2(n+1)} \end{pmatrix}=\begin{pmatrix} g_0^{(n)} \\ g_1^{(n)} \end{pmatrix}$$

starting from initial values $\sigma_0^{2(0)}$ and $\sigma_1^{2(0)}$ which can be taken as guessed values or estimations of a quadratic method.

This procedure can be illustrated by the following numerical application pertaining to a two way crossclassified design with factor A as fixed and B as random according to the model

$$y_{ijk}=\mu+a_i+b_j+e_{ijk},$$

where a_i is the fixed effect of level i, $b_j \sim_{iid}\left(0,\sigma_1^2\right)$ and $e_{ijk} \sim_{iid}\left(0,\sigma_0^2\right)$.

Table *Distribution of data according to levels of factors A and B*

A	B	*n*	*y*
1	1	2	8, 2
1	2	2	10, 4
1	3	6	3, 9, 9, 9, 10, 8
2	1	5	3, 9, 6, 4, 8
2	2	9	14, 8, 6, 9, 6, 6, 10, 12, 10

If data are sorted by levels of B first $\left(j=1,2,3\right)$, the matrices $(\mathbf{X},\mathbf{Z},\mathbf{y})$ are as follows for $j=1,2,3$

For $\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 8 \\ 1 & 1 & 1 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 0 & 3 \\ 1 & 0 & 1 & 0 & 0 & 9 \\ 1 & 0 & 1 & 0 & 0 & 6 \\ 1 & 0 & 1 & 0 & 0 & 4 \\ 1 & 0 & 1 & 0 & 0 & 8 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 10 \\ 1 & 1 & 0 & 1 & 0 & 4 \\ 1 & 0 & 0 & 1 & 0 & 14 \\ 1 & 0 & 0 & 1 & 0 & 8 \\ 1 & 0 & 0 & 1 & 0 & 6 \\ 1 & 0 & 0 & 1 & 0 & 9 \\ 1 & 0 & 0 & 1 & 0 & 6 \\ 1 & 0 & 0 & 1 & 0 & 6 \\ 1 & 0 & 0 & 1 & 0 & 10 \\ 1 & 0 & 0 & 1 & 0 & 12 \\ 1 & 0 & 0 & 1 & 0 & 10 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 & 1 & 9 \\ 1 & 1 & 0 & 0 & 1 & 9 \\ 1 & 1 & 0 & 0 & 1 & 10 \\ 1 & 1 & 0 & 0 & 1 & 8 \end{pmatrix}.$

The algorithm is initiated from starting values of the parameters. For instance with $\sigma_0^2 = 10$ $\sigma_1^2 = 5$, one obtains the iterative scheme

#	f_{00}	f_{01}	f_{11}	g_0	g_1	σ_0^2	σ_1^2
1	0.211355	0.009810	0.075337	1.611703	0.143796	7.5828	0.9213
2	0.379308	0.105504	0.843795	2.978514	1.441847	7.6430	0.7531
3	0.376191	0.125895	1.014146	2.971685	1.707407	7.6539	0.7334
4	0.375509	0.128598	1.036909	2.968671	1.742377	7.6554	0.7309
5	0.375416	0.128948	1.039857	2.968245	1.746897	7.6556	0.7306

As shown on this table displaying the first five iterations, the algorithm converges very rapidly to the same final solutions whatever are the starting values chosen. According to (3.23), $2\hat{\mathbf{F}}^{-1}$ also provides an estimate of the asymptotic sampling variance covariance matrix of the ML estimators that is, for $\hat{\sigma}_0^2 = 7.6556$ and $\hat{\sigma}_1^2 = 0.7306$

$$2\hat{\mathbf{F}}^{-1} = \begin{pmatrix} 5.5647 & -0.6900 \\ -0.6897 & 2.0081 \end{pmatrix}.$$

These results can be checked using some standard software.

For instance, SAS-Proc Mixed on this data set with the same starting values and the Fisher "scoring" option gives $\hat{\sigma}_0^2 = 7.6554$, $\hat{\sigma}_1^2 = 0.7310$ and $2\hat{\mathbf{F}}^{-1} = \begin{pmatrix} 5.5640 & -0.6897 \\ -0.6897 & 2.0090 \end{pmatrix}$.

Despite the high speed of convergence, this algorithm involves heavy computations which make it difficult to apply to large data sets and we will see later on how to cope with this problem. However, it can be applied to all kinds of mixed linear models involving discrete and/or continuous covariates and even correlated random effects.

3.2.3 Variants

Profile likelihood

The principle underlying this procedure lies in maximizing the log-likelihood by successive steps. First, we maximize $\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y})$ with respect to $\boldsymbol{\beta}$. Then, the function so obtained $\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y}) = \mathrm{L}\left(\hat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\gamma};\mathbf{y}\right)$ of the single vector $\boldsymbol{\gamma}$ is maximized with respect to this parameter. It is called the profile (Cox and Reid, 1987) or concentrated (Harville and Callanan, 1990) log-likelihood. In short, we can summarize the process as follows

$$\begin{aligned} \mathrm{Max}_{\boldsymbol{\beta},\boldsymbol{\gamma}}\,\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) &= \mathrm{Max}_{\boldsymbol{\gamma}}\left[\mathrm{Max}_{\boldsymbol{\beta}}\,\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y})\right] \\ &= \mathrm{Max}_{\boldsymbol{\gamma}}\,\mathrm{L}\left(\hat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\gamma};\mathbf{y}\right) \\ &= \mathrm{Max}_{\boldsymbol{\gamma}}\,\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y}) \end{aligned} \quad , \qquad (3.29)$$

where $\hat{\boldsymbol{\beta}}_\gamma = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is a GLS solution for $\boldsymbol{\beta}$.

On account of (3.8), minus twice the log-likelihood is

$$-2\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y}) = N\ln(2\pi) + \ln|\mathbf{V}| + \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_\gamma\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_\gamma\right),$$

or, alternatively,

$$-2\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y}) = N\ln(2\pi) + \ln|\mathbf{V}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y}\,. \qquad (3.30)$$

On using the identity $\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}}$ (see appendix), one immediately obtains the expression of the gradient

$$\frac{\partial \left[-2\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y}) \right]}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \gamma_k} \right) - \mathbf{y}' \underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}} \mathbf{y}, \qquad (3.31)$$

which actually coincides with (3.17).

Two remarks are worth mentioning at this stage:

a) the profile likelihood reduces the dimension of the unknowns by "concentrating" the likelihood function onto the parameters of interest, after eliminating the nuisance parameters;

b) however, strictly speaking, this function is not a likelihood function even if, occasionally, it holds some of its properties (Berger et al, 1999).

Example 3.2 *Likelihood and Profile likelihood for a N Gaussian sample* $y_i \sim_{iid} \mathcal{N}(\mu, \sigma^2)$

Since the observations are independent, the density of the data vector, $\mathbf{y} = \{y_i\}$ is the product of the elementary densities

$$\mathrm{p}(\mathbf{y}; \mu, \sigma^2) = \prod\nolimits_{i=1}^{N} \mathrm{p}(y_i; \mu, \sigma^2).$$

Therefore, the loglikelihood $\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = \ln \mathrm{p}(\mathbf{y}; \mu, \sigma^2)$ for this sample can be written as a sum $\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = \sum_{i=1}^{N} \mathrm{L}(\mu, \sigma^2; y_i)$ where $\mathrm{L}(\mu, \sigma^2; y_i) = \ln \mathrm{p}(y_i; \mu, \sigma^2)$. By definition of the standard Gaussian distribution,

$$-2\mathrm{L}(\mu, \sigma^2; y_i) = \ln 2\pi + \ln \sigma^2 + (y_i - \mu)^2 / \sigma^2,$$

so that

$$-2\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = N(\ln 2\pi + \ln \sigma^2) + \sum\nolimits_{i=1}^{N} (y_i - \mu)^2 / \sigma^2.$$

One can decompose the last sum of squares into $\sum_{i=1}^{N}(y_i-\mu)^2 = N\left[s^2+(\bar{y}-\mu)^2\right]$, where $\bar{y}=\left(\sum_{i=1}^{N} y_i\right)/N$ is the sample mean, and $s^2=\sum_{i=1}^{N}(y_i-\bar{y})^2/N$ is the sample variance.

Hence,

$$-2\mathrm{L}\left(\mu,\sigma^2;\mathbf{y}\right) = N\left[\ln 2\pi + \ln\sigma^2 + \frac{s^2+(\bar{y}-\mu)^2}{\sigma^2}\right].$$

Since $\partial(-2\mathrm{L})/\partial\mu = -2N(\bar{y}-\mu)/\sigma^2$, the first step of maximization with respect to μ leads to $\hat{\mu}=\bar{y}$ which does not depend on σ^2. Therefore, minus twice the log-likelihood reduces simply to

$$-2\mathrm{L}_P\left(\sigma^2;\mathbf{y}\right) = N\left(\ln 2\pi + \ln\sigma^2 + s^2/\sigma^2\right).$$

Differentiating this with respect to σ^2 leads to

$$\frac{\partial}{\partial\sigma^2}\left[-2\mathrm{L}_P\left(\sigma^2;\mathbf{y}\right)\right] = N\left(\sigma^2-s^2\right)/\sigma^4$$

that gives for $N\geq 2$, $\hat{\sigma}^2=s^2$, the usual ML estimator of the variance.

However, it is important to notice that contrarily to a regular likelihood, the expectation of this score function is not zero since $\mathrm{E}\left(s^2\right)=(N-1)\sigma^2/N$.

The Hartley-Rao form

Hartley and Rao (1967) consider linear mixed models as described in (3.4) and (3.5). But, instead of parameterizing $\mathbf{V}$ in terms of the variance components $\boldsymbol{\sigma}^2=\left({}_c\sigma_k^2\right)_{0\leq k\leq K}$,they single out the residual variance σ_0^2 and introduce the vector $\boldsymbol{\eta}=\left({}_c\eta_k=\sigma_k^2/\sigma_0^2\right)_{1\leq k\leq K}$, of variance ratios. To that respect, they write $\mathbf{V}$ as $\mathbf{V}=\mathbf{H}\sigma_0^2$ where $\mathbf{H}=\mathbf{I}_N+\sum_{k=1}^{K}\eta_k\mathbf{Z}_k\mathbf{Z}_k^{'}$ is a function of only $\boldsymbol{\eta}$. Since $|\mathbf{V}|=|\mathbf{H}|\sigma_0^{2N}$, the log-likelihood becomes

$$\begin{aligned}-2\mathrm{L}\left(\boldsymbol{\beta},\sigma_0^2,\boldsymbol{\eta};\mathbf{y}\right) = {} & N\ln(2\pi)+\ln|\mathbf{H}|+N\ln\sigma_0^2 \\ & +(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma_0^2\end{aligned}, \qquad (3.32)$$

Then, differentiating this function with respect to the parameters gives

$$\frac{\partial(-2\mathrm{L})}{\partial\boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma_0^2, \tag{3.33}$$

$$\frac{\partial(-2\mathrm{L})}{\partial\sigma_0^2} = \frac{N}{\sigma_0^2} - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma_0^4}, \tag{3.34}$$

$$\frac{\partial(-2\mathrm{L})}{\partial\eta_k} = \mathrm{tr}\left(\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial\eta_k}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{H}^{-1}\frac{\partial\mathbf{H}}{\partial\eta_k}\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma_0^2. \tag{3.35}$$

The ML equations are obtained by equating (3.33), (3.34) and (3.35) to zero

$$\mathbf{X}'\hat{\mathbf{H}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\hat{\mathbf{H}}^{-1}\mathbf{y}, \tag{3.36}$$

$$\hat{\sigma}_0^2 = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{H}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/N \tag{3.37}$$

$$\mathrm{tr}\left(\hat{\mathbf{H}}^{-1}\mathbf{H}_k\right) - (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{H}}^{-1}\mathbf{H}_k\hat{\mathbf{H}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}_0^2 = 0, \tag{3.38}$$

where $\mathbf{H}_k = \partial\mathbf{H}/\partial\eta_k = \mathbf{Z}_k\mathbf{Z}_k^{'}$.

Equation (3.36) looks very similar to the classical one in (3.14) except that it involves the variance ratios η_k instead of the variance components σ_k^2. The Hartley-Rao form leads to a particular equation for the residual variance shown in (3.37) for which, as will be seen later on, Henderson (1973) proposed a simple algorithm of calculation. As previously, on using

$$\mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{H}_k\right) = \sum_{l=1}^{K}\mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}\mathbf{H}_l\right)\eta_l + \mathrm{tr}\left(\mathbf{H}^{-2}\mathbf{H}_k\right),$$

(3.38) can be replaced by a quasi-linear system

$$\mathbf{D}\left(\boldsymbol{\eta}^{[n]}\right)\boldsymbol{\eta}^{[n+1]} = \mathbf{e}\left(\boldsymbol{\eta}^{[n]}\right), \tag{3.39}$$

where $\mathbf{D}(.) = \left({}_m d_{kl}\right)$, $\mathbf{e}(.) = \left({}_c e_k\right)$ and

$$d_{kl} = tr\left(\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}\mathbf{H}_l\right)$$

$$e_k = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{H}^{-1}\mathbf{H}_k\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}_0^2 - \mathrm{tr}\left(\mathbf{H}^{-2}\mathbf{H}_k\right)$$

In the same way, solving for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_0^2$ can be carried out via iterating with

$$\mathbf{X}'\left[\mathbf{H}^{(n)}\right]^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}^{(n+1)} = \mathbf{X}'\left[\mathbf{H}^{(n)}\right]^{-1}\mathbf{y},$$

$$\hat{\sigma}_0^{2(n+1)} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n+1)})' \left[\mathbf{H}^{(n)}\right]^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(n+1)}) / N .$$

Letting $\boldsymbol{\theta} = \left(\sigma_0^2, \boldsymbol{\eta}'\right)'$ and $\boldsymbol{\sigma}^2 = \left({}_c\sigma_k^2\right)_{0 \le k \le K}$, one can derive the expression of the information matrix $\mathbf{J}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ given that of $\mathbf{J}\left(\boldsymbol{\sigma}^2\right) = \mathbf{F}/2$ for $\boldsymbol{\sigma}^2$ using

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\sigma}^{2\,\prime}}{\partial \boldsymbol{\theta}} \mathbf{J}\left(\boldsymbol{\sigma}^2\right) \frac{\partial \boldsymbol{\sigma}^2}{\partial \boldsymbol{\theta}'} .$$

3.2.4 Numerical aspects

Henderson's algorithm

Very early, Henderson (1973) had in mind to derive an algorithm that avoids computing directly such terms as $\mathbf{V}^{-1}$ and $\underline{\mathbf{P}}$. He started from the same kind of linear mixed models as previously and the derivative of $-2\mathrm{L}_P$ with respect to σ_k^2 which is, from (3.31)

$$\partial(-2\mathrm{L}_P) / \partial\sigma_k^2 = \mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_k\mathbf{Z}_k'\right) - \mathbf{y}'\underline{\mathbf{P}}\mathbf{Z}_k\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{y} .$$

Remember the BLUP $\hat{\mathbf{u}}_k$ of $\mathbf{u}_k$ can be written as $\hat{\mathbf{u}}_k = \mathrm{Cov}\left(\mathbf{u}_k, \mathbf{y}'\right)\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$, viz, $\hat{\mathbf{u}}_k = \sigma_k^2\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{y}$ so that the quadratic form $\mathbf{y}'\underline{\mathbf{P}}\mathbf{Z}_k\mathbf{Z}_k'\underline{\mathbf{P}}\mathbf{y}$ reduces simply to a sum of squares $\hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k / \sigma_k^4$.

Similarly, it can be shown that

$$\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}_k\mathbf{Z}_k'\right) = \frac{q_k}{\sigma_k^2} - \frac{\mathrm{tr}\left(\underline{\mathbf{C}}_{kk}\right)\sigma_0^2}{\sigma_k^4},$$

where $\underline{\mathbf{C}}_{kk} = \left[\left(\mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1}\right)^{-1}\right]_{kk}$ is the $(q_k \,\mathrm{x}\, q_k)$ block pertaining to the random factor k in the inverse of the random factor part of the MME, and $\mathbf{G} = \bigoplus_{k=1}^{K} \sigma_k^2 \mathbf{I}_{q_k}$.

Therefore, setting to zero the derivatives of $-2\mathrm{L}_P$ with respect to σ_k^2 leads to the equations

$$q_k\hat{\sigma}_k^2 = \hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k + \mathrm{tr}\left(\hat{\underline{\mathbf{C}}}_{kk}\right)\hat{\sigma}_0^2 . \qquad (3.40)$$

As far as the residual variance σ_0^2 is concerned, the reasoning is based on the following profile log-likelihood $-2\mathrm{L}_p(\boldsymbol{\eta};\mathbf{y}) = -2\mathrm{L}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}), \hat{\sigma}_0^2(\boldsymbol{\eta}), \boldsymbol{\eta}; \mathbf{y}\right]$ pertaining to the Hartley-Rao form, and which is

$$-2\mathrm{L}_p(\boldsymbol{\eta};\mathbf{y}) = N(\ln 2\pi + 1) + \ln|\mathbf{H}| + N\ln\hat{\sigma}_0^2(\boldsymbol{\eta}),$$

where

$$\hat{\sigma}_0^2(\boldsymbol{\eta}) = \left[\mathbf{y} - \hat{\boldsymbol{\beta}}(\boldsymbol{\eta})\right]'\mathbf{H}^{-1}\left[\mathbf{y} - \hat{\boldsymbol{\beta}}(\boldsymbol{\eta})\right] / N$$

and $\hat{\boldsymbol{\beta}}(\boldsymbol{\eta})$ is a solution to

$$\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\eta}) = \mathbf{X}'\mathbf{H}^{-1}\mathbf{y}.$$

Now, the BLUP $\hat{\mathbf{e}}$ of $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}$ can be expressed as $\hat{\mathbf{e}} = \mathbf{R}\underline{\mathbf{P}}\mathbf{y}$, (Here $\mathbf{R} = \sigma_0^2\mathbf{I}_N$), so that (ignoring $\boldsymbol{\eta}$ within parentheses)

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \sigma_0^2\mathbf{y}'\underline{\mathbf{P}}\mathbf{y} = \mathbf{y}'\hat{\mathbf{e}}$$

and

$$\hat{\sigma}_0^2 = \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'\mathbf{Z}'\mathbf{y}\right) / N. \qquad (3.41)$$

Notice the similarity between this formula and the one obtained for the ML estimator of σ^2 in the fixed model case: $\hat{\sigma}^2 = \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}\right) / N$, the numerator of both being obtained as the difference between the total sum of squares $\mathbf{y}'\mathbf{y}$ and the product of the solution of either the MME equations or the LS system times their respective right-hand sides.

Henderson (1973) proposed to utilize (3.40) and (3.41) as a basis for an iterative algorithm for computing ML estimations of variance components in models such as (3.4),

$$\sigma_k^{2[n+1]} = \left[\hat{\mathbf{u}}_k^{[n]'}\hat{\mathbf{u}}_k^{[n]} + \mathrm{tr}\left(\underline{\mathbf{C}}_{kk}^{[n]}\right)\sigma_0^{2[n]}\right] / q_k \qquad (3.42)$$

$$\sigma_0^{2[n+1]} = \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}^{[n]'}\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}^{[n]'}\mathbf{Z}'\mathbf{y}\right) / N, \qquad (3.43)$$

where $\hat{\boldsymbol{\beta}}^{[n]}$, $\hat{\mathbf{u}}^{[n]}$ are solutions to the HMME using $\sigma_0^{2[n]}$ and $\sigma_k^{2[n]}$ in the coefficient matrix.

Formula (3.43) can be easily extended to the case of correlated random vectors

$$Var\begin{pmatrix}\mathbf{u}_k\\ \mathbf{u}_l\end{pmatrix}=\begin{pmatrix}\sigma_k^2\mathbf{I}_q & \sigma_{kl}\mathbf{I}_q\\ \sigma_{kl}\mathbf{I}_q & \sigma_l^2\mathbf{I}_q\end{pmatrix}$$

with the same dimension. Then

$$\sigma_{kl}^{[n+1]}=\left\{\hat{\mathbf{u}}_k^{[n]\prime}\hat{\mathbf{u}}_l^{[n]}+\mathrm{tr}\left[\underline{\mathbf{C}}_{kl}^{[n]}\right]\sigma_0^{2[n]}\right\}/q \qquad (3.44)$$

An appealing variant of this algorithm was formulated by Harville (1977). The idea is to rewrite (3.40) in substituting σ_k^2/η_k to σ_0^2 so that, after factorizing σ_k^2 on the left, one obtains

$$\sigma_k^{2[n+1]}=\hat{\mathbf{u}}_k^{[n]\prime}\hat{\mathbf{u}}_k^{[n]}/\left[q_k-\mathrm{tr}\left(\underline{\mathbf{C}}_{kk}^{[n]}\right)/\eta_k^{[n]}\right], \qquad (3.45)$$

that is used jointly with (3.42).

Apart from their simplicity, these two algorithms have the pleasing property of yielding positive values for the estimations of variance components provided they are started with strictly positive values $\sigma_k^{2[0]}>0$ for all variance components. In addition, in many examples, Harville's version of the algorithm turned out to be faster than that of Henderson.

Calculation of $-2\mathrm{L}_p$

Let us start with the expression of the profile log-likelihood as a function of $\boldsymbol{\gamma}$ after maximization for $\boldsymbol{\beta}$,

$$-2\mathrm{L}_P=N\ln(2\pi)+\ln|\mathbf{V}|+\mathbf{y}'\underline{\mathbf{P}}\mathbf{y}\,. \qquad (3.46)$$

First, we have already shown that in a linear mixed model such that $\mathbf{V}=\mathbf{ZGZ}'+\mathbf{R}$, one has $\underline{\mathbf{P}}\mathbf{y}=\mathbf{R}^{-1}\hat{\mathbf{e}}$ so that

$$\mathbf{y}'\underline{\mathbf{P}}\mathbf{y}=\mathbf{y}'\mathbf{R}^{-1}\mathbf{y}-\hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}\,, \qquad (3.47)$$

where $\mathbf{T}=(\mathbf{X},\mathbf{Z})$, and $\hat{\boldsymbol{\theta}}=\left(\hat{\boldsymbol{\beta}}',\hat{\mathbf{u}}'\right)'$ are solutions to the HMME.

Furthermore applying the rules of determinant computation by blocks to the following partition

$$\mathbf{A} = \begin{pmatrix} \mathbf{R}^{-1} & \mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}$$

gives

$$|\mathbf{A}| = |\mathbf{R}^{-1}|\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\right| = 1/|\mathbf{R}||\mathbf{G}|,$$

and symmetrically

$$|\mathbf{A}| = \left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|\left|\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{R}^{-1}\right| = \left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|/|\mathbf{V}|$$

which, being equal, leads to

$$|\mathbf{V}| = |\mathbf{R}||\mathbf{G}|\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|. \qquad (3.48)$$

Substituting (3.47) and (3.48) into (3.46), we can express $-2\mathrm{L}_p$ under a more explicit and computable form which can be applied to any Gaussian linear mixed model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})$

$$-2\mathrm{L}_P = N\ln 2\pi + \ln|\mathbf{R}| + \ln|\mathbf{G}| + \ln\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right| + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}. \qquad (3.49)$$

By definition, this formula can be employed at any stage of the search of $\boldsymbol{\gamma}$ which maximizes the log-likelihood, and in particular to determine the value of its maximum L_m

$$-2\mathrm{L}_m = -2\mathrm{L}_P(\mathbf{G} = \hat{\mathbf{G}}_{ML}, \mathbf{R} = \hat{\mathbf{R}}_{ML})$$

using the elements of the HMME.

Further simplifications arise in formula (3.49) for special structures of $\mathbf{R}$ and $\mathbf{G}$

Example 3.3 *Expression of the loglikehood when* $\mathbf{R} = \sigma_0^2\mathbf{I}_N$

We first consider this usual case of independent residuals with a homogeneous variance σ_0^2. Then,

$$\ln|\mathbf{R}| = N\ln\sigma_0^2 \;;\; \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} = \left(\mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1}\right)/\sigma_0^2,$$

and

$$\mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y} = \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y}\right)/\sigma_0^2.$$

Hence, $-2\mathrm{L}_P$ as a function of σ_0^2 and of the parameters (say $\mathbf{g}$) determining $\mathbf{G}$ is

$$-2\mathrm{L}_P\left(\mathbf{g},\sigma_0^2;\mathbf{y}\right) = N\ln 2\pi + (N-q)\ln\sigma_0^2 + \ln|\mathbf{G}| + \ln\left|\mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1}\right| + \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y}\right)/\sigma_0^2 \quad .(3.50)$$

Moreover, let us assume that $\mathbf{G} = \bigoplus_{k=1}^{K}\mathbf{G}_k$ with $\mathbf{G}_k = \sigma_k^2\mathbf{A}_k$ for $k = 1,...,K$ i.e., that the K random factors are also independent, then the previous expression becomes

$$-2\mathrm{L}_P = N\ln 2\pi + \left(N - \sum_{k=1}^{K} q_k\right)\ln\sigma_0^2 + \sum_{k=1}^{K} q_k\ln\sigma_k^2 + \sum_{k=1}^{K}\ln|\mathbf{A}_k| + \ln\left|\mathbf{Z}'\mathbf{Z} + \bigoplus_{k=1}^{K}\mathbf{A}_k^{-1}\left(\sigma_0^2/\sigma_k^2\right)\right| + \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y}\right)/\sigma_0^2$$

We can go a step further by setting as parameters σ_0^2 and $\boldsymbol{\eta} = \left({}_c\eta_k = \sigma_k^2/\sigma_0^2\right)_{1\le k\le K}$ as before, and consider the profile log-likelihood $\mathrm{L}_P^*\left(\boldsymbol{\eta};\mathbf{y}\right) = \mathrm{Max}_{\sigma_0^2}\,\mathrm{L}_P\left(\sigma_0^2,\boldsymbol{\eta};\mathbf{y}\right)$ obtained after maximization with respect to σ_0^2. Now, $\hat{\sigma}_0^2(\boldsymbol{\eta}) = \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{y}\right)/N$ as in Henderson's algorithm, and substituting this in the previous expression gives the profile log-likelihood

$$-2\mathrm{L}_P^*\left(\boldsymbol{\eta};\mathbf{y}\right) = N\left(\ln 2\pi + 1\right) + N\ln\hat{\sigma}_0^2(\boldsymbol{\eta}) + \sum_{k=1}^{K} q_k\ln\eta_k + \sum_{k=1}^{K}\ln|\mathbf{A}_k| + \ln\left|\mathbf{Z}'\mathbf{Z} + \bigoplus_{k=1}^{K}\mathbf{A}_k^{-1}/\eta_k\right| \quad . \quad (3.51)$$

Example 3.4 *ML estimation in the random coefficients models for growth data* (Example 1.6 continued)

The data considered he are a subsample (girls) of set due to Pothoff and Roy (1964) about facial growth measurements made on 27 children at 4 equidistant ages (8,10,12 and 14 years) with 4 missing values as defined by Little and Rubin (1987).

Table 3.1: Facial growth* measurements taken on 11 girls at 4 ages

	Age (years)			
Girl	8	10	12	14
1	210	200	215	230
2	210	215	240	255
3	205	NA	245	260
4	235	245	250	265
5	215	230	225	235
6	200	NA	210	225
7	215	225	230	250
8	230	230	235	240
9	200	NA	220	215
10	165	NA	190	195
11	245	250	280	280

*distance from the centre of the pituary to the pteryomaxillary fissure (unit 10^{-4}m)

Letting $t_j^* = \left(t_j - 8\right)/2$, the model can be written as

$$y_{ij} = \alpha + \beta t_j^* + a_i + b_i t_j^* + e_{ij}, \qquad (3.52)$$

where y_{ij} represents the j^{th} measurement taken at age t_j^* on the i^{th} individual. The fixed part $\alpha + \beta t_j^*$ describes the overall population profile with α being the (fixed intercept) at an age of 8 years, β the rate of growth for a 2-year period and the random part $a_i + b_i t_j^*$ corresponds to the subject-specific i deviation counterpart such that

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim_{\text{iid}} \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}\right].$$

The residuals e_{ij} are assumed $e_{ij} \sim_{\text{iid}} \mathcal{N}\left(0, \sigma_e^2\right)$.

Let us begin with the random intercept model $y_{ij} = \alpha + \beta t_j^* + a_i + e_{ij}$.

Here the HMME are

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \sigma_e^{2[n]} / \sigma_a^{2[n]} \mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}}^{[n]} \\ \hat{\mathbf{u}}^{[n]} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{pmatrix} \qquad (3.53)$$

where $\mathbf{X'X} = \begin{pmatrix} 40 & 62 \\ 62 & 150 \end{pmatrix}$, $\mathbf{X'Z} = \begin{pmatrix} 4 & 4 & 3 & 4 & 4 & 3 & 4 & 4 & 3 & 3 & 4 \\ 6 & 6 & 5 & 6 & 6 & 5 & 6 & 6 & 5 & 5 & 6 \end{pmatrix}$,

$\mathbf{Z'Z} = Diag\begin{pmatrix} 4 & 4 & 3 & 4 & 4 & 3 & 4 & 4 & 3 & 3 & 4 \end{pmatrix}$, $\mathbf{X'y} = \begin{pmatrix} 9115 & 14625 \end{pmatrix}'$, and

$\mathbf{Z'y} = \begin{pmatrix} 855 & 920 & 710 & 995 & 905 & 635 & 920 & 935 & 635 & 550 & 1055 \end{pmatrix}'$.

Starting values for the variance components can be chosen somewhat arbitrarily ; here we took $\sigma_e^{2[0]} = 200$ and $\sigma_a^{2[0]} = 100$. Based on these values, we solve (3.53) in $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ followed by (3.42) and (3.43) and obtain the following iterative scheme.

Table 3.2. Solutions to ML and MME equations by round of iteration

n	$\alpha^{[.]}$	$\beta^{[.]}$	$a_1^{[.]}$	$\sigma_e^{2[.]}$	$\sigma_a^{2[.]}$
1	211.91	9.64	-8.41	192.1207	207.2596
2	211.48	9.74	-10.02	129.0611	308.9038
3	211.19	9.81	-11.01	86.8348	374.0731
4	211.07	9.85	-11.42	68.4920	400.3699
5	211.02	9.86	-11.56	62.0439	408.8686
6	211.00	9.86	-11.63	59.9569	411.4967
10	211.00	9.86	-11.63	59.0118	412.6650
15	211.00	9.86	-11.63	59.0022	412.6756

Iterations were stopped when $\left(\left\|\boldsymbol{\theta}^{[n+1]}-\boldsymbol{\theta}^{[n]}\right\|^2 / \left\|\boldsymbol{\theta}^{[n]}\right\|^2\right)^{1/2} < \varepsilon$ where $\boldsymbol{\theta} = \left(\sigma_e^2, \sigma_a^2\right)'$ and $\varepsilon = 10^{-6}$.

The procedure provides in the same time as a by product solutions to the HMME for variance components set equal to their ML estimations: here we have

$\hat{\alpha} = 211.002 \pm 6.457$, $\hat{\beta} = 9.863 \pm 1.048$, $\hat{a}_1 = -11.631 \pm 7.114$,…,

$\hat{a}_{11} = 36.643 \pm 7.114$ and $-2\mathrm{L}_m = 312.5446$ (formula 3.51).

For instance, SAS-Proc Mixed on this data set gives $\hat{\sigma}_a^2 = 412.6400$ and $\hat{\sigma}_e^2 = 59.0040$, $\hat{\alpha} = 211.000 \pm 6.457$, $\hat{\beta} = 9.863 \pm 1.049$,

$\hat{a}_1 = -11.631 \pm 7.114$,…, $\hat{a}_{11} = 36.643 \pm 7.114$ and $-2\mathrm{L}_m = 312.5446$.

This algorithm extends very easily to the case of the "intercept+slope" model in (3.52). In the HMME given in (3.53) $\sigma_e^{2[n]}\left(\mathbf{G}^{[n]}\right)^{-1}$ replace $\sigma_e^{2[n]} / \sigma_a^{2[n]}\mathbf{I}_{11}$ where $\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \otimes \mathbf{I}_{11}$ if random effects are listed as $\mathbf{u}' = \left({}_r\mathbf{u}_i'\right)_{1 \le i \le 11}$ with $\mathbf{u}_i' = \left(a_i, b_i\right)$. The formulae for the variance components are (3.42) and (3.43) as before, the covariance estimate of σ_{ab} being updated via (3.44). Starting from $\sigma_e^{2[0]} = 50$, $\sigma_a^{2[0]} = 400$, $\sigma_{ab}^{[0]} = 0$ and $\sigma_b^{2[0]} = 20$, one obtains

Table 3.3. Solutions to Henderson's equations equations by round of iteration

n	$\sigma_e^{2[.]}$	$\sigma_a^{2[.]}$	$\sigma_{ab}^{[.]}$	$\sigma_b^{2[.]}$
1	38.1878	368.4004	3.1625	15.4757
10	36.9055	364.4838	7.7950	12.0136
20	37.2109	363.2651	8.0972	11.8291
30	37.2381	363.8096	8.1243	11.8126
43	37.2407	363.4243	8.1270	11.8110

The numerical process needs more iterations than previously (43 as compared to 14) for the same level of accuracy. Final estimations are $\hat{\sigma}_e^2 = 37.2408$,

$\hat{\sigma}_a^2 = 363.8043$, $\hat{\sigma}_{ab} = 8.1270$, $\hat{\sigma}_b^2 = 11.8110$ and $-2\mathrm{L}_m = 308.4897$ (formula 3.50). Again, these values are very similar to those obtained with SAS-Proc Mixed that are $\hat{\sigma}_e^2 = 37.2421$, $\hat{\sigma}_a^2 = 363.7800$, $\hat{\sigma}_{ab} = 8.1293$ and $\hat{\sigma}_b^2 = 11.8108$ $-2\mathrm{L}_m = 308.4898$. Finally, the reader is encouraged to replicate these procedures on the "Boy" sample (Table 3.4) of this data set (See Exercise 3.8).

3.2.5 Hypothesis testing

Tests of hypotheses may concern either fixed or random effects. Regarding fixed effects, since data are correlated with possibly heterogeneous variances, the usual Fisher-Snedecor statistic, which was derived for iid Gaussian data and linear models no longer applies. Regarding variance components, ratios of mean squares, which prevail in the context of ANOVA with balanced data, are also inappropriate in the unbalanced case even for purely random models.

For conciseness reasons, testing procedures for random effects are presented just once in the next part, since REML is precisely the method devoted to estimation of variance components.

Before going into the procedures themselves, let us recall the basic results of the asymptotic normality properties of maximum likelihood estimators which underlie all this testing theory.

Asymptotic normality

Let $\hat{\boldsymbol{\alpha}}_N$ be the ML estimator of $\boldsymbol{\alpha} \in \mathrm{A}$ from a data sample $\mathbf{y}_N$ of size N. Under the usual regularity conditions (compact parameter space, continuous log-likelihood function and continuously differentiable up to the third order, existing information matrix and its inverse), the sequence $\sqrt{N}\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}\right)$ converges in distribution to a centered Gaussian distribution having variance-covariance matrix $\mathrm{Lim}\, N\left[\mathbf{J}_N\left(\boldsymbol{\alpha}\right)\right]^{-1}$ when $N \to \infty$ (Sweeting, 1980; Mardia and Marshall, 1984). In short,

$$\sqrt{N}\left(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}\right) \underset{\mathcal{L}}{\to} \mathcal{N}\left(0, \mathrm{Lim}\, N\left[\mathbf{J}_N\left(\boldsymbol{\alpha}\right)\right]^{-1}\right), \qquad (3.54)$$

where $\mathbf{J}_N(\boldsymbol{\alpha}) = \mathrm{E}\left[-\partial^2 \mathrm{L}(\boldsymbol{\alpha};\mathbf{y}_N)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'\right]$ is the Fisher information matrix pertaining to $\boldsymbol{\alpha}$.

Since Lim $N\left[\mathbf{J}_N(\boldsymbol{\alpha})\right]^{-1}$ can be consistently estimated by $N\left[\mathbf{J}_N(\hat{\boldsymbol{\alpha}})\right]^{-1}$, one can construct the following asymptotic pivot (Leonard and Hsu, 1999, page 35):

$$\hat{\mathbf{J}}_N^{T/2}(\hat{\boldsymbol{\alpha}}_N - \boldsymbol{\alpha}) \underset{\mathcal{L}}{\rightarrow} \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{3.55}$$

where $\hat{\mathbf{J}}_N^{T/2}$ is the condensed notation for the Cholesky decomposition $\mathbf{J}_N(\hat{\boldsymbol{\alpha}}) = \hat{\mathbf{J}}_N = \hat{\mathbf{J}}_N^{1/2}\hat{\mathbf{J}}_N^{T/2}$.

The asymptotic property in (3.54) can be extended to a continuously differentiable function $\mathbf{g}(\boldsymbol{\alpha})$ (from $\mathbb{R}^p$ to $\mathbb{R}^q$)

$$\sqrt{N}\left[\mathbf{g}(\hat{\boldsymbol{\alpha}}_N) - \mathbf{g}(\boldsymbol{\alpha})\right] \underset{\mathcal{L}}{\rightarrow} \mathcal{N}\left(0, \mathrm{Lim}\, N \frac{\partial \mathbf{g}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'}\left[\mathbf{J}_N(\boldsymbol{\alpha})\right]^{-1}\frac{\partial \mathbf{g}'(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right), \tag{3.56}$$

where $\partial\mathbf{g}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}'$ indicates the $(q \mathrm{x}\, p)$ matrix having $\partial g_i(\boldsymbol{\alpha})/\partial\alpha_j$ as element (ij) and $\partial\mathbf{g}'(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ is its transpose.

The Wald statistic

Let us consider the test of the null hypothesis $\mathrm{H}_0 : \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ against its alternative $\mathrm{H}_1 : \mathbf{C}'\boldsymbol{\beta} \neq \mathbf{m}$ where $\mathbf{C}'$ is a $(r \mathrm{x}\, p)$ matrix whose r rows are linearly independent and $\mathbf{m}$ is a $(r \mathrm{x} 1)$ vector of constants often nil but not necessarily (see chapter 1, section 1.4).

We have seen that the asymptotic distributions of the ML estimators of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are independent and that the information matrix pertaining to $\boldsymbol{\beta}$ is $\mathbf{J}_\beta = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$. In such conditions, we can apply the previous results (3.54) and (3.55) to $\mathbf{C}'\hat{\boldsymbol{\beta}}$ so that, under H_0

$$\sqrt{N}\left(\mathbf{C}'\hat{\boldsymbol{\beta}} - \mathbf{m}\right) \underset{\mathcal{L}}{\rightarrow} \mathcal{N}\left(0, \mathrm{Lim}\, N\, \mathbf{C}'\mathbf{J}_\beta^{-}\mathbf{C}\right), \tag{3.57}$$

and, letting $\hat{\mathbf{J}}_\beta = \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}$,

$$\left[\left(\mathbf{C}'\hat{\mathbf{J}}_{\beta}^{-}\mathbf{C}\right)^{-1}\right]^{T/2}\left(\mathbf{C}'\hat{\boldsymbol{\beta}}-\mathbf{m}\right)\underset{\mathcal{L}}{\rightarrow}\mathcal{N}\left(\mathbf{0},\mathbf{I}_{r}\right). \qquad (3.58)$$

Hence, on multiplying (3.58) by its transpose, one obtains an asymptotic Chi-square distribution with r degrees of freedom

$$\left(\mathbf{C}'\hat{\boldsymbol{\beta}}-\mathbf{m}\right)'\left[\mathbf{C}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{C}\right]^{-1}\left(\mathbf{C}'\hat{\boldsymbol{\beta}}-\mathbf{m}\right)\underset{d}{\rightarrow}\chi_{r}^{2} \qquad (3.59)$$

that is precisely the distribution of the Wald statistic under H_0 for the test considered.

Notice the close similarity between this statistic and the one that would have been used if $\mathbf{V}$ had been known. The difference lies in $\mathbf{V}$ replaced here by its ML estimation $\hat{\mathbf{V}}$ and consequently by the Chi-square so obtained being now not the true distribution but an asymptotic one. This is why the property is often symbolized under the classical form

$$\mathbf{C}'\hat{\boldsymbol{\beta}}\rightarrow\mathcal{N}\left[\mathbf{C}'\boldsymbol{\beta},\ \mathbf{C}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{C}\right], \qquad (3.60)$$

Strictly speaking, although practical, this notation is an abuse since the limiting distribution of $\mathbf{C}'\hat{\boldsymbol{\beta}}$ is degenerate with a covariance matrix equal to zero. Only the notations shown in (3.58) and (3.59) are correct.

Several software (see e.g., Littell et al., 2006, page 756) offer a Fisher-Snedecor option for this test by analogy with the case when $\mathbf{V}$ is known a part from a constant. As a matter of fact when $\mathbf{V}=\mathbf{H}\sigma_0^2$ and $\mathbf{H}$ is known, and if one defines by $W\left(\sigma_0^2\right)$ the quantity $\left(\mathbf{C}'\hat{\boldsymbol{\beta}}-\mathbf{m}\right)'\left[\mathbf{C}'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{C}\right]^{-1}\left(\mathbf{C}'\hat{\boldsymbol{\beta}}-\mathbf{m}\right)$, then, under H_0, the statistic $W\left(\hat{\sigma}_0^2\right)/r$ has a Fisher-Snedecor distribution $\mathrm{F}\left[r,N-\mathrm{r}(\mathbf{X})\right]$ (see section 1.5, formulae 1.84 and 1.85). Here $\hat{\sigma}_0^2$ is the usual ML estimator $\left(\mathbf{y}'\mathbf{H}^{-1}\mathbf{y}-\hat{\boldsymbol{\beta}}'\mathbf{H}^{-1}\mathbf{y}\right)/N$ based on $\hat{\boldsymbol{\beta}}$, the GLS (or ML) estimator of $\boldsymbol{\beta}$. Similarly, in the general case, one forms $\hat{W}/r$, $\hat{W}$ being the Wald statistic in (3.59); its

value is compared with that of a $\mathrm{F}(r,d)$ where the number of degrees of freedom d is computed according to some approximation One of the most popular technique for calculating d is the Satterthwaite approximation. For $r=1$, it reduces to $d \approx 2SE^2 / \left[\mathrm{Var}\left(SE^2\right)\right]$ where SE is the asymptotic standard error of $\mathbf{C}'\hat{\boldsymbol{\beta}}$. SE^2 can be approximated by $\mathbf{C}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{C}$ or its MME equivalent. $\mathrm{Var}\left(SE^2\right)$is more tricky. Technical details are given in (Giesbrecht and Burns, 1985; Fai and Cornelius, 1996; Littell et al, 2006). Except in some special cases of balanced designs, the resulting distribution under H_0 is no longer an exact Fisher-Snedecor (or square of a Student for $r=1$). Therefore, although this procedure based on certain calculations of d might be practically efficient as proved by simulation (McBride, 2000; Shaalje et al., 2002), it remains an approximation with no clear theoretical foundation.

The likelihood ratio statistic

An alternative to Wald's test lies in the likelihood ratio test also known as the Neyman-Pearson test. It can be formulated as follows: see e.g., Mood et al., (1974), page 419; Cox and Hinkley, (1974), page 322

$$\mathrm{H}_0 : \{\boldsymbol{\beta} \in \mathrm{B}_0\} \mathrm{x} \{\boldsymbol{\gamma} \in \Gamma\} \text{ versus } \mathrm{H}_1 : \{\boldsymbol{\beta} \in (\mathrm{B} \setminus \mathrm{B}_0)\} \mathrm{x} \{\boldsymbol{\gamma} \in \Gamma\}.$$

For instance, in the previous example B corresponds to $\mathbb{R}^p$ and B_0 is the real subspace of dimension $p-r$ pertaining to $\mathbb{R}^p$ constrained by the r relationships $\mathrm{H}_0 : \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$.

Now, let us consider the maximum of the log-likelihood $\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) = \log \mathrm{p}(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma})$ under the two conditions H_0 (reduced model) and $\mathrm{H}_0 \cup \mathrm{H}_1$ (complete model)

$$L_{R,m} = \mathrm{Max}_{\boldsymbol{\beta}\in\mathrm{B}_0,\boldsymbol{\gamma}\in\Gamma} \mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) \ ; \ L_{C,m} = \mathrm{Max}_{\boldsymbol{\beta}\in\mathrm{B}_0,\boldsymbol{\gamma}\in\Gamma} \mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}).$$

It has been shown (Cox and Hinkley, 1974) that the statistic $\lambda = -2\mathbf{L}_{R,m} + 2\mathbf{L}_{C,m}$ contrasting $-2L_m$ between the reduced and the complete models respectively,

has under H_0 an asymptotic Chi-square distribution, the number of degrees of freedom of which being the difference between the dimension of B and that of B_0. In short

$$\lambda = -2\mathbf{L}_{R,m} + 2\mathbf{L}_{C,m}\Big|_{H_0} \underset{d}{\to} \chi^2_{\dim(B)-\dim(B_0)}. \tag{3.61}$$

Let us go back to the usual test of $H_0 : \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ against its alternative $H_1 : \mathbf{C}'\boldsymbol{\beta} \neq \mathbf{m}$. How do we test such assumptions via the likelihood ratio test? This is obvious when $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ can be translated as a simple reduced model (see forthcoming example). However, this is not always the case as for instance in a μ_{ij} model or when testing some specific interaction contrasts. When the reduced model cannot be made explicit directly, we have to specify it as $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with the constraint $\mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$. Thus, maximizing the log-likelihood function requires to take into account such a constraint. This is accomplished by forming the function

$$L^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}; \mathbf{y}) = L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) + \boldsymbol{\theta}'(\mathbf{C}'\boldsymbol{\beta} - \mathbf{m}),$$

where $\boldsymbol{\theta}$ is a $(r \times 1)$ vector of Lagrange multipliers.

Differentiating $L^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}; \mathbf{y})$ with respect to $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ gives

$$\partial L^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \mathbf{C}\boldsymbol{\theta},$$

$$\partial L^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\theta} = \mathbf{C}'\boldsymbol{\beta},$$

$$\partial L^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\gamma} = \partial L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\gamma}.$$

Setting the first two derivatives to zero enables to construct the profile likelihood $L^*_P(\boldsymbol{\gamma}; \mathbf{y}) = \max_{\boldsymbol{\beta}, \boldsymbol{\theta}} L^*(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}; \mathbf{y})$ i.e.,

$$-2L^*_P(\boldsymbol{\gamma}; \mathbf{y}) = N \ln(2\pi) + \ln|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

where $\tilde{\boldsymbol{\beta}}$ is a solution to the system

$$\begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{C} \\ \mathbf{C}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\theta}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{m} \end{pmatrix}. \tag{3.62}$$

It remains to maximize $\mathrm{L}_P^*(\boldsymbol{\gamma};\mathbf{y})$ with respect to $\boldsymbol{\gamma}$. In the case of standard linear mixed models such that $\mathbf{V}=\sum_{k=1}^{K}\sigma_k^2\mathbf{Z}_k\mathbf{Z}_k^{'}+\sigma_0^2\mathbf{I}_N$, this can be done for instance via the Henderson algorithm provided some appropriate modifications are carried out.

First, on account of (3.62), the mixed model equations become

$$\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{C}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\sigma_0^2\mathbf{G}^{-1} & \mathbf{0}\\ \mathbf{C}' & \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\mathbf{u}}\\ \tilde{\boldsymbol{\theta}}^*\end{pmatrix}=\begin{pmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\\ \mathbf{m}\end{pmatrix}, \qquad (3.63)$$

where $\tilde{\boldsymbol{\theta}}^*=\tilde{\boldsymbol{\theta}}\sigma_0^2$.

Two, the algorithm for σ_0^2 also needs some adaptation. Remember this algorithm is based on expressing $\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)'\mathbf{H}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)$ as a function of $\mathbf{y}$ and of the BLUP $\tilde{\mathbf{e}}$ of $\mathbf{e}$ Here $\tilde{\mathbf{e}}=\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}-\mathbf{Z}\tilde{\mathbf{u}}$ with $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ solutions to (3.63). Given $\mathbf{R}=\sigma_0^2\mathbf{I}_N$ and $\mathbf{V}=\sigma_0^2\mathbf{H}$, $\tilde{\mathbf{e}}$ can also be written as $\tilde{\mathbf{e}}=\mathbf{H}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)$ so that

$$\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)'\mathbf{H}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right)=\mathbf{y}'\tilde{\mathbf{e}}-\tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{H}^{-1}\left(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}\right).$$

In the standard case of a model without constraint, the last term is zero. But here, it is not as shown in (3.62) and is equal to $\sigma_0^2\tilde{\boldsymbol{\beta}}'\mathbf{C}\tilde{\boldsymbol{\theta}}=\tilde{\boldsymbol{\beta}}'\mathbf{C}\tilde{\boldsymbol{\theta}}^*$ or alternatively $\mathbf{m}'\tilde{\boldsymbol{\theta}}^*$. Hence, the formula to iterate with is

$$\sigma_0^{2[t+1]}=\left(\mathbf{y}'\mathbf{y}-\tilde{\boldsymbol{\beta}}^{[t]'}\mathbf{X}'\mathbf{y}-\tilde{\mathbf{u}}^{[t]'}\mathbf{Z}'\mathbf{y}-\tilde{\boldsymbol{\theta}}^{*[t]'}\mathbf{m}\right)/N. \qquad (3.64)$$

Notice in passing how remarkable the extension of Henderson's formula to this case is. The formula for $\sigma_k^{2[t+1]}$ remains unchanged as in (3.42). After having iterated with $\tilde{\boldsymbol{\beta}}\left(\boldsymbol{\gamma}^{[k]}\right)$ and $\boldsymbol{\gamma}^{[k]}$, we can easily calculate the maximum $\mathrm{L}_{R,m}$ of $\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y})$ under H_0

$$-2\mathrm{L}_{R,m}=N\left(\ln 2\pi+1\right)+N\ln\tilde{\sigma}_0^2+\sum_{k=1}^{K}q_k\ln\tilde{\eta}_k+\ln\left|\mathbf{Z}'\mathbf{Z}+\bigoplus_{k=1}^{K}\mathbf{I}_{q_k}/\tilde{\eta}_k\right| \qquad (3.65)$$

and contrast it with $-2\mathrm{L}_{C,m}$ of the complete model.

The score statistic

Under the same conditions as previously, the score test proposed by Rao (1973) relies on the following statistic

$$\tilde{U} = \tilde{\mathbf{S}}_{\beta}^{'} \tilde{\mathbf{J}}_{\beta}^{-} \tilde{\mathbf{S}}_{\beta}, \tag{3.66}$$

where $\tilde{\mathbf{S}}_{\beta}$ is the value of the score function $\mathbf{S}_{\beta} = \partial L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\beta}$ evaluated at the point, $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$ of the ML estimations obtained under the reduced model, and $\tilde{\mathbf{J}}_{\beta}$ is the corresponding value of the Fisher information matrix $\mathbf{J}_{\beta} = -\mathrm{E}\left[\partial^2 \mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'\right]$ pertaining to $\boldsymbol{\beta}$ evaluated at the same point $\tilde{\mathbf{J}}_{\beta} = \mathbf{J}_{\beta}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$.

The basic idea underlying this test is very simple. If $\mathrm{U}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{S}_{\beta}^{'} \mathbf{J}_{\beta} \mathbf{S}_{\beta}$ were evaluated at the point $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$ of the ML estimations obtained under the complete model, then $\mathrm{U}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ would be zero as from definition of the score: $\mathbf{S}_{\beta}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}; \mathbf{y}) = \mathbf{0}$. Evaluated at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$, this quadratic form is a measure of distance from its nil reference value. If the value is close to zero, one will tend to accept H_0; on the contrary, the higher the value of U, the higher the probability to reject H_0.

As before, this statistic has, under the null hypothesis, an asymptotic Chi-square distribution, the number of degrees of freedom of which being the difference between the numbers of parameters of the complete and reduced models. In short,

$$\mathrm{U}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})\Big|_{\mathrm{H}_0} \underset{d}{\rightarrow} \chi^2_{\dim(\mathrm{B}) - \dim(\mathrm{B}_0)}. \tag{3.67}$$

Let us come back to testing $H_0 : \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ against $H_1 : \mathbf{C}'\boldsymbol{\beta} \neq \mathbf{m}$. From the definition of $L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y})$, the score function is $S_\beta = \mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and the Fisher information matrix $\mathbf{J}_\beta = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$. Evaluated under H_0, the score function reduces to $\tilde{\mathbf{S}}_\beta = \mathbf{C}\tilde{\boldsymbol{\theta}}$. This is so because following (3.62) $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{C}\tilde{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. Hence, the expression for the score statistic is

$$U = \tilde{\boldsymbol{\theta}}'\mathbf{C}'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{C}\tilde{\boldsymbol{\theta}}. \qquad (3.68)$$

As previously, U can be easily computed from the elements of HMME in (3.63) as

$$U = \left(\tilde{\boldsymbol{\theta}}^{*'}\mathbf{C}'\tilde{\mathbf{C}}_{\beta\beta}\mathbf{C}\tilde{\boldsymbol{\theta}}^{*}\right) / \tilde{\sigma}_0^2 \qquad (3.69)$$

where $\tilde{\mathbf{C}}_{\beta\beta}$ is the block pertaining to $\tilde{\boldsymbol{\beta}}$ in a generalized inverse of the coefficient matrix of (3.63) and $\tilde{\sigma}_0^2$ is the ML estimator of the residual variance calculated as in (3.64).

Example 3.5 *Testing for* $\boldsymbol{\beta}_2 = \mathbf{0}$ in $E(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$

The test for the absence of effects of some covariates is especially interesting to consider as this question often arises in practice. Due to the formulation of H_0, one can easily derive the expression of the likelihood ratio statistic. This is done by contrasting the reduced model (R) $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}$ to the complete model (C) $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$ assuming both models have the same covariance structure $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$. For the sake of simplicity, $\mathbf{X}_1$ and $\mathbf{X}_2$ are taken as full rank matrices with sizes $(N \times p_1)$ and $(N \times p_2)$ respectively.

Letting $L_{R,m}$ and $L_{C,m}$ be the maximum of the log-likelihood function under the reduced and complete models respectively, one can write

$$-2L_{R,m} = N \ln 2\pi + \ln\left|\tilde{\mathbf{V}}\right| + \mathbf{y}'\underline{\tilde{\mathbf{P}}}_1\mathbf{y}, \qquad (3.70)$$

$$-2L_{C,m} = N\ln 2\pi + \ln\left|\hat{\mathbf{V}}\right| + \mathbf{y}'\underline{\hat{\mathbf{P}}}\mathbf{y}\,, \tag{3.71}$$

where $\tilde{\mathbf{V}} = \mathbf{V}(\tilde{\boldsymbol{\gamma}})$, $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\gamma}})$ are the estimations of the variance covariance functions under the R and C models respectively, $\underline{\mathbf{P}}_1 = \mathbf{V}^{-1}\left(\mathbf{I}_N - \mathbf{Q}_1\right)$ and $\underline{\mathbf{P}} = \mathbf{V}^{-1}\left(\mathbf{I}_N - \mathbf{Q}\right)$ with $\mathbf{Q}_1 = \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1\right)^{-}\mathbf{X}_1'\mathbf{V}^{-1}$ and $\mathbf{Q} = \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$, $\mathbf{X}$ being $\left(\mathbf{X}_1, \mathbf{X}_2\right)$.

Finally, the test statistic λ is obtained as $\lambda = -2L_{R,m} + 2L_{C,m}$, the P-value of which is determined from its asymptotic Chi-square distribution with p_2 degrees of freedom. Expressions (3.70) and (3.71) can also be computed from the elements of the HMME: see (3.49) and example 3.3.

Conceptually, the simplest test to apply in this situation is the Wald test. By application of (3.59) in the case of $\mathrm{H}_0 : \mathbf{C}'\boldsymbol{\beta} = \mathbf{m}$ (here $\boldsymbol{\beta}_2 = \mathbf{0}$), the test statistic turns out to be:

$$W = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2\,, \tag{3.72}$$

where $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}})$ is the solution to $\left(\mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{X}_2\right)\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\gamma}}) = \mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{y}$, with $\underline{\hat{\mathbf{P}}}_1$ being as before but evaluated at $\underline{\mathbf{P}}_1 = \underline{\mathbf{P}}_1(\hat{\boldsymbol{\gamma}})$, the ML estimation of $\hat{\boldsymbol{\gamma}}$ under the complete model.

From the definition of $\hat{\boldsymbol{\beta}}_2$, one can derive an alternative expression for W i.e.,

$$W = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\underline{\hat{\mathbf{P}}}_1\mathbf{y}\,, \tag{3.73}$$

which will be useful later on for comparing W to the score statistic.

Notice that again, (3.72) or (3.73) can be easily computed from the HMME. Assuming for instance that $\mathbf{R} = \sigma_0^2\mathbf{I}_N$, these are

$$\begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \mathbf{X}_1'\mathbf{Z} \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \mathbf{X}_2'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X}_1 & \mathbf{Z}'\mathbf{X}_2 & \mathbf{Z}'\mathbf{Z} + \sigma_0^2\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}. \tag{3.74}$$

Thus, (3.72) is simply $W = \hat{\sigma}_0^2 \hat{\boldsymbol{\beta}}_2' \hat{\mathbf{C}}_{22}^{-1} \hat{\boldsymbol{\beta}}_2$ where $\hat{\mathbf{C}}_{22}$ is the block pertaining to $\hat{\boldsymbol{\beta}}_2$ in the inverse of the coefficient matrix.

Here the value of the score function reduces to

$$\tilde{\mathbf{S}}_\beta = \begin{pmatrix} \mathbf{0} \\ \mathbf{X}_2' \tilde{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_1) \end{pmatrix},$$

where $\tilde{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_1(\tilde{\boldsymbol{\gamma}})$ and $\tilde{\mathbf{V}} = \mathbf{V}(\tilde{\boldsymbol{\gamma}})$ are the ML estimations of $\boldsymbol{\beta}_1$ and $\mathbf{V}(\boldsymbol{\gamma})$ under the reduced model $\mathrm{E}(\mathbf{y}) = \mathbf{X}_1 \boldsymbol{\beta}_1$.

The zero term in $\tilde{\mathbf{S}}_\beta$ comes from the nil value of the score function i.e., $\mathbf{X}_1' \tilde{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_1) = \mathbf{0}$ under the reduced model. This result can also be obtained by applying (3.68) to the system (3.62) with $\mathbf{C} = \left(\mathbf{0}_{(p_2 \times p_1)}, \mathbf{I}_{p_2} \right)$ and $\tilde{\boldsymbol{\beta}}_2 = \mathbf{0}$ so that $\mathbf{C}\boldsymbol{\theta} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\theta}_2 \end{pmatrix}$ with $\boldsymbol{\theta}_2$ being equal to $\mathbf{X}_2' \tilde{\mathbf{V}}^{-1} \mathbf{X}_1 \tilde{\boldsymbol{\beta}}_1 + \boldsymbol{\theta}_2 = \mathbf{X}_2' \tilde{\mathbf{V}}^{-1} \mathbf{y}$.

Now $\mathbf{X}_2' \tilde{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}_1)$ is precisely $\mathbf{X}_2' \underline{\tilde{\mathbf{P}}}_1 \mathbf{y}$ with $\underline{\tilde{\mathbf{P}}}_1$ evaluated at $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$. Then, since $\mathbf{J}_\beta = \mathbf{X}' \mathbf{V}^{-1} \mathbf{X}$, the score statistic turns out to be

$$\tilde{U} = \mathbf{y}' \underline{\tilde{\mathbf{P}}}_1 \mathbf{X}_2 \left(\mathbf{X}_2' \underline{\tilde{\mathbf{P}}}_1 \mathbf{X}_2 \right)^{-1} \mathbf{X}_2' \underline{\tilde{\mathbf{P}}}_1 \mathbf{y}.$$

Letting $\tilde{\boldsymbol{\beta}}_2$ be the solution to $\left(\mathbf{X}_2' \underline{\tilde{\mathbf{P}}}_1 \mathbf{X}_2 \right) \tilde{\boldsymbol{\beta}}_2 = \mathbf{X}_2' \underline{\tilde{\mathbf{P}}}_1 \mathbf{y}$, then $\tilde{U}$ can also be written as

$$\tilde{U} = \tilde{\boldsymbol{\beta}}_2' \mathbf{X}_2' \underline{\tilde{\mathbf{P}}}_1 \mathbf{y}. \qquad (3.75)$$

Thus, the score statistic in (3.75) presents the same form as the Wald statistic in (3.73), the only difference being that the first is based on the ML estimation $\tilde{\mathbf{V}} = \mathbf{V}(\tilde{\boldsymbol{\gamma}})$ of $\mathbf{V}$ under the reduced model while the second utilises $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\gamma}})$, the estimation under the complete model. Therefore, computationally speaking, there is little if anything to gain in applying the score test as compared to that of Wald.

Example 3.6 *Testing the homogeneity of slopes between genders in the analysis of growth data* (Example 1.6 and 3.4 continued)

Here, the data comprises both the girl and boy samples measured at 4 equidistant ages (8,10,12 and 14 years) with 9 missing values as defined by Little and Rubin (1987). We will use the same notations as in example 1.6 with i refering to gender ($i=1,2$ for boys and girls respectively), j to measurement $j=1,...,4$ and k to individual within gender ($k=1,...,16$ for $i=1$, and $k=1,...,11$ for $i=2$).

Table 3.4. Facial growth measurements taken on 16 boys at 4 ages

	Age (years)			
Boy	8	10	12	14
1	260	250	290	310
2	215		230	265
3	230	225	240	275
4	255	275	265	270
5	200		225	260
6	245	255	270	285
7	220	220	245	265
8	240	215	245	255
9	230	205	310	260
10	275	280	310	315
11	230	230	235	250
12	215		240	280
13	170		260	295
14	225	255	255	260
15	230	245	260	300
16	220		235	250

Let $\mathbf{y}_{ik}=\left(y_{ijk}\right)$, $\mathbf{e}_{ik}=\left(e_{ijk}\right)$, $\boldsymbol{\beta}=(\alpha_1,\alpha_2-\alpha_1,\beta_1,\beta_2-\beta_1)'$, $\mathbf{u}_{ik}=(a_{ik},b_{ik})'$, $\mathbf{X}_{ik}=(\mathbf{1}_4,\mathbf{0}_4,\mathbf{t},\mathbf{0}_4)$ if $i=1$, $\mathbf{X}_{ik}=(\mathbf{1}_4,\mathbf{1}_4,\mathbf{t},\mathbf{t})$ if $i=2$ and $\mathbf{Z}_{ik}=(\mathbf{1}_4,\mathbf{t})$ with

$\mathbf{t} = (t_1, t_2, t_3, t_4)'$, then the model can be written under its typical linear mixed model form

$$\mathbf{y}_{ik} = \mathbf{X}_{ik}\boldsymbol{\beta} + \mathbf{Z}_{ik}\mathbf{u}_{ik} + \mathbf{e}_{ik}, \qquad (3.76)$$

where $\mathbf{u}_{ik} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{G}_0)$ and $\mathbf{e}_{ik} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{R}_0)$. We will assume that $\mathbf{R}_0 = \sigma_e^2 \mathbf{I}_4$ and $\mathbf{G}_0 = \sigma_a^2 \mathbf{I}_{27}$ (random intercept model). Data are shown on tables 3.1 and 3.4 Estimating parameters under the complete (C) model (Gender + Age + Gender*Age) via the Henderson's algorithm gives:

Table 3.5. ML estimates under two models for growth measurements

Parameters	Complete model	Reduced model
α_1	211.3364±6.5109	205.8594±6.1090
$\alpha_2 - \alpha_1$	14.6318±8.4495	23.8888±7.4950
β_1	9.7795±1.9381	13.3062±1.2826
$\beta_2 - \beta_1$	5.9550±2.5160	
σ_e^2	201.7364	217.3041
σ_a^2	309.5281	305.8440
$-2L_m$	857.2247	862.6231

In order to test the nul hypothesis that $\Delta_\beta = \beta_2 - \beta_1 = 0$, we can apply first the Wald test based on the ML estimation $\hat{\Delta}_\beta$ of Δ_β and its standard error SE under the complete model. This gives $\hat{\Delta}_\beta / SE = 2.3669$ giving a P-value of $P(\chi_1^2 \geq 2.3669^2) = 0.0179$. We can also use the F-type test (here equivalent to a t-test on the square root of the corresponding statistics). With the Satterthwaite approximation, $d = 72$ (according to SAS-Proc Mixed) so that the adjusted P-value becomes $P(F_{1,72} \geq 2.3669^2) = 0.0206$.

By contrasting $-2L_m$ between the two models, one has $\lambda = 862.6231 - 857.2247$ that is $\lambda = 5.3984$ leading to a P-value of $P\left(\chi_1^2 \geq 5.3984\right) = 0.0201$. Finally, as far as the score test is concerned, we just have to pick the variance component estimations from the reduced model $\sigma_e^2 = 217.3041$ and $\sigma_a^2 = 305.8440$, plug them into the HMME (3.74) and solve them. Doing so gives $\tilde{\Delta}_\beta = 5.9570$ and $SE_\beta = 2.6111$ so that the corresponding P-value is $P\left(\chi_1^2 \geq 2.2813^2\right) = 0.0225$ which is very close to the F-type Wald test and the LRT.

Discussion

The three statistics (Wald, Neyman-Pearson, Rao) have the same asymptotic properties under the null hypothesis (Rao, 1973; Gourieroux and Montford, 1989).

First, with respect to the asymptotic conditions, it is important to question their applicability given the data and model structures. Is the number N of observations or experimental units large enough? What is expected when this number increases? Does the size p of $\boldsymbol{\beta}$ increase accordingly or not? If so, what happens to the ratio N / p. Most authors assumed p being fixed or bounded or, at least, increasing at a slower rate than N to establish the asymptotic properties of ML estimators. One must ask such questions before applying the corersponding tests blindly.

Secondly, the debate remains open about the relative merits of these three tests for finite samples with however, a tendency to prefer the likelihood ratio test. The likelihood ratio test requires contrasting two models: the complete and reduced models whereas the Wald test only requires running the complete model. But, it also has the disadvantage of not being functionally invariant; it is actually a quadratic Taylor expansion of the log-likelihood function about the parameter value around its maximum. This point might be important to ponder in some regression analysis problems. We have seen here that there is little to

gain in computations by applying the score statistic for testing the absence of effects of some covariates.

A further point deserves careful attention. By construction, the reduced model corresponding to H_0 must be nested within the complete model for applying the likelihood ratio test which imposes some restrictions on the kind of hypotheses that can be tested.

For instance, comparison of models with respect to fixed effects must be carried out assuming the same variance covariance structure (see Exercise 3.9). Similarly, comparison of $\mathbf{V}$ structures requires the same expectation model. These constraints of the testing procedure raise critical issues on how to eventually choose the two structures in linear mixed models? One way to circumvent such a circularity may be to recourse to robust testing procedures. For instance, in the case of repeated data $\mathbf{y}_i = \left({}_c y_{ij} \right)_{1 \le j \le n_i}$ on the same experimental unit i, the testing procedure developed by Liang and Zeger (1986) allows to get rid of the uncertainty in the true variance covariance structure. The procedure known as GEE (Generalized Estimating Equations) is based on the so-called "sandwich" estimator of the sampling variance of the LS estimator $\hat{\boldsymbol{\beta}}$ with

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-} \sum_{i=1}^{I} \mathbf{X}_i' \mathbf{W}_i \mathbf{y}_i \,, \tag{3.77}$$

$$\begin{aligned} \mathrm{Var}(\mathbf{C}'\hat{\boldsymbol{\beta}}) = \mathbf{C}' & \left(\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-} \left(\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i \mathbf{X}_i \right) \\ & \left(\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-} \mathbf{C} \end{aligned} \,, \tag{3.78}$$

where $\mathbf{W}_i$ is a working matrix of weights and $\mathbf{V}_i$ is the variance of $\mathbf{y}_i$ which is replaced by a consistent estimator $\hat{\mathbf{V}}_i = \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} \right) \left(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} \right)'$.

The simplest choice for the working matrix is $\mathbf{W}_i = \mathbf{I}_{n_i}$ leading to OLS. As already seen in chapter 1, OLS provides an unbiased estimator of $\boldsymbol{\beta}$ whatever

the true unknown $\mathbf{V}$ is (see Exercise 3.10 for an application to the same data set and models as in example 3.6). However, in that case of longitudinal data, there is an additional important condition required to preserve the unbiasedness property which consists of no missing data for each subject or missing data patterns completely at random.

3.3 Restricted maximum likelihood

3.3.1 Classical presentation

A simple example

At first glance, the question arises as to why introducing a new procedure for estimating variance components given all the desirable properties of maximum likelihood. One way to tackle that issue is to consider the very simple example of estimating the variance from a N-sample of independent Gaussian observations $y_i \sim_{iid} \mathcal{N}\left(\mu,\sigma^2\right)$ with expectation μ and variance σ^2

Letting $\overline{y}=\left(\sum_{i=1}^{N} y_i\right)/N$ designate the mean of the observations, and $s^2=\sum_{i=1}^{N}\left(y_i-\overline{y}\right)^2/N$, the so-called sample variance. As shown in example (3.2), $-2\mathrm{L}\left(\mu,\sigma^2;\mathbf{y}\right)$ can be written as

$$-2\mathrm{L}\left(\mu,\sigma^2;\mathbf{y}\right)=N\left[\ln 2\pi+\ln\sigma^2+\frac{s^2+\left(\overline{y}-\mu\right)^2}{\sigma^2}\right]$$

with its partial derivatives with respect to μ and σ^2 :

$$\partial\left(-2\mathrm{L}/N\right)/\partial\mu=-2\left(\overline{y}-\mu\right)/\sigma^2,$$

$$\partial\left(-2\mathrm{L}/N\right)/\partial\sigma^2=\frac{1}{\sigma^2}-\frac{s^2+\left(\overline{y}-\mu\right)^2}{\sigma^4}.$$

Setting them to zero yields

$$\hat{\mu}=\overline{y}\ , \tag{3.79}$$

and, for $N\geq 2$,

$$\hat{\sigma}^2=s^2+\left(\overline{y}-\hat{\mu}\right)^2=s^2. \tag{3.80}$$

Now $\mathrm{E}\left(s^2\right)=\left[\sum_{i=1}^{N}\mathrm{E}\left(y_i-\bar{y}\right)^2\right]/N$, and $\mathrm{E}\left(y_i-\bar{y}\right)^2=\mathrm{Var}\left(y_i-\bar{y}\right)=\frac{N-1}{N}\sigma^2$,

so that

$$\mathrm{E}\left(\hat{\sigma}^2\right)=\left(N-1\right)\sigma^2/N \tag{3.81}$$

As a result, the ML estimator s^2 of σ^2 est biased downwards and underestimates σ^2 with a value of the bias $-\sigma^2/N$ being a decreasing function of the sample size N.

More generally for $y_{ij}\sim_{id}\mathcal{N}\left(\mathbf{x}_{ij}^{'}\boldsymbol{\beta},\sigma^2\right)$ for $1\le i\le I$ and $1\le j\le J$ where $\boldsymbol{\beta}$ is a $\left(p\,\mathrm{x}1\right)$ vector of real valued coefficients, the ML estimator of σ^2 is

$$\hat{\sigma}^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij}-\mathbf{x}_{ij}^{'}\hat{\boldsymbol{\beta}}\right)^2/N \tag{3.82}$$

with expectation

$$E\left(\hat{\sigma}^2\right)=\left(N-p\right)\sigma^2/N \tag{3.83}$$

where $N=IJ$ and $\hat{\boldsymbol{\beta}}$ is the OLS estimator of $\boldsymbol{\beta}$ and

Clearly, the ratio p/N is critical in this matter and the bias may be become very large in some particular situations (see exercise 3.11). That observation was at the origin of the development of this new concept of residual (restricted) maximum likelihood (REML).

How to correct this bias? As shown clearly from (3.82), the issue is due to some inteference between the estimation of μ and that of σ^2. Two procedures can be envisioned to avoid it that prefigure the methods used later on for the general linear mixed model.

i) Factorization of the likelihood

The principle is as follows. The likelihood is written as the product of two parts and only that part which does not depend on μ is kept to estimate σ^2. To that respect, one introduces the one to one transformation

$$\mathbf{y}=\left(y_i\right)_{1\le i\le N}\leftrightarrow\mathbf{y}^{*}=\left(\mathbf{z}',\bar{y}\right), \tag{3.84}$$

where $\mathbf{z} = \left({}_{c} z_i = y_i - \overline{y} \right)_{1 \le i \le N-1}$ is a vector made of $N-1$ deviations from the average ; here the $N-1$ first ones are taken but the choice of these $N-1$ values does not matter. Due to the nature of this transformation, one can relate the distribution of $\mathbf{y}$ to that of $\mathbf{y}^*$

$$f_Y(\mathbf{y}) = f_{Y^*}\left(\mathbf{y}^*\right)\left|\mathbf{J}\right|, \tag{3.85}$$

where $\left|\mathbf{J}\right|$ is the absolute value of the jacobian determinant $\mathbf{J} = \det\left({}_{m}\partial y_j^* / \partial y_i \right)_{1 \le i \le N,\ 1 \le j \le N}$.

First, $\left|\mathbf{J}\right|$ does not depend of the parameters as seen from the definition of $\mathbf{y}^*$; secondly, $\overline{y}$ and $\mathbf{z}$ are independent, and thirdly, the distribution of $\mathbf{z}$ does not depend on μ so that

$$f_Y(\mathbf{y}) \propto f_Z\left(\mathbf{z} \mid \sigma^2\right) f_{\overline{Y}}\left(\overline{y} \mid \mu, \sigma^2\right), \tag{3.86}$$

or equivalently in terms of loglikelihood

$$\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = \mathrm{L}_1(\sigma^2; \mathbf{z}) + \mathrm{L}_2(\mu, \sigma^2; \overline{y}) + \text{constant}, \tag{3.87}$$

where $\mathrm{L}_1(\sigma^2; \mathbf{z}) = \log f_Z\left(\mathbf{z} \mid \sigma^2\right)$, $\mathrm{L}_2(\mu, \sigma^2; \overline{y}) = \log f_{\overline{Y}}\left(\overline{y} \mid \mu, \sigma^2\right)$ and the constant being equal to $\log\left|\mathbf{J}\right|$.

The idea behind REML consists of only using $\mathrm{L}_1(\sigma^2; \mathbf{z})$ to estimate σ^2 thus explaining the term “residual likelihood” given by Thompson to this function as it literally designates the likelihood of such quantities $\mathbf{z}$. There are different ways to compute $\mathrm{L}_1(\sigma^2; \mathbf{z})$. By direct specification of the distribution of $\mathbf{z}$ that is $\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_Z\right)$ with

$$\mathbf{V}_Z = \sigma^2\left(\mathbf{I}_{N-1} - \mathbf{J}_{N-1} / N\right), \tag{3.88}$$

one arrives at

$$-2\mathrm{L}_1(\sigma^2; \mathbf{z}) = (N-1)\left(\ln 2\pi + \ln \sigma^2\right) - \ln N + Ns^2 / \sigma^2. \tag{3.89}$$

Differentiating with respect to σ^2 gives

$$\frac{\partial(-2\mathrm{L}_1)}{\partial\sigma^2}=\left[(N-1)\sigma^2-Ns^2\right]/\sigma^4$$

and setting to zero gives for $N \geq 2$

$$\hat{\sigma}^2 = Ns^2/(N-1), \qquad (3.90)$$

which is the usual unbiased estimator of σ^2 as already seen in (1.34).

ii) Treating μ as missing (Foulley, 1993)

If μ were known, then the ML estimator of σ^2 would be $\hat{\sigma}^2 = s^2 + (\bar{y}-\mu)^2$ whose value is always higher or equal than s^2. But μ being generally unknown, one may think to predict its contribution to $(\bar{y}-\mu)^2$ by replacing that term by its conditional expectation $\mathrm{E}\left[(\bar{y}-\mu)^2 \middle| \mathbf{y},\sigma^2\right]$ given the data. Remember that writing $\frac{\bar{y}-\mu}{\sqrt{\sigma^2/N}} \sim \mathcal{N}(0,1)$ can be interpreted, either as $\bar{y} \mid \mu,\sigma^2 \sim \mathcal{N}(\mu,\sigma^2/N)$, or as $\mu \mid \bar{y},\sigma^2 \sim \mathcal{N}(\bar{y},\sigma^2/N)$. Then, using this last "fiducial" interpretation

$$\mathrm{E}\left[(\bar{y}-\mu)^2 \middle| \mathbf{y},\sigma^2\right] = \mathrm{Var}\left(\mu \middle| \bar{y},\sigma^2\right) = \sigma^2/N.$$

Now the equation to solve is $\hat{\sigma}^2 = s^2 + \hat{\sigma}^2/N$ that has the same solution as in (3.90) by maximization of the logresidual likelihood. Again, this approach shows that ML does not take properly into account the effect of the incertainty in μ when replacing it by $\bar{y}$ in the loglikelihood function.

General case

For the linear Gaussian model defined by $\mathbf{y} \sim \mathcal{N}(\mathbf{X\beta},\mathbf{V})$, Patterson and Thompson (1971) proposed the following transformation:

$$\mathbf{y} = (y_i)_{1\leq i\leq N} \leftrightarrow (\mathbf{u}',\mathbf{v}')'. \qquad (3.91)$$

where $\mathbf{u} = \mathbf{Py}$ and $\mathbf{v} = \mathbf{Sy} = (\mathbf{I}_N - \mathbf{H})\mathbf{y}$ with $\mathbf{H}$ (the so called hat matrix) being the usual LS projector $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$.

By definition $\mathbf{v}$ does not depend on $\boldsymbol{\beta}$ but only on $\boldsymbol{\gamma}$ and is the basis of the residual likelihood $L_1(\boldsymbol{\gamma};\mathbf{v})$. Two comments are worthwhile at this stage.

1) One may question whether some information on $\boldsymbol{\gamma}$ is not lost by basing inference only on $L_1(\boldsymbol{\gamma};\mathbf{v})$ and ignoring $L_2(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{u})$. This issue has been somewhat controversial in the seventies and eighties. Some people argued that "there is no avalaible information on $\boldsymbol{\gamma}$ in the absence of knowledge of $\boldsymbol{\beta}$" whilst others said that "this information is inextricably mixed up with the nuisance parameters": see Kalbfleisch and Sprott (1970) and discussants;

Actually, most specialists agree that there is no loss of information by doing so "though it is difficult to give a totally satisfactory justification of this claim" (McCullagh and Nelder, 1989, page 247).

2) Vector $\mathbf{v}$ has N elements with some of them linearly dependent. To get rid of this redundant information, Harville has suggested to only consider a subvector noted $\mathbf{K}'\mathbf{y}$ of $N-r_X$ LIN elements of $\mathbf{v}$ called « error contrasts ». that is $\mathbf{K}'=\mathbf{T}(\mathbf{I}_N-\mathbf{H})$ for any $(N-r_X)\text{x}N$ transformation matrix $\mathbf{T}$ having full row rank and thus verifying $\mathbf{K}'\mathbf{X}=\mathbf{0}$. A possibility consists of building $\mathbf{K}$ from the $N-r_X$ eigenvectors of $\mathbf{S}$. Let $\mathbf{A}$ be this $N\text{x}(N-r_X)$ matrix, it can be checked that this matrix meets the conditions assigned to $\mathbf{K}$ (See Exercise 3.12)

Now, we can write the residual loglikelihood as the loglikelihood $L(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y})$ of $\boldsymbol{\gamma}$ based on $\mathbf{K}'\mathbf{y}$

$$-2L(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y})=(N-r_X)\ln 2\pi+\ln|\mathbf{K}'\mathbf{V}\mathbf{K}|+\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}\,. \quad (3.92)$$

This expression considerably simplifies given the two following identities:

$$\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'=\underline{\mathbf{P}}, \quad (3.93)$$

$$|\mathbf{K}'\mathbf{V}\mathbf{K}|=|\mathbf{V}||\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}||\underline{\mathbf{X}}'\underline{\mathbf{X}}|^{-1}|\mathbf{K}'\mathbf{K}|, \quad (3.94)$$

where $\mathbf{P}=\mathbf{V}^{-1}(\mathbf{I}-\mathbf{Q})$ as previously, and $\underline{\mathbf{X}}$ corresponds to any matrix formed by r_X LIN columns of $\mathbf{X}$.

Hence, inserting (3.93) and (3.94) into (3.92) gives

$$-2\mathrm{L}(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y}) = C + \ln|\mathbf{V}| + \ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y}, \qquad (3.95)$$

where C is a constant equal literally to $(N - r_X)\ln 2\pi - \ln|\underline{\mathbf{X}}'\underline{\mathbf{X}}| + \ln|\mathbf{K}'\mathbf{K}|$.

By definition of a loglikelihood this term does not matter. Most software uses in their computation either

$$C = (N - r_X)\ln 2\pi - \ln|\underline{\mathbf{X}}'\underline{\mathbf{X}}|$$

or, simply, as SAS-Proc Mixed

$$C = (N - r_X)\ln 2\pi .$$

In any case, formula (3.95) makes it clear that the restricted loglikelihood does not depend on a specific value of $\mathbf{K}$ so that the choice of the projector is immaterial. We could have chosen as well $\tilde{\mathbf{v}} = \tilde{\mathbf{S}}\mathbf{y} = (\mathbf{I}_N - \tilde{\mathbf{H}})\mathbf{y}$ with $\tilde{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}$, $\mathbf{W}$ being any positive definite matrix of known coefficients.

As compared to the expression of the profile loglikelihood given in (3.30),

$$-2\mathrm{L}_P(\boldsymbol{\gamma};\mathbf{y}) = N\ln(2\pi) + \ln|\mathbf{V}| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y}$$

it turns out minus twice the restricted loglikelihood - noted often as $-2\mathrm{RL}(\boldsymbol{\gamma};\mathbf{y})$- adds a term $\ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|$ that automatically adjust for the sampling variance of the GLS estimator $\hat{\boldsymbol{\beta}}$ since $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}})^{-1}$.

Now by differentiating (3.95) with respect to $\boldsymbol{\gamma}$, one has

$$\frac{\partial[-2\mathrm{L}(\boldsymbol{\gamma};\mathbf{K}'\mathbf{y})]}{\partial\gamma_k} = \frac{\partial\ln|\mathbf{V}|}{\partial\gamma_k} + \frac{\partial\ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|}{\partial\gamma_k} + \mathbf{y}'\frac{\partial\underline{\mathbf{P}}}{\partial\gamma_k}\mathbf{y} \qquad (3.96)$$

After making simple algebraic manipulation

$$\frac{\partial\ln|\mathbf{V}|}{\partial\gamma_k} + \frac{\partial\ln|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}|}{\partial\gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\right).$$

Similarly, it can be shown (see appendix 3.6)

$$\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}},$$

Then, (3.96) becomes

$$\frac{\partial \left[-2\mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}) \right]}{\partial \gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \right) - \mathbf{y}' \underline{\mathbf{P}} \frac{\partial \mathbf{V}}{\partial \gamma_k} \underline{\mathbf{P}} \mathbf{y} . \qquad (3.97)$$

If $\mathbf{V}$ has a linear structure, $\mathbf{V} = \sum_{l=0}^{K} \mathbf{V}_l \gamma_l$ with $\partial \mathbf{V} / \partial \gamma_k = \mathbf{V}_k$, and knowing that $\underline{\mathbf{P}} \mathbf{V} \underline{\mathbf{P}} = \underline{\mathbf{P}}$, then $\mathrm{tr}\left(\underline{\mathbf{P}} \mathbf{V}_k \right) = \sum_{l=0}^{K} \mathrm{tr}\left(\underline{\mathbf{P}} \mathbf{V}_k \underline{\mathbf{P}} \mathbf{V}_l \right) \gamma_l$ and the system of REML equations obtained by setting (3.97) to zero, can be written as

$$\sum_{l=0}^{K} \mathrm{tr}\left(\hat{\underline{\mathbf{P}}} \mathbf{V}_k \hat{\underline{\mathbf{P}}} \mathbf{V}_l \right) \hat{\gamma}_l = \mathbf{y}' \hat{\underline{\mathbf{P}}} \mathbf{V}_k \hat{\underline{\mathbf{P}}} \mathbf{y} . \qquad (3.98)$$

Letting

$$\tilde{\mathbf{F}} = \left({}_m \tilde{f}_{kl} = \mathrm{tr}\left(\underline{\mathbf{P}} \mathbf{V}_k \underline{\mathbf{P}} \mathbf{V}_l \right) \right) \qquad (3.99)$$

$$\mathbf{g} = \left({}_c g_k = \mathbf{y}' \underline{\mathbf{P}} \mathbf{V}_k \underline{\mathbf{P}} \mathbf{y} \right). \qquad (3.100)$$

The non linear system (3.98) can be solved iteratively as a linear system at each iteration using

$$\tilde{\mathbf{F}}\left(\boldsymbol{\gamma}^{[n]} \right) \boldsymbol{\gamma}^{[n+1]} = \mathbf{g}\left(\boldsymbol{\gamma}^{[n]} \right). \qquad (3.101)$$

Similar comments as those about ML can be made here about REML on how to simplify calculations required for elements of $\tilde{\mathbf{F}}$ and $\mathbf{g}$.

By comparing ML equations in (3.27) with those for REML (3.98), everything goes off as if $\underline{\mathbf{P}}$ is substituted to $\mathbf{V}^{-1}$ in the coefficient matrix. But, this substitution matters a lot since the expectation of the score of the restricted likelihood $\partial \mathrm{L}(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}) / \partial \gamma_k$ is zero as expected from a true likelihood whereas that pertaining to the profile likelihood $\partial \mathrm{L}_P(\boldsymbol{\gamma}; \mathbf{y}) / \partial \gamma_k$ cannot be nil. This is another distinctive characteristic of REML vs ML that can explained why REML is less biased than ML.

As far as precision is concerned, one will proceed as with ML by using the Hessian matrix of the restricted loglikelihood or the Fisher information matrix which are (see appendix 3.6)

$$-\frac{\partial^2 \mathrm{L}}{\partial\gamma_k\partial\gamma_l}=\frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial^2\mathbf{V}}{\partial\gamma_k\partial\gamma_l}\right)-\frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\right) \\ -\frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\left(\frac{\partial^2\mathbf{V}}{\partial\gamma_k\partial\gamma_l}-2\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\right)\underline{\mathbf{P}}\mathbf{y} \qquad (3.102)$$

$$\mathrm{E}\left(-\frac{\partial^2 \mathrm{L}}{\partial\gamma_k\partial\gamma_l}\right)=\frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\right). \qquad (3.103)$$

The complementarity between formulae (3.102) and (3.103) prompted Gilmour et al. (1995) to propose for linear mixed models a second order algorithm called AI-REML based on the "**A**verage" of these two "**I**nformation" matrices

$$AI_{kl}=\frac{1}{2}\mathbf{y}'\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_k}\underline{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\gamma_l}\underline{\mathbf{P}}\mathbf{y}. \qquad (3.104)$$

Example 3.7 *REML equations for a single random factor model.(Example 31 continued)*

Let us consider the same model as in example 3.1. We can make explicit the elements of the system $\tilde{\mathbf{F}}\boldsymbol{\gamma}=\mathbf{g}$ as follows. The two elements of $\mathbf{g}=(g_0,g_1)'$ are the same quadratic forms as with ML: $g_0=\mathbf{y}'\underline{\mathbf{P}}^2\mathbf{y}$ and $g_1=\mathbf{y}'\underline{\mathbf{P}}\mathbf{Z}\mathbf{Z}'\underline{\mathbf{P}}\mathbf{y}$, but the coefficients of $\tilde{\mathbf{F}}$ have changed; they are

$$\tilde{f}_{00}=\mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{V}_0\underline{\mathbf{P}}\mathbf{V}_0\right)=\mathrm{tr}\left(\underline{\mathbf{P}}^2\right),$$

$$\tilde{f}_{01}=\mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{V}_0\underline{\mathbf{P}}\mathbf{Z}\mathbf{Z}'\right)=\mathrm{tr}\left(\mathbf{Z}'\underline{\mathbf{P}}^2\mathbf{Z}\right),$$

$$\tilde{f}_{11}=\mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{Z}\mathbf{Z}'\underline{\mathbf{P}}\mathbf{Z}\mathbf{Z}'\right)=\mathrm{tr}\left[\left(\mathbf{Z}'\underline{\mathbf{P}}\mathbf{Z}\right)^2\right].$$

The system to be solved iteratively to get the ML estimations of σ_0^2 and σ_1^2 is

$$\begin{pmatrix}\tilde{f}_{00}^{(n)} & \tilde{f}_{01}^{(n)}\\ \tilde{f}_{01}^{(n)} & \tilde{f}_{11}^{(n)}\end{pmatrix}\begin{pmatrix}\sigma_0^{2(n+1)}\\ \sigma_1^{2(n+1)}\end{pmatrix}=\begin{pmatrix}g_0^{(n)}\\ g_1^{(n)}\end{pmatrix}$$

starting from initial values $\sigma_0^{2(0)}$ and $\sigma_1^{2(0)}$ which can be taken as guessed values or estimations of a quadratic method.

This procedure can be illustrated by the same numerical application as in Example 3.1 pertaining to a two way crossclassified design with factor A as fixed and B as random according to the model

$$y_{ijk} = \mu + a_i + b_j + e_{ijk},$$

where a_i is the fixed effect of level i, $b_j \sim_{iid} \left(0, \sigma_1^2\right)$ and $e_{ijk} \sim_{iid} \left(0, \sigma_0^2\right)$.

The algorithm is initiated from starting values of the parameters. For instance with $\sigma_0^2 = 10$ $\sigma_1^2 = 5$, one obtains the iterative scheme

#	f_{00}	f_{01}	f_{11}	g_0	g_1	σ_0^2	σ_1^2
1	0.201643	0.007692	0.042659	1.611703	0.143796	7.9187	1.9430
2	0.325138	0.030465	0.178624	2.629509	0.614693	7.8910	2.0954
3	0.326921	0.028021	0.163208	2.639224	0.560306	7.8949	2.0776
4	0.326655	0.028288	0.164889	2.637600	0.566231	7.8945	2.0796
5	0.326685	0.028257	0.164696	2.637778	0.566231	7.8945	2.0794

As shown on this table displaying the first five iterations, the algorithm converges very rapidly to the same final solutions, and this holds whatever are the starting values chosen. According to (3.103), $2\tilde{\mathbf{F}}^{-1}$ also provides an estimate of the asymptotic sampling variance covariance matrix of the REML estimators that is, for $\hat{\sigma}_0^2 = 7.8945$ and $\hat{\sigma}_1^2 = 2.0794$

$$2\tilde{\mathbf{F}}^{-1} = \begin{pmatrix} 6.2144 & -1.0662 \\ -1.0662 & 12.3250 \end{pmatrix}.$$

These results can be checked using some standard software. For instance, SAS-Proc Mixed on this data set gives $\hat{\sigma}_0^2 = 7.8945$ and $\hat{\sigma}_1^2 = 2.0794$. The difference between the ML and REML estimations of σ_1^2 (0.7306 vs 2.0794 respectively) is striking while the residual variance σ_0^2 does not change very much (7.6546 vs

7.8945 respectively). From these figures, one may suspect a biased estimation of ML in the estimation of σ_1^2 as corroborated by values obtained with other procedures as Henderson III (see Exercise 3.13).

3.3.2 Bayesian interpretation

For Bayesians, the natural way to treat nuisance parameters is by integration thus resulting in what is called a marginal or integrated likelihood (Berger et al, 1999). In our case, the parameter of interest is represented by $\boldsymbol{\gamma}$ and that of nuisance by $\boldsymbol{\beta}$ so that a general integrated likelihood is defined as

$$\begin{aligned} f^I(\mathbf{y}|\boldsymbol{\gamma}) &= \int f(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\beta} \\ &= \int f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\beta} \end{aligned}, \tag{3.105}$$

where $\mathrm{d}\boldsymbol{\beta}$ is the symbol standind for $\mathrm{d}\beta_1\,\mathrm{d}\beta_2 ... \mathrm{d}\beta_p$.

Following Harville (1974), we are considering the most commonly used default conditional prior that is the uniform $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = 1$. Then the uniform-integrated likelihood is defined by

$$f^U(\mathbf{y}|\boldsymbol{\gamma}) = \int f(\mathbf{y},|\boldsymbol{\beta},\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\beta}\,. \tag{3.106}$$

Doing that, we can establish equivalence between the residual and uniform-integrated likelihood as shown below.

For the model $\mathbf{y} \sim \mathcal{N}(\underline{\mathbf{X}}\boldsymbol{\beta}, \mathbf{V})$, the density $f(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma})$ is

$$f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}) = (2\pi)^{-N/2}\left|\mathbf{V}\right|^{-1/2}\exp\left[-(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})/2\right].$$

Now, in the same way as we decomposed $\sum_{i=1}^{N}(y_i-\mu)^2$ into

$$\sum_{i=1}^{N}(y_i-\mu)^2 = \sum_{i=1}^{N}(y_i-\bar{y})^2 + N(\bar{y}-\mu)^2,$$

we can write

$$\begin{aligned} (\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\boldsymbol{\beta}) &= (\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}) + \\ &\quad (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \end{aligned}, \tag{3.107}$$

where $\hat{\boldsymbol{\beta}}$ is the GLS estimator of $\boldsymbol{\beta}$.

The first term on the right-hand side does not depend on $\boldsymbol{\beta}$ and as a constant can be factorized out. Hence,

$$f^U(\mathbf{y}|\boldsymbol{\gamma}) = (2\pi)^{-N/2}|\mathbf{V}|^{-1/2}\exp\left[-\left(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)/2\right] \cdot \int\exp\left[-\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)/2\right]\mathrm{d}\boldsymbol{\beta} \quad (3.108)$$

The expression under the sum symbol is the kernel density of $\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\gamma}$ distributed as $\mathcal{N}\left[\hat{\boldsymbol{\beta}},\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-1}\right]$ implying by definition that

$$(2\pi)^{-r_X/2}\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|^{1/2}\int\exp\left[-\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)/2\right]\mathrm{d}\boldsymbol{\beta}=1.$$

Consequently, the integral on the right side in (3.108) is equal to $(2\pi)^{r_X/2}\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|^{-1/2}$ so that

$$f^U(\mathbf{y}|\boldsymbol{\gamma}) = (2\pi)^{-(N-p)/2}|\mathbf{V}|^{-1/2}\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|^{-1/2} \cdot \exp\left[-\left(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}\right)/2\right]. \quad (3.109)$$

and minus twice the log of this density exactly gives the same expression as in (3.95) with a constant $C=\left(N-r_X\right)\log 2\pi$.

This provides another interpretation of REML as the maximum uniform-integrated likelihood or in short maximum marginal likelihood (MML) (Harville, 1974, 1977). If, in addition, one assumes a flat prior on $\boldsymbol{\gamma}$ v.i.z $\pi(\boldsymbol{\gamma})=1$, then REML turns out to be the mode of the posterior distribution of $\boldsymbol{\gamma}$ (MAP) since $\pi(\boldsymbol{\gamma}|\mathbf{y}) \propto f^U(\mathbf{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})$.

$$\hat{\boldsymbol{\gamma}}_{REML} = \mathrm{argmax}_{\boldsymbol{\gamma}\in\Gamma}\log\pi(\boldsymbol{\gamma}|\mathbf{y}). \quad (3.110)$$

3.3.3 Numerical aspects

Henderson-type and Harville's algorithm

Without entering in details, it can be shown that the Henderson algorithm for computing the ML estimations of variance components can be readily extended to REML estimation. The formulae are as follows:

$$\sigma_k^{2[n+1]} = \left[\hat{\mathbf{u}}_k^{[n]'} \hat{\mathbf{u}}_k^{[n]} + \mathrm{tr}\left(\mathbf{C}_{kk}^{[n]} \right) \sigma_0^{2[n]} \right] / q_k \qquad (3.111)$$

$$\sigma_0^{2[n+1]} = \left(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}^{[n]'} \mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}^{[n]'} \mathbf{Z}'\mathbf{y} \right) / \left(N - r_X \right) \qquad (3.112)$$

And, regarding Harville's variant,

$$\sigma_k^{2[n+1]} = \hat{\mathbf{u}}_k^{[n]'} \hat{\mathbf{u}}_k^{[n]} / \left[q_k - \mathrm{tr}\left(\mathbf{C}_{kk}^{[n]} \right) / \eta_k^{[n]} \right]. \qquad (3.113)$$

Here as previously $\hat{\boldsymbol{\beta}}^{[n]}$, $\hat{\mathbf{u}}^{[n]}$ are solutions to the HMME using $\sigma_0^{2[n]}$ and $\sigma_k^{2[n]}$ in the coefficient matrix. But $\mathbf{C}_{kl}$ now represents the block corresponding to vector $\mathbf{u}_k$ in the inverse $\mathbf{C}$ of the HMME coefficient matrix defined as $\mathbf{C} = \begin{bmatrix} \underline{\mathbf{X}}'\underline{\mathbf{X}} & \underline{\mathbf{X}}'\mathbf{Z} \\ \mathbf{Z}'\underline{\mathbf{X}} & \mathbf{Z}'\mathbf{Z} + \sigma_0^2 \mathbf{G}^{-1} \end{bmatrix}^{-1}$. Apart from this specificity of $\mathbf{C}$, formulae for variance-covariance components are unchanged. Regarding the residual variance, we have to divide by $N - r_X$ instead of N.

These algorithms as their ML analogs share the nice property of providing non negative values for variance components if starting values are strictly positive.

Again, (3.111) can be generalized to correlated random vectors of the same size using

$$\sigma_{kl}^{[n+1]} = \left[\hat{\mathbf{u}}_k^{[n]'} \hat{\mathbf{u}}_l^{[n]} + \mathrm{tr}\left(\mathbf{C}_{kl}^{[n]} \right) \sigma_0^{2[n]} \right] / q \,. \qquad (3.114)$$

Calculation of –2RL

Let us get back to the expression of the residual loglikelihood (RL) in (3.95)

$$-2\,\mathrm{RL} = \left(N - r_X \right) \ln 2\pi + \ln \left| \mathbf{V} \right| + \ln \left| \underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}} \right| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y}$$

We already proved that:

$$\mathbf{y}'\underline{\mathbf{P}}\mathbf{y} = \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y}\,, \qquad (3.115)$$

where $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}', \hat{\mathbf{u}}'\right)'$ is a the solution to HMME $\left(\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right)\hat{\boldsymbol{\theta}} = \mathbf{T}'\mathbf{R}^{-1}\mathbf{y}$ where $\mathbf{T} = (\underline{\mathbf{X}}, \mathbf{Z})$ and $\boldsymbol{\Sigma}^{-} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{pmatrix}$.

Using well-known results on determinants of partititioned matrices, we find

$$\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right| = \left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|. \qquad (3.116)$$

As shown previously (3.48)

$$\left|\mathbf{V}\right| = \left|\mathbf{R}\right|\left|\mathbf{G}\right|\left|\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right|.$$

Then

$$\left|\mathbf{V}\right|\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| = \left|\mathbf{R}\right|\left|\mathbf{G}\right|\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right|. \qquad (3.117)$$

Substituting (3.115) and (3.117) into (3.95) yields the following result valuable for any Gaussian linear mixed model described as $\mathbf{y} \sim \mathcal{N}(\underline{\mathbf{X}}\boldsymbol{\beta}, \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}')$

$$\begin{aligned} -2\,\mathrm{RL} = \left(N - r_X\right)\ln 2\pi + \ln\left|\mathbf{R}\right| + \ln\left|\mathbf{G}\right| + \ln\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right| \\ + \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{T}'\mathbf{R}^{-1}\mathbf{y} \end{aligned}. \qquad (3.118)$$

As with the profile loglikelihood, this formula greatly simplifies the calculation of the maximum of the restricted loglikelihood function, in particular by resorting to HMME and their inputs-outputs. We just have to plug in REML estimations of $\mathbf{R}$ and $\mathbf{G}$ into (3.118) to get

$$-2\,\mathrm{RL}_m = -2\,\mathrm{RL}(\mathbf{G} = \hat{\mathbf{G}}_{REML}, \mathbf{R} = \hat{\mathbf{R}}_{REML}).$$

This formula can be simplified in many instances using the particular structures of $\mathbf{R}$ and $\mathbf{G}$. The only quantity raising some difficulties is $\ln\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right|$. These can be solved by resorting to a Cholesky decomposition $\mathbf{E}\mathbf{E}'$ of the coefficient matrix $\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}$ so that $\ln\left|\mathbf{T}'\mathbf{R}^{-1}\mathbf{T} + \boldsymbol{\Sigma}^{-}\right| = 2\sum_{j=1}^{\dim E} \ln \varepsilon_{jj}$ where ε_{jj} is the j^{th} diagonal element of $\mathbf{E}$.

3.3.4. Residual likelihood and testing procedures

Fixed effects via likelihood ratio

In the case of ML, testing fixed effects via the likelihood ratio statistics consists of contrasting $-2\mathrm{L}_m$ between a reduced model (R) and a complete (C) model corresponding to the null hypothesis H_0 and the union of the null and alternative H_1 respectively and of referring to the distribution of this statistic under H_0. Unfortunately, it is not possible to directly extend this procedure to the residual loglikelihood $-2\mathrm{R}\,\mathrm{L}_m$ since the reduced and complete models use different kinds of information. For instance, if one wants to test H_0 : «$\boldsymbol{\beta}_1 = \mathbf{0}$ » vs its complementary alternative H_1 : «$\boldsymbol{\beta}_1 \neq \mathbf{0}$ », R uses $\mathbf{S}_0\mathbf{y}$ such that $\mathrm{E}_R(\mathbf{y}) = \underline{\mathbf{X}}_0\boldsymbol{\beta}_0$ while C is based on $\mathbf{S}\mathbf{y}$ such that $\mathrm{E}_C(\mathbf{y}) = \underline{\mathbf{X}}\boldsymbol{\beta} = \underline{\mathbf{X}}_0\boldsymbol{\beta}_0 + \underline{\mathbf{X}}_1\boldsymbol{\beta}_1$. Therefore, directly contrasting $-2\mathrm{R}\,\mathrm{L}_m$ between these two models is meaningless as far as testing fixed effects is concerned despite some "heuristic" justifications contrary to assertion (Gurka, 2006).

To make the approach coherent, on can contrast these two models on the basis of the same set of residuals $\mathbf{S}_0\mathbf{y}$ as proposed by Welham et Thompson (1997) that is

$$-2\mathrm{L}\left(\boldsymbol{\beta}_0, \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = \left(N - p_0\right)\ln 2\pi + \ln\left|\mathbf{K}_0'\mathbf{V}\mathbf{K}_0\right| + \left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}_0\boldsymbol{\beta}_0\right)'\left[\mathrm{Var}\left(\mathbf{K}_0'\mathbf{y}\right)\right]^{-1}\left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}_0\boldsymbol{\beta}_0\right)$$

and

$$-2\mathrm{L}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = \left(N - p_0\right)\ln 2\pi + \ln\left|\mathbf{K}_0'\mathbf{V}\mathbf{K}_0\right| + \left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}\boldsymbol{\beta}\right)'\left[\mathrm{Var}\left(\mathbf{K}_0'\mathbf{y}\right)\right]^{-1}\left(\mathbf{K}_0'\mathbf{y} - \mathbf{K}_0'\mathbf{X}\boldsymbol{\beta}\right)$$

where $p_0 = \mathrm{r}(\underline{\mathbf{X}}_0)$ and $\mathbf{K}_0'\mathbf{y}$ are $N - p_0$ LIN elements of $\mathbf{S}_0\mathbf{y}$.

Since $\mathbf{K}_0'\mathbf{X}_0 = \mathbf{0}$, the first expression reduces to that of an usual restricted loglikelihood

$$-2\mathrm{L}\left(\boldsymbol{\gamma}; \mathbf{K}_0'\mathbf{y}\right) = C(\underline{\mathbf{X}}_0) + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0\right| + \left(\mathbf{y} - \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \underline{\mathbf{X}}_0\hat{\boldsymbol{\beta}}_0\right). \qquad (3.119)$$

where $C(.)$ is a function of the design matrix for fixed effects as defined in (3.95).

Let $\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}),\boldsymbol{\gamma};\mathbf{K}_0'\mathbf{y}\right]$ designate the profile likelihood $\mathrm{L}_P\left(\boldsymbol{\gamma};\mathbf{K}_0'\mathbf{y}\right)$ of $\mathrm{L}\left(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{K}_0'\mathbf{y}\right)$ where $\tilde{\boldsymbol{\beta}}$ is a GLS solution to $\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma})=\underline{\mathbf{X}}'\mathbf{V}^{-1}\mathbf{y}$, it can be shown (See Exercise 3.14) that the appropriate "restricted" likelihood ratio statistics for testing H_0 :"$\boldsymbol{\beta}_1=\mathbf{0}$" is

$$A=-2\max_{\gamma}\mathrm{L}\left(\boldsymbol{\gamma};\mathbf{K}_0'\mathbf{y}\right)+2\max_{\gamma}\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}),\boldsymbol{\gamma};\mathbf{K}_0'\mathbf{y}\right]. \qquad (3.120)$$

In this formula, $\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}),\boldsymbol{\gamma};\mathbf{K}_0'\mathbf{y}\right]$ can be expressed as

$$\begin{aligned}-2\mathrm{L}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}),\boldsymbol{\gamma};\mathbf{K}_0'\mathbf{y}\right]&=C(\underline{\mathbf{X}}_0)+\ln\left|\mathbf{V}\right|+\ln\left|\underline{\mathbf{X}}_0'\mathbf{V}^{-1}\underline{\mathbf{X}}_0\right|\\&\quad+\left(\mathbf{y}-\underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right)'\mathbf{V}^{-1}\left(\mathbf{y}-\underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right)\end{aligned}. \qquad (3.121)$$

Fixed effects via Wald's statistic

When $\mathbf{V}$ depends of unknown parameters $\boldsymbol{\gamma}$, the sampling variance of $\hat{\boldsymbol{\beta}}$ is computed as the inverse of the Fisher information matrix evaluated at $\boldsymbol{\gamma}$ being equal to its (RE)ML estimate. This procedure does not take into account the uncertainty due to estimating $\boldsymbol{\gamma}$ so that the precision of $\hat{\boldsymbol{\beta}}$ is overestimated (SE underestimated). Therefore, the properties of Wald's test are affected for small samples. Because sampling variances are underestimated, the corresponding Wald statistics are overestimated and the corresponding test turns out to be too liberal (ie. rejects H_0 too often). That is why Kenward and Roger (1997) (later on referred as KR) proposed some adjustment on how to compute the precision and to construct the test statistic.

KR considered a REML-based GLS estimator

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}})=\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\mathbf{X}'\left[\mathbf{V}(\hat{\boldsymbol{\gamma}})\right]^{-1}\mathbf{y} \qquad (3.122)$$

where $\hat{\boldsymbol{\gamma}}$ stands in short for $\hat{\boldsymbol{\gamma}}_{REML}$, and

$$\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})=\left(\mathbf{X}'\left[\mathbf{V}(\hat{\boldsymbol{\gamma}})\right]^{-1}\mathbf{X}\right)^{-1}. \qquad (3.123)$$

The true sampling variance of (3.122) can be decomposed into

$$\mathrm{var}\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}})\right] = \mathrm{var}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right] + \mathrm{E}\left(\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right]\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right]'\right) \quad (3.124)$$

This formula clearly highlights the deficiency of the usual estimator (3.123) since the last term is ignored and $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ differs from the first term of (3.124) $\mathrm{var}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right] = \boldsymbol{\Phi}(\boldsymbol{\gamma})$, the difference $\boldsymbol{\Phi}(\boldsymbol{\gamma}) - \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ being a definitive positive matrix. . In short

$$\mathrm{var}\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}})\right] > \mathrm{var}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right] > \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}).$$

On the basis of (3.124) written as $\boldsymbol{\Phi}_A(\boldsymbol{\gamma}) = \boldsymbol{\Phi}(\boldsymbol{\gamma}) + \boldsymbol{\Lambda}(\boldsymbol{\gamma})$, KR build an estimator $\hat{\boldsymbol{\Phi}}_A$ of $\boldsymbol{\Phi}_A(\boldsymbol{\gamma})$ based on the usual estimator $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ and an estimator $\hat{\boldsymbol{\Lambda}}$ of the correction term $\boldsymbol{\Lambda}$. In particular, since $\mathbf{E}\left[\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\right] \neq \boldsymbol{\Phi}(\boldsymbol{\gamma})$, one has to correct for the bias $\mathbf{B} = \mathrm{E}\left[\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})\right] - \boldsymbol{\Phi}(\boldsymbol{\gamma})$.

Following Kackar and Harville (1984), one can evaluate this bias via a second order Taylor expansion of $\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}})$ about $\boldsymbol{\Phi}(\boldsymbol{\gamma})$:

$$\begin{aligned}\boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) \approx \boldsymbol{\Phi}(\boldsymbol{\gamma}) &+ \sum_{k=1}^{K}(\hat{\gamma}_k - \gamma_k)\frac{\partial \boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial \gamma_k} \\ &+ \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}(\hat{\gamma}_k - \gamma_k)(\hat{\gamma}_l - \gamma_l)\frac{\partial^2 \boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l}\end{aligned} \quad (3.125)$$

resulting in

$$\mathbf{B} = \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl}\frac{\partial^2 \boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l}, \quad (3.126)$$

where W_{kl} is the kl element of $\mathbf{W} = \mathrm{Var}(\hat{\boldsymbol{\gamma}})$

$$\frac{\partial^2 \boldsymbol{\Phi}(\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} = \boldsymbol{\Phi}(\mathbf{P}_k \boldsymbol{\Phi} \mathbf{P}_l + \mathbf{P}_l \boldsymbol{\Phi} \mathbf{P}_k - \mathbf{Q}_{kl} - \mathbf{Q}_{lk} + \mathbf{R}_{kl})\boldsymbol{\Phi} \quad (3.127)$$

with $\mathbf{P}_k = \mathbf{X}'\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k}\mathbf{X}$, $\mathbf{Q}_{kl} = \mathbf{X}'\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k}\mathbf{V}\frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k}\mathbf{X}$, $\mathbf{R}_{kl} = \mathbf{X}'\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}(\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l}\mathbf{V}^{-1}\mathbf{X}$.

One can proceed along the same lines as far as $\boldsymbol{\Lambda}$ is concerned with a 1st order Taylor expansion of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}})$ about $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$ viz.

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\gamma}}) \approx \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}) + \sum_{k=1}^{K} (\hat{\gamma}_k - \gamma_k) \partial \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}) / \partial \gamma_k .$$

Now

$$\frac{\partial \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})}{\partial \gamma_k} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-} \mathbf{X}' \frac{\partial \mathbf{V}^{-1}}{\partial \gamma_k} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \quad \mathrm{Var}\left[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})\right] = \mathbf{V} - \mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-} \mathbf{X}',$$

yielding as in Kackar and Harville (1984)

$$\boldsymbol{\Lambda} \approx \boldsymbol{\Phi}\left[\sum_{k=1}^{K}\sum_{l=1}^{K} W_{kl} \left(\mathbf{Q}_{kl} - \mathbf{P}_k \boldsymbol{\Phi} \mathbf{P}_l\right)\right]\boldsymbol{\Phi} \qquad (3.128)$$

For a linear mixed model structure $\mathbf{V} = \sum_{k=0}^{K} \mathbf{V}_k \gamma_k$, the elements $\mathbf{R}_{kl} = 0$ yielding $\mathbf{B} = -\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}(\boldsymbol{\gamma}) = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) + \boldsymbol{\Lambda}$ and finally, due to $\boldsymbol{\Phi}_A = \boldsymbol{\Phi}(\boldsymbol{\gamma}) + \boldsymbol{\Lambda}$

$$\hat{\boldsymbol{\Phi}}_A = \boldsymbol{\Phi}(\hat{\boldsymbol{\gamma}}) + 2\hat{\boldsymbol{\Lambda}} . \qquad (3.129)$$

Remember that $\mathbf{W}$ can be consistently approximated by the inverse $\mathbf{J}^{-1}(\boldsymbol{\gamma})$ of the Fisher information matrix $\mathbf{J}(\boldsymbol{\gamma}) = 1/2\left({}_m \mathrm{tr}\left(\underline{\mathbf{P}}\mathbf{V}_k \underline{\mathbf{P}}\mathbf{V}_l\right)\right)_{kl}$, but one can also uses the observed information or the average information matrix (see formulae 3.102 and 3.104).

Coming back to the question of testing $\mathrm{H}_0 : \mathbf{C}'\boldsymbol{\beta} = \mathbf{0}$ against its contrary alternative H_1 with $\mathbf{C}'$ being a $(r \times p)$ full row rank matrix, KR propose to build a statistic under the following form

$$F^* = \lambda F \qquad (3.130)$$

where λ is a positive scale factor lower than 1, and F is the statistic $F = \hat{W} / r$ based on Wald's pivot $\hat{W} = \hat{\boldsymbol{\beta}}'\mathbf{C}\left(\mathbf{C}'\hat{\boldsymbol{\Phi}}_A\mathbf{C}\right)^{-1}\mathbf{C}'\hat{\boldsymbol{\beta}}$ adjusted for the precision of $\hat{\boldsymbol{\beta}}$ as explained previously.

KR provided formulae for computing m and λ (see page 987). The values of m and λ are calibrated by matching the first two moments of both sides of (3.130) such that F^* has an approximative $\mathrm{F}(r, m)$ distribution under the null

hypothesis. In addition, they impose that this distribution turns out to be the true one when $\hat{W}$ reduces to the Hotelling T^2, or in other ANOVA F-ratios situations.

A typical situation coming under Hotelling's statistic arises when testing $\mathrm{H}_0 : \mathbf{C}'\boldsymbol{\mu} = \mathbf{0}$ under the Gaussian model $\mathbf{y}_i \sim_{iid} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; $i = 1, 2, ..., N$ (Rao, 1973). Then

$$T^2 = \min_{H_0} (\bar{\mathbf{y}} - \boldsymbol{\mu})' (\mathbf{S} / N)^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}), \qquad (3.131)$$

where $\bar{\mathbf{y}} = \left(\sum_{i=1}^{N} \mathbf{y}_i \right) / N$ and $\mathbf{S} = (N-1)^{-1} \sum_{i=1}^{N} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ are the usual moment estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Then $F^* = \lambda T^2 / r$ with $\lambda = (N - r) / (N - 1)$ and under the null hypothesis F^* has a $\mathrm{F}(r, N - r)$ distribution. It can be shown that $T^2 = \hat{W} = \hat{\boldsymbol{\mu}}' \mathbf{C} \left[\mathbf{C}' \hat{\mathrm{V}}(\hat{\boldsymbol{\mu}}) \mathbf{C} \right]^{-1} \mathbf{C}'$ where $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$ and $\hat{\mathrm{V}}(\hat{\boldsymbol{\mu}}) = \mathbf{S} / N$.

In any case, the proposed adjustement improves the usual asymptotic approximation of precision of estimations of fixed effects and a better adequation between the nominal and actual alpha- levels than with the Wald Chi-square asymptotic test and the non-adjusted F-type procedures for small samples. Brown and Prescott (2006) mentioned a good behaviour down to 5 subjects per treatment group.

From a practical point of view, this procedure is now implemented in SAS-Proc Mixed using the option DDFM=KENWARDROGER within the MODEL statement.

Example 3.8 *Testing period by time interaction on repeated data*

The small data set chosen to illustrate the KR adjustement and Hotteling's testing procedure is part of a planified experiment on skin biophysical parameters (here skin capacitance) recorded on 8 Caucasian young and non pregnant women at 6 occasions: two successive 24h periods with three measurements every 8 hours in each period (Latreille et al., 2006). One of the

objectives of such studies is to test the Circadian rythmicity i.e. whether the skin properties show the same pattern from one period to the other.

Table 3.6: Records on 8 subjects measured at 6 occasions (period by hour combination)

Subject	11	12	13	21	22	23
1	840	855	825	795	855	790
2	785	760	710	770	735	695
3	650	650	685	565	580	535
4	695	685	740	665	675	670
5	800	710	675	660	620	670
6	730	745	710	765	735	700
7	720	615	595	470	570	410
8	805	755	735	690	675	665

The model can be written using the same notations as in (3.131) $\mathbf{y}_i = \left({}_c y_{it} \right)_{1 \le t \le 6}$ with $\boldsymbol{\mu} = \left({}_c \mu_t \right)_{1 \le t \le 6}$ and $\boldsymbol{\Sigma} = \left({}_m \sigma_{st} \right)_{1 \le s \le 6, 1 \le t \le 6}$. Let $\mu_t = \mu_{jk}$ where j stands for period and k for hour within period, testing the period x hour interaction reduces to $\mathrm{H}_0 : \mathbf{C}'\boldsymbol{\mu} = \mathbf{0}$ with

$$\mathbf{C}' = \begin{pmatrix} 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}.$$

From the data we can compute $\bar{\mathbf{y}}$ and $\mathbf{S}$. Here, we have

$$\bar{\mathbf{y}}' = \begin{pmatrix} 753.125 & 721.875 & 709.375 & 672.500 & 680.625 & 641.875 \end{pmatrix};$$

The diagonal elements of $\mathbf{S}$ are :

$$\begin{pmatrix} 4156.6964 & 5556.6964 & 4260.2678 & 12307.1428 & 8917.4107 & 13635.2679 \end{pmatrix}$$

The off diagonal elements , under a correlation matrix form, are :

1	0.7878	0.4851	0.6024	0.6302	0.6225
	1	0.8671	0.9076	0.9404	0.9018
		1	0.8032	0.8600	0.9774
			1	0.8947	0.9593
				1	0.8588
					1

Then, we can compute T^2 from $T^2 = N\overline{\mathbf{y}}'\mathbf{C}\left[\mathbf{C}'\mathbf{S}\mathbf{C}\right]^{-1}\mathbf{C}'\overline{\mathbf{y}}$. Here $N=8$ and $r=2$ so that

$$\mathbf{C}'\overline{\mathbf{y}} = \begin{pmatrix} 13.125 \\ -26.250 \end{pmatrix},\ \mathbf{C}'\mathbf{S}\mathbf{C}/N = \begin{pmatrix} 554.4084821 & 229.1294643 \\ 229.1294643 & 584.5982143 \end{pmatrix} \text{ and}$$

$T^2/r = 1.1793$ which is equivalent to the non adjusted F statistic $\hat{W}/r$. By definition, the scale parameter is equal to $\lambda = (N-r)/(N-1)$ that is here $\lambda = 6/7$ so that $F^* = \lambda F = 1.0108$. This has to be compared to a Fisher distribution $\mathrm{F}(r, N-r)$ with $r=2$ and $N-r=6$ degrees of fredom resulting in a P-value of 0.42.

On the other hand we can view this test in a purely mixed model context as $y_{ijk} = \mu_{jk} + \varepsilon_{ijk}$ with μ_{jk} being the fixed effect part and ε_{ijk} being the residual term. In turn μ_{jk} can be classically decomposed into $\mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}$ where μ corresponds to a mean, α_j to a period j effect, β_k to a time k effect and γ_{jk} to a period j by time k interaction effect. The resulting variance covariance of the residual terms has the so called "unstructured' form and can be written as $\mathbf{V} = \mathbf{I}_8 \otimes \mathbf{\Sigma}$ when data sorted by subject and time x period measurement within subject. Testing that $\gamma_{jk} = 0$ can be easily carried out under SAS-Proc Mixed using the following code:

```
proc mixed data=skin.facehydra ;
class individual period time;
model face=period time period*time/solution ddfm=kenwardroger chisq;
repeated/sub=individual type=UN;
run;
```

The ouputs for "Type 3 tests of fixed effects" indicate a F statistics of 1.01 and a P value of 0.4185 for a F distribution with 2 and 6 degrees of freedom indicating a perfect agreement with the Hotelling exact testing procedure as far as the statistic and its true distribution are concerned. This is especially clear as the unadjusted analysis (without the option ddfm=kr) gives a value of the F statistic

equal to 1.18 that is exactly our $\hat{W}/r$ term and refers to a F distribution with 2 and 7 degrees of freedom giving a P-value of 0.3619.

Variance covariance structures

As pointed out earlier, testing for mean and variance structures should be carried out separetely. In particular, comparison of $\mathbf{V}$ structures requires the same expectation model. Because misspecification of the mean model may pollute estimation of $\mathbf{V}$, it is usually recommended to start comparison of $\mathbf{V}$ structures with the so called maximal model i.e., the most elaborate model we can conceive on the subject-matter considerations. This can be in some special designed experiments (treatment, blocks, fixed time measurements), the saturated model. But in many instances, involving e.g., a lot and a mixture of discrete and continuous covariates, setting the maximal model can be a complicated task and there will be no objective rule to build it. Remember, we can also iterate between the mean and variance structures comparison stages starting e.g., with selecting an elaborate mean model using when possible a non-parametric estimation of $\mathbf{V}$ based on a robust type approach and then switch to comparing different $\mathbf{V}$ structures.

In any case, given a maximal for the mean, we can compare models for the variance using the REML-based likelihood ratio test based on contrasting maximized restricted loglikelihoods of the two models

$$\lambda = -2\mathrm{RL}_{R,m} + 2\mathrm{RL}_{C,m}\Big|_{\mathrm{H}_0} \underset{d}{\rightarrow} \chi^2_{\dim(C)-\dim(R)}, \qquad (3.133)$$

which, under the reduced model, has an asymptotic Chi-square distribution with degrees of freedom equal to the difference between the number of parameters determining the variance structure under the complete (C) and reduced (R) models respectively.

Two comments are worthwhile at this stage:

a) We could have used the classical likelihood as well since both being likelihood functions have the same limiting distribution for λ. However

the RL should be preferred as its actual rejection level is closer to the nominal one than for the classical likelihood.

b) As in the case of compaison of mean models, the LR applies only if the R and C models are nested viz. that R is a special case of C.

Example 3.9. Comparing variance structures for repeated data. Example 3.8 continued

For instance, in the repeated data example (3.8) we can compare the unstructured (UN) V model vs the "intra class" or "compound symmetry (CS)" V model such as $\mathbf{V}_i = \sigma^2\left[\rho\mathbf{I}_6 + (1-\rho)\mathbf{J}_6\right]$ assuming a constant variance σ^2 across time occasions and a constant correlation ρ between pairs of measurements. The likelihood ratio test is $\lambda = 469.6 - 425.1 = 44.5$. This value should be compared with a Chi-square of 13 degrees of freedom (15 - 2 for the UN and CS models respectively) yielding a P value of 10^{-7} thus indicating that the CS assumption does not adequately fit the data. For more details about the comparison of covariance structures, see e.g. Wolfinger (1993), Verbeke G. and Molenberghs G. (2000) and West et al. (2007).

Notice that in this simple case $2\mathrm{RL}_m$ can be simply computed from

$$(N-p)\log 2\pi + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \quad \text{(see 3.95)}$$

using the appropriate estimation of $\mathbf{V}$ under the two models.

Actually, the standard theory for the distribution of the likelihood ratio statistic under the null hypothesis does not always apply. There are some complications due when the null hypothesis specifies parameter values which are on the boundary of the parameter space. For instance, we may like to test whether we need a random intercept in a model or equivalently that $H_0 : \sigma_1^2 = 0$ against $H_1 : \sigma_1^2 > 0$ with zero being obviously a lower bound for the values of σ_1^2. The general theory of such testing procedures has been developed by Self and Liang (1987) and Delmas and Foulley (2007), and some of its application to mixed

models for longitudinal data presented by Stram et Lee (1994, 1995), Verbeke and Molenbergs (2003) and Zhang and Lin (2008).

To practically illustrate this issue, let us consider the simple random intercept model mentioned before $\mathbf{y} \sim \mathcal{N}\left(\mathbf{X\beta}, \mathbf{ZZ}'\sigma_1^2 + \mathbf{I}_N\sigma_0^2\right)$ with the corresponding test $H_0 : \sigma_1^2 = 0$ vs $H_1 : \sigma_1^2 > 0$. The LR statistic can be written in short as:

$$\lambda = -2RL_R + 2RL_C, \tag{3.134}$$

where $RL_R = \text{Max}_{\sigma_0^2>0} \text{RL}(\sigma_1^2 = 0, \sigma_0^2; \mathbf{y})$ and $RL_C = \text{Max}_{\sigma_1^2 \geq 0, \sigma_0^2>0} \text{RL}(\sigma_1^2, \sigma_0^2; \mathbf{y})$.

By blindly applying the result in (3.133), we infer that under H_0, λ has an asymptotic Chi-square distribution with 1 degree of freedom. But this claim is wrong. Actually, it is possible that under the complete model, the REML estimation σ_1^2 is going to be 0 so that $RL_C = RL_R$ and $\lambda = 0$. How often can this happen? Under H_0 and in asymptotic conditions, this will happen half time due to the normal asymptotic distribution of the unconstrained REML estimator about its true value, here 0. Therefore, the correct distribution we have to refer to (3.134) is a 50:50 mixture of a Dirac probability mass in 0 (sometimes noted as a chi-square with 0 degrees of freedom χ_0^2) and of a chi-square with 1 degree of freedom χ_1^2.

$$\lambda \underset{\mathcal{L}}{\rightarrow} 1/2\chi_0^2 + 1/2\chi_1^2. \tag{3.135}$$

Now, what does it imply making the right vs wrong decision? How often do we reject H_0 when it is true? It is rejected when i) λ is positive which occurs with probability $1/2$, and ii) given it is positive, when λ is higher than a given threshold s so that $1/2\Pr(\lambda \geq s) = \alpha$, where α is by definition the significance level of the test. This implies that s must be computed as

$$\Pr(\chi_1^2 \geq s) = 2\alpha, \tag{3.136}$$

i.e. $s = \chi^2_{1,1-2\alpha}$ being the $(1-2\alpha)$ quantile of the chi-square distribution with 1 degree of freedom.

Practically, since the true threshold is lower than the naïve one, this means that the appropriate LRT will reject more often the null hypothesis than the naïve one which is therefore too conservative. As far as the P-value, it should be calculated as $1/2\Pr(\chi_1^2 \geq \lambda_{obs})$ that is half the value of the naive one.

The case one vs two correlated random also deserves some attention. This corresponds to the following linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}, \qquad (3.137)$$

where $\mathbf{X}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are known $(N\text{x}p)$, $(N\text{x}q)$ and $(N\text{x}q)$ incidence matrices; $\boldsymbol{\beta}$ is the $(p\text{x}1)$ vector of fixed effect, $\mathbf{u}_1$, $\mathbf{u}_2$ are the two $(q\text{x}1)$ vectors of random effects and $\mathbf{e}$ is the $(N\text{x}1)$ vector of residuals independent of $\mathbf{u}_1$ and $\mathbf{u}_2$. It is assumed that $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)'$ has a centered Gaussian distribution with variance

$$\text{V}ar\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \begin{pmatrix}\sigma_1^2\mathbf{I}_q & \sigma_{12}\mathbf{I}_q\\ \sigma_{12}\mathbf{I}_q & \sigma_2^2\mathbf{I}_q\end{pmatrix},$$

which can be written in condensed notations as $\text{var}(\mathbf{u}) = \boldsymbol{\Sigma} \otimes \mathbf{I}_q$ where

$\boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}$. Similarly, $\mathbf{e}$ is assumed $\mathbf{e} \sim \mathcal{N}(0, \sigma_0^2\mathbf{I}_N)$.

This is a common model in longitudinal data analysis as already seen in examples (1.6) and (3.6).

We want to test $H_0 : \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & 0\\ 0 & 0\end{pmatrix}$ with $\sigma_1^2 > 0$ by construction vs

$H_1 : \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}$ with i) $\sigma_2^2 > 0$ and ii) $\sigma_1^2\sigma_2^2 - \sigma_{12}^2 \geq 0$ which is equivalent to

$\boldsymbol{\Sigma}$ being either positive-definite (ii strictly positive) or positive-semidefinite (ii positive or nil).

Now, if as previously one relies on the LR statistic $\lambda = -2RL_R + 2RL_C$, its asymptotic distribution under H_0 is no longer a chi-square with 2 (3-1) degrees of freedom but a 50:50 mixture of a chi-square with 1 degree of freedom χ_1^2. and of a chi-square with 2 degrees of freedom χ_2^2.

$$\lambda \underset{\mathcal{L}}{\rightarrow} 1/2\chi_1^2 + 1/2\chi_2^2. \qquad (3.138)$$

Again, besides the theory establishing that result, we can try to figure out how the first component χ_1^2 arises. This component resorts from a submodel having a variance covariance matrix depending on 2 parameters only. However, this cannot be $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$, since no REML estimation can yield $\hat{\sigma}_{12} = 0$, this event having a nul probability of occurrence. Actually, the only submodel that can yield such a component is the one in (3.137) having the two random effects proportional $\mathbf{u}_2 = \delta \mathbf{u}_1$, thus resulting in $\mathbf{\Sigma} = \sigma_1^2 \begin{pmatrix} 1 & \delta \\ \delta & \delta^2 \end{pmatrix}$ which is a positive-semidefinite (psd) matrix (condition $\sigma_1^2\sigma_2^2 - \sigma_{12}^2 = 0$).

The 1 vs 2 random effect case can be extended in the same way to m vs $m+1$ with $H_0 : \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{mm} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix}$ vs $H_1 : \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{mm} & \mathbf{\Sigma}_{m,m+1} \\ \mathbf{\Sigma}'_{m,m+1} & \sigma_{m+1}^2 \end{pmatrix}$. Under H_0, the LR statistic is now asymptotically distributed as $1/2\chi_m^2 + 1/2\chi_{m+1}^2$.

In practice these properties presupposed that the REML (or ML) algorithm used for estimating $\mathbf{\Sigma}$ can provide solutions which are either pd or psd matrices. This has to be checked as some software reduces computation of REML estimates to only pd $\mathbf{\Sigma}$matrices.

Other kinds of tests such as m vs $m+k$ random effects do not resort to a simple mixture of chi-squares but to complex mixtures of distributions, the specification of which is beyond the scope of this book.

Corresponding procedures based on the score test were considered by Verbeke and Molenbergs (2003) and Zhang and Lin (2008).

3.3.5. Residual likelihood and Information criteria

It may happen that models compared are not nested. For instance, in the analysis of repeated data, one might be interested in contrasting a random intercept model with a constant correlation $r_{st} = r$ among any pair of measurements y_{is} and y_{it} made on the same individual i, and a first order autoregressive model with a correlation equal to $r_{st} = \rho^{|s-t|}$. Obviously, the first model is not a special case of the second one. A simple way to handle that is to recourse to information criteria.

Historically, the first and the most popular one is the Akaike Information Criterion. This criterion is derived from a measure of the Kullback-Leibler distance $I(f, g_m)$ between the true model generating the data with distribution $f(\mathbf{y})$ and a class of candidate model with distribution $g_m(\mathbf{y} \mid \boldsymbol{\theta}_m)$

$$I(f, g_m) = E_f\left[\log f(\mathbf{y})\right] - E_f\left[\log g_m(\mathbf{y} \mid \boldsymbol{\theta}_m)\right] \qquad (3.139)$$

where $E_f(.)$ stands for an expectation taken with respect to the true distribution $f(.)$ generating the data.

When comparing two different candidate models, only the last term matters as the first one is common to both of them.

Comparing two different candidate models relies on

$$\Delta I_{m,m'} = I(f, g_m) - I(f, g_{m'}) = E_f\left[\log g_{m'}(\mathbf{y} \mid \boldsymbol{\theta}_{m'})\right] - E_f\left[\log g_m(\mathbf{y} \mid \boldsymbol{\theta}_m)\right],$$

so that only the last term matters.

The true value $\boldsymbol{\theta}_m$ is unknown but it can be estimated consistently by its MLE $\hat{\boldsymbol{\theta}}_m(\mathbf{y})$ based on the observed data $\mathbf{y}$ and one may think to replace $\boldsymbol{\theta}_m$ by its

MLE $\hat{\boldsymbol{\theta}}_m(\mathbf{y})$ and compare models on $-E_f\left[\log g_m\left(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}_m\right)\right]$. However if one model is nested into the other one $\mathcal{M}\supset\mathcal{M}'$, the difference

$$I(f,\hat{g}_m)-I(f,\hat{g}_{m'})=E_f\left[\log g_{m'}\left(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}_{m'}\right)\right]-E_f\left[\log g_{m'}\left(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}_{m'}\right)\right]$$

will always be negative or nil as $\log g_m\left(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}_m\right)\geq\log g_{m'}\left(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}_{m'}\right)$ and preference will always be given to the larger model.

This too optimistic view in favour of the larger model is due to the fact that $\log g\left(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}\right)$ (with notation ignoring model subscript) utilizes the data twice: once to get $\hat{\boldsymbol{\theta}}(\mathbf{y})$ and a second time in evaluating the fitting ability of the model by the observed loglikelihood $\log g\left(\mathbf{y}\,|\,\hat{\boldsymbol{\theta}}(\mathbf{y})\right)$.

To avoid this double use of the data, one may consider the following quantity

$$AI=\int f(y)\left\{E_{\tilde{\mathbf{y}}}\left[-2\log g\left(\tilde{\mathbf{y}}\,|\,\boldsymbol{\theta}(\mathbf{y})\right)\right]\right\}dy$$

or, in short

$$AI=E_yE_{\tilde{y}}\left[-2\log g\left(\tilde{\mathbf{y}}\,|\,\hat{\boldsymbol{\theta}}(\mathbf{y})\right)\right] \qquad (3.140)$$

where the expectations have to be taken with respect to different and independent sample spaces: (y) and ($\tilde{y}$) having the same true distribution. The multiplication factor 2 in (3.140) is introduced by convenience so that quantities are on a deviance (or likelihood ratio) scale. Then, AI can be consistently estimated by

$$-2\log g\left(\tilde{\mathbf{y}}\,|\,\hat{\boldsymbol{\theta}}(\mathbf{y})\right)+\text{Penalty} \qquad (3.141)$$

where

$$\text{Penalty}=-E_{\tilde{y}}\left[-2\log g\left(\tilde{\mathbf{y}}\,|\,\hat{\boldsymbol{\theta}}(\mathbf{y})\right)\right]+E_yE_{\tilde{y}}\left[-2\log g\left(\tilde{\mathbf{y}}\,|\,\hat{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$$

If one assumes that the observations $y_i\ \ i=1,..,N$ are iid and that $f=g(.\,|\,\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is a member of the parametric class of candidate models so that

consistency of the MLE estimator guarantees the convergence of $\hat{\boldsymbol{\theta}}(\mathbf{y})$ to $\boldsymbol{\theta}_0$. Then it has been shown that $\text{Penalty} = 2k$ and one obtains in (3.141) the Akaike Information Criterion (Akaike, 1973) for comparing models on the basis of the following value

$$AIC = -2L_{\max} + 2k\,, \tag{3.142}$$

where $L_{\max}$ is the maximised loglikelihood of the candidate model and k is the number of unknown parameters of this model.

The empirical idea behind (3.141) is to compare models on account of their fitting ability ($-2L_{\max}$) but with a penalty term $(2k)$ against models with too many parameters thus allowing to evaluate their predictive ability more fairly. This is especially clear due to the expectations taken with respect to two different data samples: (y) and ($\tilde{y}$) acting as testing and learning samples respectively.

The value k for the penalty terms relies theoretically on the assumption that the true distribution generating the data belong to the class of candidate models. However, this assumption can be relaxed leading to the so-called Tageuchi Information Criterion (TIC) (Takeuchi, 1976)

$$TIC = -2L_{\max} + 2tr\left(\mathbf{K}\mathbf{J}^{-1}\right)$$

where $\mathbf{K} = Var_f\left(\dfrac{\partial \log g\left(y\,|\,\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right)$ is the variance of the score function and $\mathbf{J} = -E_f\left(\dfrac{\partial^2 \log g\left(y\,|\,\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)$ the Fisher information matrix with expectation and variance taken with respect to the true density.

Estimation of $tr\left(\mathbf{K}\mathbf{J}^{-1}\right)$ will rely on the true density replaced by $g\left(y\,|\,\tilde{\boldsymbol{\theta}}_0\right)$ with $\tilde{\boldsymbol{\theta}}_0 = \arg\min I(f, g(.\,|\,\boldsymbol{\theta}_k))$ the best approximation in the parametric class of candidate models to the true model (sometimes called quasi true model). Notice

that if the true model is a candidate model $\mathbf{K} = \mathbf{J}Var\left(\hat{\boldsymbol{\theta}}\right)\mathbf{J}$ and $tr\left(\mathbf{K}\mathbf{J}^{-1}\right) = \dim\left(\boldsymbol{\theta}\right)$.

But the accuracy of estimation of this penalty term is generally poor then precluding its direct use in practice as compared to the simple AIC penalty. To overcome these limitations, bootstrap versions of the TIC penalty have been proposed (Shang and Cavanaugh, 2008).

If we want to compare models on their variance covariance structures $\mathbf{V}$ on a REML basis, the models must have the same fixed effects and the maximized residual loglikelihood $RL_{\max}$ should be substituted to $L_{\max}$ in (3.142) with k being now the number of parameters specifying $\mathbf{V}$. AIC is negatively biased and thus tends to favour models with high dimension. There have been several propositions to adjust for this underestimation of AIC due to small sample size. For instance, Hurviwch and Tsai (1989) proposed a sample size N^* corrected version of the AIC known as AICC defined as

$$AIC = -2L_{\max} + 2kN^* / \left(N^* - k - 1\right), \tag{3.143}$$

thus implying an additional penalty term since

$$AICC = AIC + 2k\left(k+1\right) / \left(N^* - k - 1\right),$$

which can be applied as soon as $N^* / k < 40$.

Another way to correct for overfitting is to recourse to Schwartz's information criterion usually called BIC as it refers to an approximative Bayes factor (BF) (see Exercise 3.15),

In Bayesian inference, if we want to confront two hypotheses H_0 and H_1 corresponding to the models $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively, we can rely on the ratio of their posterior probabilities $\dfrac{\Pr\left(\mathcal{M}_0 \mid \mathbf{y}\right)}{\Pr\left(\mathcal{M}_1 \mid \mathbf{y}\right)}$.

Since $$\frac{\Pr(\mathcal{M}_0 \mid \mathbf{y})}{\Pr(\mathcal{M}_1 \mid \mathbf{y})} = \frac{m(\mathbf{y} \mid \mathcal{M}_0)}{m(\mathbf{y} \mid \mathcal{M}_1)} \frac{\Pr(\mathcal{M}_0)}{\Pr(\mathcal{M}_1)} \tag{3.144}$$

we can also compare models on the ratio of the marginal distributions of the data $m(\mathbf{y} \mid \mathcal{M}) = \int f(\mathbf{y} \mid \boldsymbol{\theta}, \mathcal{M}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ under these two models (Jeffreys, 1961, page 248) known as the Bayes Factor

$$BF_{01} = \frac{m(\mathbf{y} \mid \mathcal{M}_0)}{m(\mathbf{y} \mid \mathcal{M}_1)} = \frac{\Pr(\mathcal{M}_0 \mid \mathbf{y})}{\Pr(\mathcal{M}_1 \mid \mathbf{y})} / \frac{\Pr(\mathcal{M}_0)}{\Pr(\mathcal{M}_1)} \tag{3.145}$$

BF can then be viewed as the ratio of the posterior odds to the priors odds and thus as a criterion for measuring the evidence for $\mathcal{M}_0$ against $\mathcal{M}_1$ brought by the data over the relative prior information on the models.

Computing the marginal likelihood remains a difficult task involving various techniques (see eg Kass and Raftery, 1955; Friel and Wyse, 2012). One can resort to integral approximations such as the Laplace' method which yields the basis for BIC (Le Barbier and Mary-Huard, 2006).

Expanding $l(\boldsymbol{\theta}) = g(\mathbf{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})$ around the posterior mode (MAP) $\tilde{\boldsymbol{\theta}}$ and exponentiating leads to

$$m(\mathbf{y}) = (2\pi)^{k/2} g(\mathbf{y} \mid \tilde{\boldsymbol{\theta}}) \pi(\tilde{\boldsymbol{\theta}}) \boldsymbol{\Lambda}^{1/2}(\tilde{\boldsymbol{\theta}}) (1 + o(N^{-1})) \tag{3.146}$$

where $\pi(\boldsymbol{\theta})$, $g(\mathbf{y} \mid \boldsymbol{\theta})$ represent the prior density and likelihood respectively and $\boldsymbol{\Lambda}(\boldsymbol{\theta}) = \left[-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1}$.

The next step is to replace the MAP $\tilde{\boldsymbol{\theta}}$ by the MLE $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ by the inverse of the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ leads to (on the deviance scale)

$$\begin{aligned} -2\log m(\mathbf{y}) = & -2\log\left[g(\mathbf{y} \mid \hat{\boldsymbol{\theta}}) \right] - 2\log\left[\pi(\hat{\boldsymbol{\theta}}) \right] - k\log(2\pi) \\ & + \log\left[\mathbf{I}_1(\hat{\boldsymbol{\theta}}) \right] + k\log(N) + o(N^{-1/2}) \end{aligned}$$

where $\mathbf{I}_1(\boldsymbol{\theta})=\mathbf{I}(\boldsymbol{\theta})/N$ is the Fisher information matrix for a single observation in the N sample of iid observations.

In order to get the BIC expression, we have to make further assumptions on the constant term: $-2\log\left[\pi\left(\hat{\boldsymbol{\theta}}\right)\right]-k\log(2\pi)+\log\left[\mathbf{I}_1\left(\hat{\boldsymbol{\theta}}\right)\right]$. If we assume that a priori $\boldsymbol{\theta}\sim\mathcal{N}\left[\hat{\boldsymbol{\theta}},n\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right)\right]$ where $\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right)=\mathbf{I}^{-1}\left(\hat{\boldsymbol{\theta}}\right)$ is the asymptotic variance of the MLE $\hat{\boldsymbol{\theta}}$ such that $n\mathbf{V}(\boldsymbol{\theta})=\left[\mathbf{I}_1(\boldsymbol{\theta})\right]^{-1}$, then $\pi\left(\hat{\boldsymbol{\theta}}\right)=(2\pi)^{-k/2}\left|\mathbf{I}_1\left(\hat{\boldsymbol{\theta}}\right)\right|^{1/2}$ the constant term cancels out and we obtained the well known expression for BIC.

$$BIC=-2L_{\max}+k\log(N). \qquad (3.147)$$

Notice the special form (so called "Unit Prior Information") of the prior distribution of $\boldsymbol{\theta}$ centered on the ML estimation $\hat{\boldsymbol{\theta}}$ and having a variance equal to the inverse of the expected information matrix provided by a single observation. The expression of this variance is the analog of the variance $g\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ of Zellner's g prior for $g=N$ in the case of the linear model.

The question arises on what N (say N^*) should be plugged in (3.147) and also in (3.143).

First, if data were independent, on should take $N^*=N$ the number of observations for ML and $N^*=N-r_X$ for REML respectively where r_X is the rank of $\mathbf{X}$. But, in practice they are not and the N^* should take care of that. At the limit, if there are I subjects with their measurements perfectly correlated, then $N^*=I$. This in fact what some people (Gurka, 2006) recommend both for ML and REML and what SAS Proc Mixed does for computing BIC. In the first option, the penalization is too strong whereas it is too small in the second one. Methods have been proposed to calculate an effective sample size which takes into account the real covariance or correlation structure of the data (covariance $\mathbf{V}_i$ and correlation $\mathbf{R}_i$ matrix for subject $i=1,2,..,I$). One possibility is based on

the information matrix $\sum_{i=1}^{I} \mathbf{1}_{n_i}^{'} \mathbf{V}_i^{-1} \mathbf{1}_{n_i}$ pertaining to the intercept in the corresponding GLS equations. After standardization for scale effects, one would replace N by $N_e = \sum_{i=1}^{I} \mathbf{1}_{n_i}^{'} \mathbf{R}_i^{-1} \mathbf{1}_{n_i}$ ie the sum of the elements of the inverse of the correlation matrix by subject cumulated over subjects (Jones, 2011). N_e lies between its boundary values $I \leq N_e \leq N$ which are attained for zero and perfect correlations respectively (see exercice 3.16).

At the opposite of AIC, BIC tends to select too parsimonious models due to a more stringent penalty. That is the reason why some people prefer to use AIC rather than BIC for covariance model selection (Fitzmaurice et al., 2004, page 177). On the positive side, one can notice that due to the penalty term involving $\log\left(N^*\right)$, BIC adjusts automatically for the phenomenon known as "dimishing significance of a fixed P-value" when sample size increases (Good, 1992). Finally it should be also noted that BIC is consistent with respect to the quasi true model whereas AIC is not. This means that for large sample size the candidate model selected by BIC will be the correct model with a probability tending to one. In addition, AIC and BIC do not exactly meet the same purpose. As seen in the previous derivations, the primary objective of AIC is a predictive one while BIC is devoted to evaluate a correct description of factors affecting the outcomes of the candidate models. For more details and technicalities, the reader can refer to the very comprehensive review of selection of linear mixed models published by Mueller et al. (2013).

Example 3.10. Comparing variance structures for repeated data via information criteria. Example 3.9 continued

In the repeated data example (3.8) and (3.9), we would like to compare the following variance covariance structure by its increasing order of complex

i) Intra-class (or Compound symmetry) where $\mathbf{V}_i = \sigma^2 \left[\rho \mathbf{I}_6 + \left(1-\rho\right) \mathbf{J}_6 \right]$

with $\rho = \sigma_1^2 / \left(\sigma_1^2 + \sigma_0^2\right)$, and $\sigma^2 = \sigma_1^2 + \sigma_0^2$, σ_1^2 and σ_0^2 being the between and within subject components of variance

ii) First order autoregressive (AR1) where $\mathbf{V}_i = \sigma^2 \mathbf{H}_i$ $\mathbf{H}_i = \left(h_{i,st} = \rho^{|s-t|}\right)$ with $\mathbf{H}_i = \left(h_{i,st} = \rho^{|s-t|}\right)$ since measurements are taken over constant periods of time (every 8 hrs)

iii) Same as in ii) but with heterogeneous variances at each time occasion (ARH1) where $\mathbf{V}_i = \left(\rho^{|s-t|}\sigma_s\sigma_t\right)$

iv) Unstructured where $\mathbf{V}_i = \left(\rho_{st}\sigma_s\sigma_t\right)$ as already seen in example (3.9).

Information criteria computed are AIC, AICC and BIC according to formulae (3.142), (3.143) and (3.147) respectively. N^* in AICC is taken as $N^* = N - r_X$ that is here $N^* = 48 - 6 = 42$. Two BIC values (indexed by 1 and 2) are reported according to whether $N^* = I = 8$ (the number of subjects) or $N^* = 42$ as in AICC.

Table 3.7 . Comparison of covariance structure models based on AIC, AICC and BIC for the skin data

Model	k	-2RL	AIC	AICC	BIC1	BIC2
Intraclass	2	469.6	473.6	473.9	473.7	477.0
AR1	2	471.1	475.1	475.5	475.3	478.6
ARH1	7	461.6	475.6	478.9	476.2	487.8
UN	21	425.1	467.1	513.3	468.8	503.6

This example deserves some attention as it illustrates very well how contrasted can be the results of comparison based on such information criteria. According to AIC and BIC1, the best model is the unstructured one, whereas for AICC and BIC2, this is the intraclass one. Again as expected, AIC chooses the most complex model whereas AICC and the BIC2 version select the simplest one.

3.3.6 ML vs REML

There has been a long debate since the beginning about the relative merits of ML vs REML for estimation and model selection in the presence of nuisance parameters (Lascar and King, 2001). As stated in the introduction, maximum likelihood procedures are a turning point in the history of variance components estimation between quadratic moment vs likelihood principle based procedures.

Both ML and REML are asymptotically equivalent but the latter shares common properties with previous methods. First for balanced settings, REML and ANOVA estimators are identical provided ANOVA estimations are within the parameter space.

Secondly, MINQUE when iterated (abbreviated as I-MINQUE) is based on the same equations as REML (formula 3.101) so they are equivalent provided I-MINQUE solutions are within the parameter space. In addition, it has been shown that I-MINQUE is consistent and asymptotically normal and we know that MINQUE does not require normality of data. This explains why REML shares the same robust properties for non Gaussian data although some adjustement has to be made on the information matrix (Jiang, 2007). In connection to that, it is not suprising that REML equations can also be interpreted also in terms of quasi-likelihood or modified profile likelihood equations. Finally, REML is also a first step towards Bayesian estimation of variance components. We also have seen that Wald's statistics for fixed effects can benefit from using REML estimations of variance components.

In addition, REML does not suffer some inconsistency properties observed with ML (see Exercise 3.11 about the Neyman-Scott problem) that occurs when the dimension of fixed effects increases grows as fast as the sample size (Jiang, 2007, page 40).

For all these theoretical features, REML should be preferred to ML in most applications although it might be more demanding computationally than ML for very large data sets. But, even in such cases, the choice has already been made to use it contrarily to what happened in other areas: see e.g. the interesting

example of the animal breeding community (Meyer, 1990; Mrode and Thompson, 2005).

3.4 Exercises

3.1 Consider a one way ANOVA model $y_{ij} = \mu + a_i + e_{ij}$, $i = 1,...,I, j = 1,...,n_i$ where $a_i \sim_{iid} \left(0, \sigma_a^2\right)$, $e_{ijk} \sim_{iid} \left(0, \sigma_e^2\right)$ and $a_i \perp e_{ij}$.

1) Specify the analytical expressions of $SSA = R\left(\mu, a\right) - R\left(\mu\right)$ and $SSE = \mathbf{y}'\mathbf{y} - R\left(\mu, a\right)$ as functions of y_{ij}, $y_{i0} = \sum_{j=1}^{J} y_{ij}$, $y_{00} = \sum_{i=1}^{I} y_{i0}$, and $N = \sum_{ij} n_{ij}$.

2) Give the formulae for the expectations of *SSA* and *SSE* under the true random model and for the moment estimators of σ_e^2 and σ_a^2 based on the mean squares $MSA = SSA / \left(I - 1\right)$ and $MSE = SSE / \left(N - I\right)$.

3) Compute $\hat{\sigma}_e^2$ and $\hat{\sigma}_a^2$ from the following small data set

A-levels	*n*	$\sum y$	$\sum y^2$
1	3	36	648
2	4	96	2512
3	2	12	80

Let $\mathbf{a} = \left({}_c a_i\right)_{1 \le i \le I}$ and assume now that random effects are correlated with $\mathbf{a} \sim \left(0, \sigma_a^2 \mathbf{P}\right)$ where $\mathbf{P} = \left({}_m p_{ij}\right)_{ij}$ is a known $\left(J \text{ x } J\right)$ symmetric definite-positive matrix such that $p_{ij} = 1$ if $i = j$, and $0 \le p_{ij} < 1$ for $i \ne j$. Such a situation arises for instance when levels of A are genetically related. Show how to change $E\left(MSA\right)$ to take into account this new assumption ?

4) Show that in the balanced case $n_i = n = N / I$, $\forall i$, the expression of $E(MSA)$ reduces to $E(MSA) = \sigma_e^2 + n(1-\overline{p})\sigma_a^2$ where $\overline{p}$ is the average value of the non diagonal elements of $\mathbf{P}$.

5) Infer from this how much relative bias in $B = \left(\hat{\sigma}_{a^*}^2 - \hat{\sigma}_a^2\right) / \hat{\sigma}_a^2$ results from ignoring elements of $\mathbf{P}$ in estimating σ_a^2 by $\hat{\sigma}_{a^*}^2$.

3.2 The same model as in 1) is envisioned but now we consider the following quadratic forms $SS_0 = \mathbf{y}'\mathbf{My}$ and $SS_1 = \mathbf{y}'\mathbf{MZZ}'\mathbf{My}$ where $\mathbf{M} = \mathbf{I}_N - \mathbf{J}_N / N$, $\mathbf{I}_N$ being the identity matrix of order N and $\mathbf{J}_N = \mathbf{1}_N\mathbf{1}_N'$, the $(N \text{x} N)$ matrix made of one's.

1) Specify the analytical expressions of SS_0 and SS_1 as functions of N, $\sum_{ij} y_{ij}^2$ and $y_{00} = \sum_{i=1}^{I} y_{i0}$ on the one hand and of n_i, $y_{i.} = \left(\sum_{j=1}^{J} y_{ij}\right) / n_i$, and $y_{..} = \left(\sum_{ij} y_{ij}\right) / N$ on the other hand.

2) How would you interpret SS_0 and how does SS_1 differ from SSA?

3) Derive the analytical expression of the coefficients of σ_e^2 and σ_a^2 in $E(SS_0)$ and $E(SS_1)$.

4) Check that the results obtained are the same as with those of the ANOVA decomposition in the balanced case i.e. $n_i = n$, $\forall i$.

5) More generally, let us define $SSA^* = \sum_{i=1}^{I} k_i \left(y_{i.} - y_{..}\right)^2$ where k_i is any positive weighing scalar. Show that SSA^* reduces to SSA for $k_i = 1$ and derive the coefficients σ_e^2 and σ_a^2 in $E(SSA^*)$ with the special cases of $k_i = 1, n_i, n_i^2$.

6) Apply the previous results to the data set shown in Exercise 3.1.

3.3 In the mixed linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k$ where $\mathbf{V} = \sum_{k=0}^{K} \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k^{'}$

1) Show that the sampling variance of the estimation $\hat{\sigma}_0^2 = SSE / (N - r_{XZ})$ of the residual variance σ_0^2 is $Var(\hat{\sigma}_0^2) = 2\sigma_0^4 / (N - r_{XZ})$ where $r_{XZ} = rank(\mathbf{X}, \mathbf{Z})$ and $SSE = \mathbf{y}'\mathbf{y} - R(\boldsymbol{\beta}, \mathbf{u})$.

2) Derive an unbiased estimator of $Var(\hat{\sigma}_0^2)$

3.4 In the toy example shown in the technical note, derive the sampling variances (and SE) of the H-III estimates of variance components $\hat{\sigma}_e^2$, $\hat{\sigma}_c^2$ and $\hat{\sigma}_b^2$ under the assumption that random effects are normally distributed and true values of variance components are $\sigma_e^2 = 800$, $\sigma_c^2 = 100$ and $\sigma_b^2 = 200$ respectively.

3.5 Consider the following data set on pelvic opening (cm^2) recorded on heifers according to a two way cross classification with A being "region x origin" (4 levels) and B "sire" of these heifers (6 levels).

1) Estimate variance components under a two factor (say A and B) mixed linear model with interaction, assuming A being fixed, B and AxB random.

2) Same question as in 1) with all effects being random.

3) Same question as in 1) with an additive model.

4) Same question as in 2) with both effects being random in applying four choices of quadratic forms proposed in the technical note.

5) Summarize the results and draw conclusions.

Table. Distribution of pelvic opening records according to region x origin (A) and sire (B) of heifers

A	B	n	$\sum y$
1	1	3	987.3
1	2	2	679.8
1	3	1	341.0
1	5	2	651.0
1	6	3	907.0
2	1	4	1239.0
2	2	3	969.0
2	4	3	915.0
2	5	5	1695.5
2	6	4	1199.3
3	1	2	710.0
3	3	2	628.0
3	4	1	333.3
3	5	2	714.0
3	6	2	607.8
4	1	1	346.5
4	2	2	573.0
4	3	3	888.3
4	5	2	664.3

$\mathbf{y}'\mathbf{y} = 4875548.33$

3.6 In the linear mixed model $\mathbf{y} = \mathbf{X\beta} + \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{u}_k + \mathbf{e}$ where $E(\mathbf{y}) = \mathbf{X\beta}$ and $Var(\mathbf{y}) = \mathbf{V} = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{Z}_k^{'} \sigma_k^2 + I_N \sigma_0^2$, consider as proposed by Schaeffer (1986) the following quadratic forms $\mathbf{y}'\mathbf{Q}_k\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{Z}_k\hat{\mathbf{u}}_k^*$, for $k = 1,..,K$ where

$\mathbf{M} = \mathbf{I}_N - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$ and $\hat{\mathbf{u}}_k^*$ is the solution of the mixed model equations with ratios of variances $\lambda_k^* = \sigma_0^{*2} / \sigma_k^{*2}$.

1) Show that if the λ_k^*'s are equal to the true ratios of variances $\lambda_k = \sigma_0^2 / \sigma_k^2$, these quadratic forms have expectations

$$E\left(\mathbf{y}'\mathbf{Q}_k\mathbf{y}\right) = tr\left(\mathbf{Z}_k'\mathbf{M}\mathbf{Z}_k\right)\sigma_k^2$$

2) Extend this result to the case where $\mathbf{V} = \sum_{k=1}^{K}\mathbf{Z}_k\mathbf{A}_k\mathbf{Z}_k'\sigma_k^2 + I_N\sigma_0^2$, where $\mathbf{A}_k$ is a symmetric pd matrix.

3) Apply the previous results to the data set of example 3.5.1 with A fixed, B random and assuming an additive model.

3.7 Same problem as in 3.6 with $\mathbf{y}'\mathbf{Q}_k\mathbf{y} = \tilde{\mathbf{u}}_k^*\mathbf{A}_k^{-1}\hat{\mathbf{u}}_k^*$ (VanRaden and Yung, 1988) where $\tilde{\mathbf{u}}_k^* = \mathbf{D}_k^*\mathbf{Z}_k'\mathbf{M}\mathbf{y}$ and $\mathbf{D}_k^* = \left({}_d d_{kl}^*\right)_{1\le l\le q_k}$ is a diagonal $\left(q_k \text{ x} q_k\right)$ matrix made of elements d_{kl}^* being the reciprocals of the diagonal elements of $\mathbf{Z}_k'\mathbf{M}\mathbf{Z}_k + \lambda_k^*\mathbf{A}_k^{-1}$

1) Show that if the ratios of variances used are the true ones, then $E\left(\mathbf{y}'\mathbf{Q}_k\mathbf{y}\right) = tr\left(\mathbf{D}_k\mathbf{Z}_k'\mathbf{M}\mathbf{Z}_k\right)\sigma_k^2$.

2) Apply these results to the same model and data set as in exercise 3.6

3.8. Using the same model as in Example 3.5, but now on the « boy » sample given in table 3.4, compute the ML estimations of variance components with Henderson's algorithm and check the results with your favorite software.

3.9. Use the same procedure as in Example 3.7 to test for homogeneity of slopes but now assuming that $\mathbf{G}_0 = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}$.

3.10. Same question as in Exercise 3.9, but now applying a Wald test with SE calculated as a sandwich estimator. Compare the results of the test for i) the random intercept model, ii) the random intercept plus random slope model and iii) the robust approach.

$i = 1,...,I$

3.11. Let consider I pairs of random variables $\left(y_{i1}, y_{i2}\right)_{1 \le i \le I}$ such that $y_{ij} = \mu_i + e_{ij}$, $j = 1,2$ with $E\left(y_{ij}\right) = \mu_i$ and $e_{ij} \sim_{iid} \mathcal{N}\left(0, \sigma^2\right)$ (Neyman Scott, 1948)

1) Derive the ML estimator σ^2_{ML} of σ^2 an its expectation.

2) Same question for the REML estimator σ^2_{REML} . Conclusion?

3) Consider the additional estimator $\tilde{\sigma}^2 = \left[\sum_{i=1}^{I}\left(y_{i1}^2 - y_{i1}y_{i2}\right)\right] / I$. How does it compare with the two previous ones?

3.12. The objective of this exercise is to build a vector of "error contrasts" $\mathbf{A}'\mathbf{y}$ for expressing the restricted likelihood of the linear model $\mathbf{y}_{N\,\mathrm{x}\,1} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. This problem can be solved according to the following steps;

1. Show that the eigenvalues of an idempotent $(m \mathrm{x} m)\,\mathbf{M}$ matrix having rank q have eigenvalues equal to 1 and 0 with multiplicity orders q and $m - q$ respectively.

2. Show that $\mathbf{S} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ can be expressed as $\mathbf{A}\mathbf{A}' = \mathbf{S}$ where the $\mathbf{A}$ matrix is constructed from the $N - r$ normed eigenvectors $\mathbf{U}_k$ pertaining to the non zero eigenvalues of $\mathbf{S} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ such that $\mathbf{A}_{[N\,\mathrm{x}\,(N-r)]} = \left(\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_k, ..., \mathbf{U}_{N-r}\right)$, r being the rank of $\mathbf{X}$.

3. Show that $\mathbf{A}'$ can be written as $\mathbf{A}' = \mathbf{T}\mathbf{S}$ where $\mathbf{T}$ is an $\left(N - r\right) \mathrm{x} N$ transformation having full row rank, and consequently that $\mathbf{A}'\mathbf{X} = \mathbf{0}$.

4. Illustrate these procedures in the simple case $y_i \sim_{iid} \mathcal{N}(\mu, \sigma^2)$, $i = 1,...,N$ by making $\mathbf{A}$ explicit for $N = 3$.

3.13. Compute the estimations of σ_0^2 and σ_1^2 for the random intercept model of Example 3.1 using Henderson III. How these estimations compare with the ML and REML ones?

3.14. Show that $-2\text{Max}_{\boldsymbol{\beta}} \text{L}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{K}_0' \mathbf{y}\right)$ reduces to formula (3.121)

$C(\underline{\mathbf{X}}_0) + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}_0' \mathbf{V}^{-1} \underline{\mathbf{X}}_0\right| + \left(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \underline{\mathbf{X}}\tilde{\boldsymbol{\beta}}\right)$, where $\tilde{\boldsymbol{\beta}}$ is a GLS solution to $\underline{\mathbf{X}}' \mathbf{V}^{-1} \underline{\mathbf{X}} \tilde{\boldsymbol{\beta}} = \underline{\mathbf{X}}' \mathbf{V}^{-1} \mathbf{y}$.

3.15. Let $y_i \sim_{iid} \mathcal{N}(\mu, \sigma^2)$, $i = 1,...,N$ with $\mu \sim \mathcal{N}(\mu_0, \tau^2)$, and consider the following two hypotheses $\text{H}_0 : \mu = \mu_0$ vs $\text{H}_1 : \mu \neq \mu_0$. Consider now the the Bayes factor in favour of H_0 against $\text{H}_0 \cup \text{H}_1$ defined as $B_{01} = f\left(\overline{y} \mid H_0\right) / f\left(\overline{y} \mid H_1\right)$ where $\overline{y} = \left(\sum_{i=1}^{N} y_i\right) / N$ and $f\left(\overline{y} \mid H_k\right)$ the corresponding density under H_k for $k = 0,1$.

1. Show that B_{01} can be expressed as $B_{01} = \sqrt{1 + \rho^{-2}} \exp\left[-z^2 / 2\left(1 + \rho^2\right)\right]$ where $z = \sqrt{N}\left(\overline{y} - \mu_0\right) / \sigma$ and $\rho = \sigma / \tau\sqrt{N}$.

2. For N large, show that B_{01} can be approximated by $B_{01} \approx \left(\tau\sqrt{N} / \sigma\right) \exp\left(-z^2 / 2\right)$.

3. How does this expression can be connected to the use of BIC for comparing the reduced model H_0 to the complete model $\text{H}_0 \cup \text{H}_1$? In particular, what condition is required on the prior distribution of μ, in addition to N large, to make $B_{01} \approx BIC_0 - BIC_1$?

3.16. Compare the models considered in example 3.10 for the skin data set (table 3.6) using in BIC the effective sample size $N_e = \mathbf{1}'\mathbf{R}^{-1}\mathbf{1}$ in place of N.

You will first, derive the theoretical expression of N_e for the random intercept model $y_{ij} = \mu + a_i + e_{ij}$, $i = 1,...,I$, $j = 1,...,n_i$ where $a_i \sim_{iid} \mathcal{N}\left(0, \sigma_a^2\right)$ and $e_{ij} \sim_{iid} \mathcal{N}\left(0, \sigma_e^2\right)$ and show that N_e reduces to N and I respectively when the intraclass correlation coefficient $\rho = \sigma_a^2 / \left(\sigma_a^2 + \sigma_e^2\right)$ equals 0 and 1 respectively. Same question for the AR(1) model $y_{ij} = \mu + \phi\left(y_{ij-1} - \mu\right) + e_{ij}$.

In the random intercept model, $\mathbf{R} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ where $\mathbf{I}$ the identity matrix and $\mathbf{J} = \mathbf{1}'\mathbf{1}$ a square matrix of ones. Knowing that, $\mathbf{R}^{-1} = \frac{1}{1-\rho}\left(\mathbf{I} - \frac{\rho}{1+\rho(n-1)}\mathbf{J}\right)$ with diagonal $\mathbf{R}^{mm} = \frac{1+\rho(n-2)}{(1-\rho)\left[1+\rho(n-1)\right]}$ and off diagonal elements $\mathbf{R}^{mm'} = -\frac{\rho}{(1-\rho)\left[1+\rho(n-1)\right]}$. Summing up the n diagonal and the $n(n-1)$ off-diagonals elements of $\mathbf{R}^{-1}$ gives $\frac{n}{1+\rho(n-1)}$. Now to get N_e, we have to take the sum of such terms over the different subjects $i = 1,...,I$ so that $N_e = \sum_{i=1}^{I} \frac{n_i}{1+\rho(n_i-1)}$ which lies between I the number of subjects ($\rho = 1$) and N the total number of observations ($\rho = 0$).

In the AR(1) model, the correlation matrix a the following structure $\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$ (example for $n = 4$) and its inverse $\mathbf{R}^{-1}$ is a

tridiagonal matrix with elements $\mathbf{R}^{-1}=\left(1-\rho^2\right)^{-1}\begin{pmatrix}1 & -\rho & 0 & 0\\ -\rho & 1+\rho^2 & -\rho & 0\\ 0 & -\rho & 1+\rho^2 & -\rho\\ 0 & 0 & -\rho & 1\end{pmatrix}$.

Letting $S=\left(1-\rho^2\right)\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}$, then $S=2+(n-2)\left(1+\rho^2\right)-2(n-1)\rho$. Now, replace $1+\rho^2$ by $(1-\rho)^2+2\rho$ so that

$$\begin{aligned}S&=2+(n-2)(1-\rho)^2+2\rho\left[(n-2)-(n-1)\right]\\&=(n-2)(1-\rho)^2+2(1-\rho)\\&=(1-\rho)\left[(n-2)(1-\rho)+2\right]\end{aligned}$$

And finally after dividing by $\left(1-\rho^2\right)$, one has $\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}=1+(n-1)\dfrac{1-\rho}{1+\rho}$ and summing over subjects leads to $N_e=\sum_{i=1}^{I}\left[1+(n_i-1)\dfrac{1-\rho}{1+\rho}\right]$ which lies between I and N values attained for ρ equal to 1 and 0 respectively.

Application to the skin example yields N_e equal to 9.90, 11.79, 11.43 and 19.28 for the intraclass, AR1, ARH1 and Unstructured models respectively with the corresponding BIC3 values shown in the table below. According to BIC3, the best model is now the intraclass one in agreement with AIC, AICC and BIC2. In fact, BIC3 is close to BIC1 for the intraclass and AR1 and ARH1 models due to a high intraclass correlation (0.77) and auto regressive parameters (0.83 and 0.84 for AR1 and ARH1 respectively) but remains strongly penalized for the Unstructured model due to some smaller correlations and a large number of parameters.

Table. Comparison of covariance structure models based on AIC, AICC and BIC for the skin data

Model	k	-2RL	AIC	AICC	BIC1	BIC2	BIC3
Intraclass	2	469.6	473.6	473.9	473.7	477.0	474.1
AR1	2	471.1	475.1	475.5	475.3	478.6	476.1
ARH1	7	461.6	475.6	478.9	476.2	487.8	478.7
UN	21	425.1	467.1	513.3	468.8	503.6	487.2

3.5 Technical note on the Henderson Method III

3.5.1 Estimation procedure

The Henderson method III (Henderson, 1953; Searle et al., 1992; later on referred as H-III) is applicable to any mixed linear model $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$ defined in (3.4) and (3.5) namely

$$\mathbf{y} = \mathbf{X\beta} + \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{u}_k + \mathbf{e} = \mathbf{X\beta} + \sum_{k=0}^{K} \mathbf{Z}_k \mathbf{u}_k \,, \qquad (3.5.1)$$

with $\boldsymbol{\mu} = \mathbf{X\beta}$ and $\mathbf{V} = \sum_{k=1}^{K} \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k^{'} + \sigma_0^2 \mathbf{I}_N = \sum_{k=0}^{K} \sigma_k^2 \mathbf{Z}_k \mathbf{Z}_k^{'}$.

The method is based on the following two steps. First, H-III relies on quadratic forms obtained as reductions due to fitting the model and some of its submodels as if they were purely fixed according to the fitting constant method of Yates (1934). These quadratics may be sums of squares or not depending mainly on the data structure (balanced vs unbalanced respectively) and on the kind of model adjusted. The second stage consists of equating the observed values (vector $\mathbf{q}$) to their expected values $E(\mathbf{q})$ expressed as linear functions of the unknown parameters and taken under the true random or mixed model considered.

Let $\mathbf{u}_S = ({}_r\mathbf{u}_k)$, $k \in S$, $S \subseteq \{\mathcal{L} : k = 1,...,K\}$ be a substet of vectors $\mathbf{u}_k$ out of the complete set of vectors , $\mathbf{u} = ({}_r\mathbf{u}_k)\ 1 \le k \le K$. For instance with 3 random vectors of random effects $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$, the potential subsets are the triplet itself $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$, doubles $(\mathbf{u}_1, \mathbf{u}_2)$, $(\mathbf{u}_2, \mathbf{u}_3)$, $(\mathbf{u}_1, \mathbf{u}_3)$ and singletons $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$.

As shown clearly later on, H-III utilizes the following quadratic: the total sum of squares, $\mathbf{y}'\mathbf{y}$; the reductions due to fitting the full (fixed) model $R(\boldsymbol{\beta}, \mathbf{u})$, different submodels $R(\boldsymbol{\beta}, \mathbf{u}_S)$ and the $\boldsymbol{\beta}$ model $R(\boldsymbol{\beta})$.

By definition, $E(\mathbf{y}'\mathbf{y}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + tr(\mathbf{V})$ v.i.z after expliciting $\mathbf{V}$

$$E(\mathbf{y}'\mathbf{y}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + N\sum_{k=0}^{K} \sigma_k^2 \,. \qquad (3.5.2)$$

The other reductions are special cases of $R(\boldsymbol{\beta},\mathbf{u}_S)$. Let $\boldsymbol{\theta}=(\boldsymbol{\beta}',\mathbf{u}_S')'$ be the concatenation of vector $\boldsymbol{\beta}$ and vector $\mathbf{u}_S$ defined previously and $\mathbf{W}_S=(\mathbf{X},\mathbf{Z}_S)$ the corresponding incidence matrix, then $R(\boldsymbol{\theta}_S)$ is the reduction due to fitting the fixed model $\mathbf{y}=\mathbf{W}_S\boldsymbol{\theta}_S+\mathbf{e}$, that is

$$R(\boldsymbol{\theta}_S)=\hat{\boldsymbol{\theta}}_S'\mathbf{W}_S'\mathbf{y}=\mathbf{y}'\mathbf{P}_S\mathbf{y},$$

where $\hat{\boldsymbol{\theta}}_S'$ is the LS estimator of $\boldsymbol{\theta}_S$, and $\mathbf{P}_S$ stands in short for the usual projector $\mathbf{P}_{W_S}=\mathbf{W}_S\left(\mathbf{W}_S'\mathbf{W}_S\right)^{-}\mathbf{W}_S$ defined in (1.16).

Applying the classical result about the expectation of quadratic forms (1.33) $E(\mathbf{y}'\mathbf{Q}\mathbf{y})=\boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu}+tr(\mathbf{Q}\mathbf{V})$ to $R(\boldsymbol{\theta}_S)$ yields

$$E(\mathbf{y}'\mathbf{P}_S\mathbf{y})=\boldsymbol{\beta}'\mathbf{X}'\mathbf{P}_S\mathbf{X}\boldsymbol{\beta}+tr(\mathbf{P}_S)\sigma_0^2+\sum_{k=1}^{K}tr(\mathbf{P}_S\mathbf{Z}_k\mathbf{Z}_k')\sigma_k^2.$$

Simplification of this expression comes from the properties of $\mathbf{P}_S$ so that

$$tr(\mathbf{P}_S)=r(\mathbf{P}_S)=r(\mathbf{W}_S),$$

$$\mathbf{P}_S\mathbf{W}_S=\mathbf{W}_S.$$

In particular, since $\mathbf{W}_S=(\mathbf{X},\mathbf{Z}_S)$, this means that

$$\mathbf{P}_S\mathbf{X}=\mathbf{X},$$

$$\mathbf{P}_S\mathbf{Z}_k=\mathbf{Z}_k \text{ for } k\in S.$$

Moreover, $\mathbf{Z}_k$ being an incidence matrix with each row of $\mathbf{Z}_k$ having its elements equal to 0 but one equal to 1, $tr(\mathbf{P}_S\mathbf{Z}_k\mathbf{Z}_k')=N$ for $k\in S$. This implies

$$\begin{aligned}E\left[R(\boldsymbol{\theta}_S)\right]=&\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+r(\mathbf{P}_S)\sigma_0^2+N\sum_{k\in S}\sigma_k^2\\&+\sum_{k\notin S}tr(\mathbf{P}_S\mathbf{Z}_k\mathbf{Z}_k')\sigma_k^2\end{aligned}\qquad(3.5.3)$$

Now, if we contrast two reductions such as $R(\boldsymbol{\beta},\mathbf{u})$ and $R(\boldsymbol{\beta},\mathbf{u}_S)$, the expectation of the difference reduces to

$$\begin{aligned}E\left[R(\mathbf{u}_{\bar{S}}\mid\boldsymbol{\beta},\mathbf{u}_S)\right]=&\left[r(\mathbf{X},\mathbf{Z})-r(\mathbf{X},\mathbf{Z}_S)\right]\sigma_0^2\\&+\sum_{k\in\bar{S}}tr(\mathbf{I}_N-\mathbf{P}_S\mathbf{Z}_k\mathbf{Z}_k')\sigma_k^2\end{aligned}\qquad(3.5.4)$$

where $\overline{S}$ is the complementary set of S with respect to $\mathcal{I}$. [2]

This means among other things that this expectation does not involve any term in $\boldsymbol{\beta}$, nor components of variance occuring in $\mathbf{u}_S$, but only coefficients for the residual σ_0^2 and for variance components being in $\mathbf{u}$ but not in $\mathbf{u}_S$. For instance, suppose that $\mathbf{u}=(\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3)$ and $\mathbf{u}_S=(\mathbf{u}_1,\mathbf{u}_2)$, then

$$E\left[R(\mathbf{u}_3 \mid \boldsymbol{\beta},\mathbf{u}_1,\mathbf{u}_2)\right]=\left[r(\mathbf{X},\mathbf{Z})-r(\mathbf{X},\mathbf{Z}_1,\mathbf{Z}_2)\right]\sigma_0^2 + \left[N-tr\left(\mathbf{I}_N-\mathbf{P}_{12}\mathbf{Z}_3\mathbf{Z}_3'\right)\right]\sigma_3^2$$

where $\mathbf{P}_{12}$ refers to the projector on $(\mathbf{X},\mathbf{Z}_1,\mathbf{Z}_2)$, and $\mathbf{Z}$ (respectively $\mathbf{Z}_3$) to the incidence matrix pertaining to $\mathbf{u}$ (respectively $\mathbf{u}_3$).

This formula clearly highlights the fact that only reductions of submodels including $\boldsymbol{\beta}$ are useful since all of them have the same quadratic form $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$, and this nuisance term vanishes when two reductions are contrasted.

Let us see now how to estimate the residual variance σ_0^2. Applying (3.5.3) to the complete model $(\boldsymbol{\beta},\mathbf{u})$, one has

$$E\left[R(\boldsymbol{\beta},\mathbf{u})\right]=\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+r(\mathbf{W})\sigma_0^2+N\sum_{k=1}^{K}\sigma_k^2 .$$

Contrasting the total sum of squares $\mathbf{y}'\mathbf{y}$ and this reduction $R(\boldsymbol{\beta},\mathbf{u})$ yields the usual *SSE* quadratic that occurred in the LS theory. On account of the expectations of $\mathbf{y}'\mathbf{y}$ (formula 3.5.2) and $R(\boldsymbol{\beta},\mathbf{u})$ given above, it turns out that the expectation of *SSE* under the true mixed model $(\boldsymbol{\beta},\mathbf{u})$ is the same as in the completely fixed model v.i.z

$$E(SSE)=\left[N-r(\mathbf{W})\right]\sigma_0^2,$$

so that an unbiased estimator of σ_0^2 is

$$\hat{\sigma}_0^2=SSE/\left[N-r(\mathbf{X},\mathbf{Z})\right]. \qquad (3.5.5)$$

There are K remaining unknowns that require K quadratic forms and their LIN expectations. In many cases, there are more possible such reductions (maximum

2^K) than unknowns so that H-III does not provide unique solutions for a given model due to some arbitrariness in choosing the reductions.

Finally, letting $\mathbf{q} = \left({}_c\mathbf{y}'\mathbf{Q}_k\mathbf{y} \right)_{0 \le k \le K}$ be the vector of the $K+1$ LIN quadratic forms chosen to estimate the variance components $\boldsymbol{\sigma}^2 = \left({}_c\sigma_k^2 \right)_{0 \le k \le K}$ such that $E(\mathbf{q}) = \mathbf{F}\boldsymbol{\sigma}^2$ with $\mathbf{F}$ non singular, H-III estimators are obtained by equating the realized value of $\mathbf{q}$ to its expectation expressed as a linear function of $\boldsymbol{\sigma}^2$, that is $\mathbf{F}\hat{\boldsymbol{\sigma}}^2 = \mathbf{q}$ leading to

$$\hat{\boldsymbol{\sigma}}^2 = \mathbf{F}^{-1}\mathbf{q}. \qquad (3.5.6)$$

3.5.2 Sampling variances

From (3.5.6), we can derive the expression of the sampling variances of $\hat{\boldsymbol{\sigma}}^2$

$$Var\left(\hat{\boldsymbol{\sigma}}^2\right) = \mathbf{F}^{-1}Var(\mathbf{q})\mathbf{F}^{-1}. \qquad (3.5.7)$$

To explicit $Var(\mathbf{q})$, we use the following result

Result. For symmetric $\mathbf{A}$ and $\mathbf{B}$ and assuming that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, then

$$Cov\left(\mathbf{y}'\mathbf{A}\mathbf{y}, \mathbf{y}'\mathbf{B}\mathbf{y}\right) = 4\boldsymbol{\mu}'\mathbf{A}\mathbf{V}\mathbf{B}\boldsymbol{\mu} + 2tr\left(\mathbf{A}\mathbf{V}\mathbf{B}\mathbf{V}\right). \qquad (3.5.8)$$

Here, the quadratic forms $\mathbf{y}'\mathbf{Q}_{kl}\mathbf{y}$ entering $\mathbf{q}$ are obtained as differences of quadratic forms defined in (3.5.4) with $\mathbf{Q}_{kl} = \mathbf{P}_k - \mathbf{P}_l$ so that $\mathbf{Q}_{kl}\boldsymbol{\mu} = \mathbf{0}$ because $\mathbf{P}_k$ and $\mathbf{P}_l$ are projectors. Therefore, the element (i, j) is

$$Cov\left(\mathbf{y}'\mathbf{Q}_i\mathbf{y}, \mathbf{y}'\mathbf{Q}_j\mathbf{y}\right) = 2tr\left(\mathbf{Q}_i\mathbf{V}\mathbf{Q}_j\mathbf{V}\right).$$

This gives complicated expressions as shown for instance with two random factors and a residual $\boldsymbol{\sigma}^2 = \left(\sigma_0^2, \sigma_1^2, \sigma_2^2\right)'$. In that case, after applying the previous formulae, the element (i, j) of $Var(\mathbf{q})$ can be expressed as $\boldsymbol{\sigma}^{2'}\mathbf{T}_{ij}\boldsymbol{\sigma}^2$ where $\mathbf{T}_{ij}$ is a $(3\text{x}3)$ symmetric matrix defined as

$$\mathbf{T}_{ij}=\begin{pmatrix} tr\left(\mathbf{Q}_i\mathbf{Z}_0\mathbf{Z}_0'\mathbf{Q}_j\mathbf{Z}_0\mathbf{Z}_0'\right) & tr\left(\mathbf{Q}_i\mathbf{Z}_0\mathbf{Z}_0'\mathbf{Q}_j\mathbf{Z}_1\mathbf{Z}_1'\right) & tr\left(\mathbf{Q}_i\mathbf{Z}_0\mathbf{Z}_0'\mathbf{Q}_j\mathbf{Z}_2\mathbf{Z}_2'\right) \\ & tr\left(\mathbf{Q}_i\mathbf{Z}_1\mathbf{Z}_1'\mathbf{Q}_j\mathbf{Z}_1\mathbf{Z}_1'\right) & tr\left(\mathbf{Q}_i\mathbf{Z}_1\mathbf{Z}_1'\mathbf{Q}_j\mathbf{Z}_2\mathbf{Z}_2'\right) \\ \textit{symmetric} & & tr\left(\mathbf{Q}_i\mathbf{Z}_2\mathbf{Z}_2'\mathbf{Q}_j\mathbf{Z}_2\mathbf{Z}_2'\right) \end{pmatrix}.$$

Since the expression of $\mathbf{y}'\mathbf{Q}_{kl}\mathbf{y}$ is generally tedious, it has been suggested to replace $\mathbf{y}'\mathbf{P}_k\mathbf{y}$ by its equivalent form $\mathbf{r}_k'\mathbf{G}_k\mathbf{r}_k$ where $\mathbf{r}_k=\mathbf{W}_k'\mathbf{y}$ is the right hand side of the LS equations due to fitting $\mathbf{y}=\mathbf{W}_k\boldsymbol{\theta}_k+\mathbf{e}$ and $\mathbf{G}_k=\left(\mathbf{W}_k'\mathbf{W}_k\right)^-$. Then, one can express $Var(\mathbf{q})$ accordingly as a function of $\boldsymbol{\sigma}^2$ and of the $\mathbf{r}_k$ and $\mathbf{G}_k$'s knowing that

$$Cov\left(\mathbf{r}_k'\mathbf{G}_k\mathbf{r}_k,\mathbf{r}_l'\mathbf{G}_l\mathbf{r}_l\right)=2tr\left(\mathbf{G}_k\mathbf{C}_{kl}\mathbf{G}_l\mathbf{C}_{kl}\right),$$

where $\mathbf{C}_{kl}=\mathbf{W}_k'\mathbf{W}_l\sigma_0^2+\sum_{i=1}^{K}\left(\mathbf{W}_k'\mathbf{Z}_i\mathbf{Z}_i'\mathbf{W}_l\right)\sigma_i^2$.

3.5.3 Example. *Estimation of variance components in the 2-way mixed linear model with interaction.*

Data are collected according to a 2-way cross classified design (see examples 1.4 and 2.4) and analyzed accordingly by a two factor (say A and B) mixed linear model with interaction but here assuming A being fixed, B and consequently AxB random

$$y_{ijk}=\mu+a_i+b_j+c_{ij}+e_{ijk},$$

where y_{ijk} is the k^{th} observation $\left(k=1,...,n_{ij}\right)$ from the i^{th} level $\left(i=1,...,I\right)$ of factor A and the j^{th} level $\left(j=1,...,J\right)$ of factor B; μ is the general mean, a_i is the fixed effect of level i, b_j is the random effect of level j, c_{ij} is the corresponding random interaction and e_{ijk} the residual term.

We assume that $b_j \sim_{iid}\left(0,\sigma_b^2\right)$, $c_{ij}\sim_{iid}\left(0,\sigma_c^2\right)$, $e_{ijk}\sim_{iid}\left(0,\sigma_e^2\right)$, and b_j, c_{ij}, e_{ijk} are uncorrelated among them. Unknown parameters are σ_b^2, σ_c^2, σ_e^2 and the

quadratic form in fixed effects $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ equal here to $\sum_{i=1}^{I} n_{i0}\left(\mu + a_i\right)^2$ as a nuisance quantity. We need four quadratic forms to estimate them which apart from the total sum of squares $\mathbf{y}'\mathbf{y}$ must all include A-effects in the corresponding reductions that are $R(\mu,a)$, $R(\mu,a,b)$ and $R(\mu,a,b,c)$.

Let the model be written in matrix notations as

$$\mathbf{y} = \mathbf{1}_N \mu + \mathbf{X}_a\mathbf{a} + \mathbf{Z}_b\mathbf{b} + \mathbf{Z}_c\mathbf{c} + \mathbf{e}.$$

Using this, we can express the reductions as

$$R(\mu,a) = \mathbf{y}'\mathbf{X}_a\left(\mathbf{X}_a'\mathbf{X}_a\right)^{-1}\mathbf{X}_a'\mathbf{y}.$$

$$R(\mu,a,b) = \mathbf{y}'\mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{y},$$

with $\mathbf{W} = \left(\mathbf{1},\mathbf{X}_a,\mathbf{Z}_b\right)$.

$$R(\mu,a,b,c) = \mathbf{y}'\mathbf{W}_F\left(\mathbf{W}_F'\mathbf{W}_F\right)^{-1}\mathbf{W}_F'\mathbf{y},$$

with $\mathbf{W}_F = \left(\mathbf{1},\mathbf{X}_a,\mathbf{Z}_b,\mathbf{Z}_c\right)$.

Notice that $R(\mu,a)$ and $R(\mu,a,b,c)$ can be simply expressed as sums of squares

$$R(\mu,a) = \sum_{i=1}^{I}\frac{y_{i00}^2}{n_{i0}}\ ,\ R(\mu,a,b,c) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{y_{ij0}^2}{n_{ij}}$$

since equivalent cell-mean models are $E\left(y_{ijk}\right) = \mu_i$ and $E\left(y_{ijk}\right) = \mu_{ij}$ respectively with LS estimations $\hat{\mu}_i = y_{i..} = y_{i00}/n_{i0}$ and $\hat{\mu}_{ij} = y_{ij.} = y_{ij0}/n_{ij}$.

The next step consists of expressing the expectation of these quadratic forms under the true model. Using (3.5.3), one has

$$E\left[R(\mu,a)\right] = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + r\left(\mathbf{X}_a\right)\sigma_e^2 + tr\left(\mathbf{P}_a\mathbf{Z}_b\mathbf{Z}_b'\right)\sigma_b^2 + tr\left(\mathbf{P}_a\mathbf{Z}_c\mathbf{Z}_c'\right)\sigma_c^2$$

where $\mathbf{P}_a = \mathbf{X}_a \left(\mathbf{X}_a' \mathbf{X}_a \right)^{-1} \mathbf{X}_a'$.

Here $r\left(\mathbf{X}_a\right) = I$; $\mathbf{X}_a' \mathbf{Z}_b$ is a $\left(I \text{x} J\right)$ matrix with ij term $\left(\mathbf{X}_a' \mathbf{Z}_b\right)_{ij} = n_{ij}$; $\mathbf{X}_a' \mathbf{Z}_b \mathbf{Z}_b' \mathbf{X}_a$ has diagonal i term $\left(\mathbf{X}_a' \mathbf{Z}_b \mathbf{Z}_b' \mathbf{X}_a\right)_{ii} = \sum_{j=1}^{J} n_{ij}^2$; same thing for $\mathbf{X}_a' \mathbf{Z}_c \mathbf{Z}_c' \mathbf{X}_a$ so that

$$E\left[R\left(\mu, a\right) \right] = \sum_{i=1}^{I} n_{i0} \left(\mu + a_i \right)^2 + I \sigma_e^2 + k \left(\sigma_b^2 + \sigma_c^2 \right),$$

with $k = \sum_{i=1}^{I} n_{i0}^{-1} \sum_{j=1}^{J} n_{ij}^2$.

Similarly since $r\left(\mathbf{W}\right) = I + J - 1$ and effect $\mathbf{b}$ belongs to the reduction,

$$E\left[R\left(\mu, a, b\right) \right] = \sum_{i=1}^{I} n_{i0} \left(\mu + a_i \right)^2 + \left(I + J - 1 \right) \sigma_e^2 + N \sigma_b^2 + h \sigma_c^2,$$

with $h = tr\left[\left(\mathbf{W}'\mathbf{W} \right)^{-} \mathbf{W}' \mathbf{Z}_c \mathbf{Z}_c' \mathbf{W} \right]$.

Finally,

$$E\left[R\left(\mu, a, b, c\right) \right] = \sum_{i=1}^{I} n_{i0} \left(\mu + a_i \right)^2 + s \sigma_e^2 + N \left(\sigma_b^2 + \sigma_c^2 \right),$$

where s represents the number of non empty cells in the design, that is IJ if the design is complete but it is not necessarily the case (see next example).

In order to eliminate the term involving the fixed effects, we can build the following quadratic forms as in an ANOVA type table:

$$SSE = \mathbf{y}'\mathbf{y} - R\left(\mu, a, b, c\right),$$

$$R\left(c \mid \mu, a, b\right) = R\left(\mu, a, b, c\right) - R\left(\mu, a, b, c\right),$$

$$R\left(b \mid \mu, a\right) = R\left(\mu, a, b\right) - R\left(\mu, a\right).$$

From what has been seen in the previous formulae, these quadratic forms have expectations

$$E\left(SSE\right) = \left(N - s\right) \sigma_e^2,$$

$$E\left[R\left(c\,|\,\mu,a,b\right)\right]=\left(s-I-J+1\right)\sigma_e^2+\left(N-h\right)\sigma_c^2,$$

$$E\left[R\left(b\,|\,\mu,a\right)\right]=\left(J-1\right)\sigma_e^2+\left(h-k\right)\sigma_c^2+\left(N-k\right)\sigma_b^2.$$

Equating these expressions to the observed values, one obtains a linear system that can be easily solved row by row starting from the first. Alternatively, under a matrix form, it looks as follows

$$\begin{pmatrix} N-s & 0 & 0 \\ s-I-J+1 & N-h & 0 \\ J-1 & h-k & N-k \end{pmatrix}\begin{pmatrix} \hat{\sigma}_e^2 \\ \hat{\sigma}_c^2 \\ \hat{\sigma}_b^2 \end{pmatrix}=\begin{pmatrix} SSE \\ R\left(c\,|\,\mu,a,b\right) \\ R\left(b\,|\,\mu,a\right) \end{pmatrix}.$$

These formulae can be easily applied to the following toy example.

Table 3.5.1 *Distribution of data (sums) according to the levels of factors A and B*

A	B	n	$\sum y$
1	1	2	100
1	2	2	140
1	3	6	480
2	1	5	300
2	2	9	810

$\mathbf{y}'\mathbf{y}=159800$

Using these data sets, we can compute the different reductions required for estimating the variance components. We have

$$R\left(\mu,a\right)=\frac{\left(100+148+480\right)^2}{12}+\frac{\left(300+810\right)^2}{14}$$
$$=139847.1429$$

$$R(\mu,a,b,c)=\frac{100^2}{2}+\frac{140^2}{2}+\frac{480^2}{6}+\frac{300^2}{5}+\frac{810^2}{9}$$
$$=144100$$

In order to calculate $R(\mu,a,b)$, the fixed effects are parameterized as $\mu+a_1+b_1,\ a_2-a_1,\ b_2-b_1,\ b_3-b_1$ so that

$$\mathbf{W'W}=\begin{pmatrix}24 & 14 & 11 & 6\\ 14 & 14 & 9 & 0\\ 11 & 9 & 9 & 11\\ 6 & 0 & 0 & 0\end{pmatrix},\ \mathbf{W'y}=\begin{pmatrix}1830 & 1110 & 950 & 48\end{pmatrix}.$$

$R(\mu,a,b)=\mathbf{y'W}\left(\mathbf{W'W}\right)^{-1}\mathbf{W'y}=144023.7288$, $R(b\,|\,\mu,a)=4176.5859$ and

$R(c\,|\,\mu,a,b)=76.2712$.

In addition, here $N=24$, $I=2$, $s=5$ since the lay-out has cell 3x2 empty and

$$MSE=(159800-144100)/(24-5)=826.3158,$$

$$k=\frac{2^2+2^2+6^2}{10}+\frac{5^2+9^2}{14}=11\tfrac{34}{35}.$$

Computing coefficient h requires $\mathbf{W'W}$ as shown previously and $\mathbf{W'Z}_c$ that is

$$\mathbf{W'Z}_c=\begin{pmatrix}2 & 2 & 6 & 5 & 9 & 0\\ 0 & 0 & 0 & 5 & 9 & 0\\ 0 & 2 & 0 & 0 & 9 & 0\\ 0 & 0 & 6 & 0 & 0 & 0\end{pmatrix};$$

the last column made of zeroes could have been ignored; however, it reminds us that there is no data for cell 3x2 although the corresponding c_{23} random effect still exists. Using the values of $\mathbf{W'W}$ and $\mathbf{W'Z}_c$ gives

$$h=tr\left[\left(\mathbf{W'W}\right)^{-1}\mathbf{W'Z}_c\mathbf{Z}_c'\mathbf{W}\right]=20.949152.$$

Then the system to solve reduces to

$$\begin{pmatrix} 19 & 0 & 0 \\ 1 & 24-h & 0 \\ 2 & h-k & 24-k \end{pmatrix}\begin{pmatrix} \hat{\sigma}_e^2 \\ \hat{\sigma}_c^2 \\ \hat{\sigma}_b^2 \end{pmatrix}=\begin{pmatrix} 826.3158 \\ 76.2712 \\ 4176.5859 \end{pmatrix}.$$

Its solutions and SE are $\hat{\sigma}_e^2 = 826.32 \pm 259.55$, $\hat{\sigma}_c^2 = -245.85 \pm 1089.42$ and $\hat{\sigma}_b^2 = 393.32 \pm 1006.70$. SE values reported here are computed under the assumption of normality of random effects and under the premice that true values of variance components are $\hat{\sigma}_e^2 = 800$, $\hat{\sigma}_c^2 = 100$ and $\hat{\sigma}_b^2 = 200$.

The results can also be displayed as in some software (e.g. SAS-Proc Mixed or Proc Varcomp) under the classical format of an ANOVA table.

Table 3.5.2 *ANOVA table for H-III estimators of variance components in a 2 way crossclassification with factor A fixed and B random*

Source	DF	Sums of Squares	Mean Square	Expected Mean square
A	1	No object	No object	No object
B	2	4176.5959	2088.2930	$\sigma_e^2 + 4.4889\sigma_c^2 + 6.0143\sigma_b^2$
C=AxB	1	76.2712	76.2712	$\sigma_e^2 + 3.0558\sigma_c^2$
Residual	19	15700	826.3158	σ_e^2

Notice that the estimation of σ_c^2 is clearly negative. This result reminds us of an important feature of this method and more generally of ANOVA-based quadratic estimators of variance components. Apart from σ_e^2, they can produce negative estimates. Although setting such estimates to zero might appear as common sense, this adjustement is not completely satisfactory since it violates the property of unbiasedness. In fact, it means that we are dealing with a new

model which ignores those random components implying to re-estimate the remaining ones under that model. In this case, we come back to an additive mixed linear model $y_{ijk} = \mu + a_i + b_j + e_{ijk}$, where a_i is the fixed effect of level i, $b_j \sim_{iid} \left(0, \sigma_b^2\right)$ and $e_{ijk} \sim_{iid} \left(0, \sigma_e^2\right)$.

Estimation of σ_e^2 and σ_b^2 are based on $SSE^* = \mathbf{y}'\mathbf{y} - R\left(\mu, a, b\right)$ and $R\left(b \mid \mu, a\right)$ respectively so that

$$E\left(SSE^*\right) = \left(N - I - J + 1\right)\sigma_e^2,$$

$$E\left[R\left(b \mid \mu, a\right)\right] = \left(J - 1\right)\sigma_e^2 + \left(N - k\right)\sigma_b^2,$$

where $k = \sum_{ij} n_{ij}^2 / n_{i0}$ as before.

This yields

$$\hat{\sigma}_e^2 = \frac{159800 - 144023.7288}{20} = 788.81,$$

$$\hat{\sigma}_b^2 = \frac{4176.5859 - 2\hat{\sigma}_e^2}{12.0285} = 216.07.$$

Happily, there is no ambiguity in the choice of quadratic forms for these two mixed models : the additive and the interactive ones. But, it is by far not always the case. For instance, consider the additive purely random model with three variance components σ_a^2, σ_b^2 and σ_e^2. In such a case, we have at least four different choices of quadratic forms in order to estimate the variance components:

a) $R\left(a \mid \mu\right)$, $R\left(b \mid \mu, a\right)$, $\mathbf{y}'\mathbf{y} - R\left(\mu, a, b\right)$

b) $R\left(b \mid \mu\right)$, $R\left(a \mid \mu, b\right)$, $\mathbf{y}'\mathbf{y} - R\left(\mu, a, b\right)$

c) $R\left(a \mid \mu, b\right)$, $R\left(b \mid \mu, a\right)$, $\mathbf{y}'\mathbf{y} - R\left(\mu, a, b\right)$

d) $R(a \mid \mu)$, $R(b \mid \mu)$, $\mathbf{y}'\mathbf{y} - R(\mu)$.

Choices a) and b) rely on asymmetric reductions corresponding to fitting explanatory variables upward one by one (the so-called type I reductions in SAS-GLM terminology) while c) and d) involve symmetric reductions due to fitting each factor either in c) after all other terms have been added (Type II) or in d) after only the intercept. None of them looks a priori better than the others : see exercise.

3.6. Appendix : Information matrices

ML Estimation

The starting point consists in the expression of $-2\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y})$ written as

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = N\ln(2\pi) + \ln|\mathbf{V}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \qquad (3.6.1)$$

where $l(\boldsymbol{\beta},\boldsymbol{\gamma}) = -2\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{y}) = -2\ln f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})$.

We already saw that the first derivatives can be expressed as:

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \qquad (3.6.2)$$

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_k} = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \ . \qquad (3.6.3)$$

From these, we can derive the expressions of the second derivatives

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \qquad (3.6.4)$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}\partial \gamma_k} = 2\mathbf{X}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \qquad (3.6.5)$$

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_l} &= \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l}\right) - \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right) \\ &\quad -(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\left(\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} - 2\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right)\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\end{aligned} \qquad (3.6.6)$$

After dividing by two, these formulae provide the terms entering the observed information matrix $\mathbf{I}(\hat{\boldsymbol{\alpha}};\mathbf{y}) = -\left.\frac{\partial^2 \mathrm{L}(\boldsymbol{\alpha};\mathbf{y})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}$ where $\boldsymbol{\alpha} = (\boldsymbol{\beta}',\boldsymbol{\gamma}')'$, that are used in the Newton-Raphson algorithm.

The Fisher information matrix $\mathbf{J}(\boldsymbol{\alpha}) = \mathrm{E}[\mathbf{I}(\boldsymbol{\alpha};\mathbf{y})]$ is obtained after taking the expectation of $\mathbf{I}(\boldsymbol{\alpha};\mathbf{y})$, that is:

$$\mathbf{J}_{\beta\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \qquad (3.6.7)$$

$$\mathbf{J}_{\beta\gamma} = \mathbf{0}, \qquad (3.6.8)$$

$$\left(\mathbf{J}_{\gamma\gamma}\right)_{kl} = \frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \qquad (3.6.9)$$

Two remarks are worthwhile at this stage. First, it turns out from (3.6.8) that the ML estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are asymptotically uncorrelated. Secondly, formulae (3.6.7-8-9) apply as well to linear and nonlinear $\mathbf{V}$ structures.

REML estimation

The logresidual likelihood can be written as

$$r(\boldsymbol{\gamma}) = \left[N - \mathrm{r}(\mathbf{X})\right]\ln 2\pi + \ln\left|\mathbf{V}\right| + \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right| + \mathbf{y}'\underline{\mathbf{P}}\mathbf{y} \qquad (3.6.10)$$

where $r(\boldsymbol{\gamma}) = -2\mathrm{L}\left(\boldsymbol{\gamma}; \mathbf{K}'\mathbf{y}\right)$.

After differentiating with respect to γ_k, one has:

$$\frac{\partial r(\boldsymbol{\gamma})}{\partial \gamma_k} = \frac{\partial \ln\left|\mathbf{V}\right|}{\partial \gamma_k} + \frac{\partial \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial \gamma_k} + \mathbf{y}'\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k}\mathbf{y} \qquad (3.6.11)$$

Now,

$$\frac{\partial \ln\left|\mathbf{V}\right|}{\partial \gamma_k} = tr\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right),$$

$$\frac{\partial \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial \gamma_k} = -\mathrm{tr}\left[\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\mathbf{V}^{-1}\underline{\mathbf{X}}\right]$$

$$= -\mathrm{tr}\left[\mathbf{V}^{-1}\underline{\mathbf{X}}\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right]$$

which allows to single out and factorize the $\underline{\mathbf{P}}$ matrix

$$\frac{\partial \ln\left|\mathbf{V}\right|}{\partial \gamma_k} + \frac{\partial \ln\left|\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right|}{\partial \gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right). \qquad (3.6.12)$$

It remains to make $\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k}$ explicit. By definition, $\mathbf{V}\underline{\mathbf{P}} = \left(\mathbf{I} - \mathbf{Q}\right)$ with $\mathbf{Q} = \underline{\mathbf{X}}\left(\underline{\mathbf{X}}'\mathbf{V}^{-1}\underline{\mathbf{X}}\right)^{-}\underline{\mathbf{X}}'\mathbf{V}^{-1}$ so that its first derivative is

$$\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}} + \mathbf{V}\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\frac{\partial \mathbf{Q}}{\partial \gamma_k}. \qquad (3.6.13)$$

Moreover, by directly differentiating the expression of $\mathbf{Q}$

$$\frac{\partial \mathbf{Q}}{\partial \gamma_k} = -\mathbf{Q}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}.$$

Hence, after substitution in (3.6.13), $\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\mathbf{V}^{-1}(\mathbf{I}-\mathbf{Q})\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}$, that is

$$\frac{\partial \underline{\mathbf{P}}}{\partial \gamma_k} = -\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}. \tag{3.6.14}$$

Therefore, the residual score is:

$$\frac{\partial r(\boldsymbol{\gamma})}{\partial \gamma_k} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - \mathbf{y}'\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\mathbf{y}. \tag{3.6.15}$$

We can check that the expectation of the score is zero.

$$\mathrm{E}\left(\frac{\partial r(\boldsymbol{\gamma})}{\partial \gamma_k}\right) = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\right) - \mathrm{tr}\left[\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\,\mathrm{E}(\mathbf{y}\mathbf{y}')\right]$$

Now $\mathrm{E}(\mathbf{y}\mathbf{y}') = \underline{\mathbf{X}}\boldsymbol{\beta}\boldsymbol{\beta}'\underline{\mathbf{X}}' + \mathbf{V}$, and because $\underline{\mathbf{P}}\underline{\mathbf{X}} = \mathbf{0}$ and $\underline{\mathbf{P}}\mathbf{V}\underline{\mathbf{P}} = \underline{\mathbf{P}}$, the second term :is equal to the first (QED).

The Hessian is derived by differentiating (3.6.15); it can be expressed as for ML in (3.6.6)

$$\begin{aligned}\frac{\partial^2 r(\boldsymbol{\gamma};\mathbf{y})}{\partial \gamma_k \partial \gamma_l} = \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l}\right) - \mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right) \\ -\mathbf{y}'\underline{\mathbf{P}}\left(\frac{\partial^2 \mathbf{V}}{\partial \gamma_k \partial \gamma_l} - 2\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right)\underline{\mathbf{P}}\mathbf{y}\end{aligned}. \tag{3.6.16}$$

The Fisher information follows

$$\left(\mathbf{J}_{\gamma\gamma}\right)_{kl} = \frac{1}{2}\mathrm{tr}\left(\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_k}\underline{\mathbf{P}}\frac{\partial \mathbf{V}}{\partial \gamma_l}\right). \tag{3.6.17}$$

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Editors), *Second International Symposium of Information Theory*, pages 267-281. Akademiai Kiado, Budapest.

Anderson R.L., & Bancroft T.A. (1952). *Statistical theory in research*. Mc Graw-Hill, New-York.

Berger, J.O., Liseo, B. & Wolpert, R. L. (1999). Integrated Likelihood methods for eliminating nuisance parameters. *Statistical Science*, **14**, 1-28.

Brown, H. K., & Prescott R. J. (2006). *Applied Mixed Models in Medicine*, 2nd edition. Wiley, New York.

Cox, D. R., & Reid N. (1987) Parameter orthogonality and approximate conditional inference, *Journal of the Royal Statistical Society* B, **49**, 1-39.

Cox, D. R., & Hinkley D.V. (1974) *Theoretical statistics*, Chapman & Hall, London.

Cressie, N., & Lahiri S.N. (1993) The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*, **45**, 217-233.

Crump, S. L. (1947). The estimation of components of variance in multiple classifications. PhD thesis, Iowa State University, Ames.

Dawid A.P. (1980). A Bayesian look at nuisance parameters. Proceedings of the first international meeting held in Valencia. In Bernardo J.M. DeGroot M.H. Lindley D.V. Smith A.F.M. (Eds), University Press, Valencia, Spain, 167-184.

Delmas C. & Foulley J-.L. (2007). On testing a class of restricted hypothesis. *Journal of Statistical Planning and Inference,* **137**, 1343-1361

Edwards, A.W. F. (1972). *Likelihood*, Cambridge University Press, Cambridge

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*, **3**, 1-21.

Fai, A. H. T. & Cornelius, P.L. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiment. *Journal of Statistical Computing and Simulation*, 54: 363-378.

Fisher R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series* A, 222, 309-368.

Fisher R. A. (1925). *Statistical methods for research workers*, Oliver and Boyd. Edinburgh and London

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2004). *Applied longitudinal analysis*. John Wiley & Sons, New York.

Foulley J-. L. (1993). A simple argument showing how to derive restricted maximum likelihood, *Journal of Dairy Science*, **76**, 2320-2324.

Friel, N., & Wyse J. (2012). Estimating the statistical evidence -a review. *Statistica Neerlandica*, **66**, 288-308.

Giesbrecht, F. G., & Burns J. C. (1985). Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics*, 41, 853-862.

Gilmour, A. R., Thompson R., & Cullis B. R. (1995). An efficient algorithm for REML estimation in linear mixed models. *Biometrics*, **51**, 1440-1450.

Goffinet, B. (1983). Risque quadratique et sélection : quelques résultats appliqués à la sélection animale et végétale, Thèse de Docteur Ingénieur, Université Paul Sabatier, Toulouse.

Good, I. J. (1992). The Bayes/Non Bayes compromise: a brief review. *Journal of the American Statistical Association*, **87**, 597-606.

Gourieroux, C., & Montfort A. (1989). *Statistique et modèles économétriques*, Economica, France.

Gurka, M (2006). Selecting the best linear mixed model under REML. *The American Statistician*, **60**, 19-26.

Hartley, H. O., and Rao J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika*, 54, 93-108.

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

Harville D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, **72**, 320-340.

Harville, D. A., & Callanan T.P. (1990). *Computational aspects of likelihood based inference for variance components*; In: Gianola D. Hammond K. (Editors), *Advances in statistical methods for genetic improvement of livestock*. New York, Heidelberg, Berlin: Springer Verlag.

Harville, D.A. (1997). *Matrix algebra from a statistician's perspective*. Springer, New York.

Henderson, C.R. (1953), Estimation of variance and covariance components. *Biometrics* 9, 226-252

Henderson, C.R. (1973). Sire evaluation and genetic trends, In: *Proceedings of the animal breeding and genetics symposium in honor of Dr J Lush*. American Society Animal Science-American Dairy Science Association, 10-41, Champaign, IL.

Henderson, C.R. (1984). *Applications of linear models in animal breeding*, University of Guelph, Guelph, 1984.

Henderson C. R., Kempthorne O., Searle S. R. & von Krosigk, C.N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **13**, 192-218

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika,* 76, 297-307.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edition. The Clarendon Press, Oxford.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.

Jones R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in. Medicine,* **30**, 3050–3056.

Kackar, A. N. & Harville, D. A. (1984). Approximation for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853-862

Kalbfleisch, J. D., & Sprott D. A. (1970). Application of the likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society* B, **32**, 175-208.

Kass R. E., & Raftery A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983-997

LaMotte, L.R. (1973). Quadratic estimation of variance components, *Biometrics,* **29**, **311-330**

Laskar, M. R., & King M. L., (2001). Modified likelihood and related method for handling nuisance parameters in the linear regression model, In Saleh A.K.M.E. (editor). *Data Analysis from Statistical Foundations*, Nova Science Publisher, Inc., Huntington, New York, 119-142.

Latreille, J., Guinot, C., Robert-Granié, C., Le Fur, I., Tenenhaus, M., & Foulley J-.L. (2004). Daily variations in skin surface using mixed model methodology. *Skin Pharmacology and physiology*, **17**, 133-140.

Lebarbier, E., & Mary-Huard, T. (2006). Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la Société Française de Statistique,* **147**, 39-57.

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

Lindley, D.V., & Smith A. F. M. (1972). Bayes Estimates for the Linear Model, *Journal of the Royal Statistical Society* B, **34**, 1-41.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (2006). SAS System for Mixed Models, 2nd edition. SAS Institute Inc, Cary, NC.

Mardia, K. V., & Marshall, R. J. (1985). Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika*, **71**, 135-146.

McCullagh, P., & Nelder J. (1989). *Generalized linear models*, 2nd edition. Chapman and Hall, London.

McBride, J. B. (2000). Adequacy of approximations to distributions of test statistics in complex mixed linear models. M.S. Project, Brigham Young University.

Meyer, K. (1990). Present status of knowledge about statistical procedures and algorithms available to estimate variance and covariance components. *Proceedings of the Fourth World Congress on Genetics Applied to Livestock Production,* Vol. **13**,407-419.

Mood, A. M., Graybill F. A., & Boes D. C. (1974). *Introduction to the theory of statistics*, 3rd edition. McGraw-Hill. Tokyo.

Mrode, R. A., & Thompson, R. (2005). *Linear models for the prediction of animal breeding values*, 2nd edition, CABI Publishing, Cambridge, MA.

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model Selection in Linear Mixed Models. *Statistical Science*, **28**, 135-167.

Patterson H. D., & Thompson R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554.

Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model usefully especially for growth curve problems. *Biometrika*, ***51***, 313-326.

Rao, C. R. (1971a). Estimation of variance components-Minque theory, *Journal of Multivariate Analysis*, **1**, 257-275.

Rao, C. R. (1971b). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis*, **1**, 445-456.

Rao C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd edition. Wiley, New-York.

Rao C. R. (1979), MIQE theory and its relation to ML and MML estimation of variance components, *Sankhya B*, **41**, 138-153.

Rao, C. R., & Kleffe, J. (1988). *Estimation of variance components and applications*. North Holland series in statistics and probability, Elsevier, Amsterdam.

Schaeffer, L.R. 1986. Pseudo expectation approach to variance component estimation. *Journal of. Dairy Science*, **69**, 2884–2889.

Shaalje, G. B., McBride, J. J., and Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological and Environmental Statistics*, **7**, 512–524.

Searle, S. R. (1971). *Linear models*. Wiley, New-York.

Searle, S. R. (1979). *Notes on variance component estimation. A detailed account of maximum likelihood and kindred methodology*, Paper BU-673-M, Cornell University, Ithaca

Searle, S. R. (1982*). Matrix algebra useful for statistics*. Wiley, New York

Searle, S. R. (1989). Variance components-some history and a summary account of estimation methods. *Journal of Animal Breeding and Genetics*, **106**, 1-29.

Searle, S. R., Casella G., & Mc Culloch, C. E. (1992). *Variance components*, Wiley, New-York

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

Stram, D.O., & Lee J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171-1177.

Stram, D.O., & Lee J.W. (1995). Correction to "Variance components testing in the longitudinal mixed effects model", *Biometrics*, **51**, 1196.

Shaalje, G. B., McBride, J. J., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological and Environmental Statistics*, **7**, 512–524.

Shang, J., & Cavanaugh. J.E.(2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational. Statistics. and Data Analysis,* **52**, 2004–2021.

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statististics*, **8**, 1375-1381.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion for model fitting. *Suri-Kagaku*, **153**, 12-18.

Van Raden, P. M., & Yung, Y. C. (1988). A general purpose approximation to restricted maximum likelihood : the tilde-hat approach. Journal of Dairy Science, **71**, 187-194

Verbeke, G. & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer Verlag, New York.

Verbeke, G. & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, **59**, 254–262.

Welham, S.J., & Thompson, R. (1997). A likelihood ratio test for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society* B, **59**, 701-714

West, B., Welch, K. E, & Galecki, A. T. (2007). *Linear Mixed Models: a practical guide using statistical software*. ChapmanHall/CRC, Boca Raton.

Wolfinger, R. D. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation*, **22**, 1079-1106.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, **29**, 51-66

Zhang, D. & Lin X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In Dunson D. (Editor). *Random Effect and Latent Variable Model Selection*. Lecture Notes in Statistics, **192**, 19–36. Springer, New York.