REML Estimation and Linear Mixed Models

1. From linear models to REML

Sue Welham

Rothamsted Research
Harpenden UK AL5 2JQ

November 18, 2008

Introduction

- This course is going to look at various aspects of REML estimation in the linear mixed model - sometimes also called multi-level models, or hierarchical models
- The material will be a mixture of theory and application, with some focus on various types of data: what models to use, how to use them, limitations . . .
- My background: statistics in agriculture & plant sciences - methods widely applicable to medical statistics and data from the social sciences
- Acknowledgements: unpublished book by Cullis, Smith & Verbyla.
- Topics covered over 5 weeks:
  - The linear model $\longrightarrow$ linear mixed models & REML estimation
  - Simple models and analysis of longitudinal data
  - Spatial models & kriging
  - Penalized spline models
  - Analysis of multi-environment trials - a complex example
- Software: GenStat, http://www.vsni.co.uk/software/genstat-teaching/

The linear model

A general form for the linear model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{e} \tag{1}$$

where

- $\boldsymbol{y} = (y_1 \dots y_n)'$ is a vector of n data values
- $\boldsymbol{\tau}$ is a vector of p unknown fixed effects
- $\boldsymbol{X}$ is an $n \times p$ design matrix with value x_{ij} in row i, column j
- $\boldsymbol{e} = (e_1 \dots e_n)'$ is a vector of n random errors (deviations) and it is assumed that

  $$\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

  so the errors are independent, identically distributed normal random variables.

The aim of a statistical analysis is to estimate $\boldsymbol{\tau}$ together with some measure of uncertainty on the estimate in order to make predictions, again with some measure of uncertainty.

The design matrix

Explanatory variables can be defined as

- quantitative variables (sometimes called variates)
  - the variable has numeric values to be related to the response
  - eg. height, weight
  - the design matrix has a single column ($p = 1$) containing the values of the variable
  - this defines a linear relationship between the variable and the response (often called linear regression)
- qualitative variables (often called factors or dummy variables)
  - define a set of g groups ($g > 1$, $p = g$), such that each unit falls into one group
  - eg. variety, colour, soil type
  - the design matrix has one column for each group, and column j takes value 1 if the ith unit ($i = 1 \ldots n$) is in group j and zero otherwise
  - the sum of the columns is then $\mathbf{1}_n$, a vector with value 1 in all units
  - this defines a model where each group can have a different mean value

The design matrix (2)

Consider the simple case of a single explanatory variable.

The design matrix usually also contains a column to represent an overall constant, i.e. $\mathbf{1}_n$ with value 1 for every unit.

In a model with a single quantitative variable (variate), , then $p = 2$ and the constant term represents the intercept, i.e. the response value when explanatory variable x=0.

In a model with a single qualitative variable (factor), then $p = g + 1$ and more care is required:

- the design matrix $\boldsymbol{X}$ is no longer of full rank as, for $i = 1 \ldots n$,

$$\sum_{j=2}^{p} x_{ij} = 1 = x_{i1}$$

  i.e. there are $g + 1$ parameters to describe g group means.

- to make parameters identifiable, some constraint is required
- interpretation of the constant depends on the constraint (return to this later)

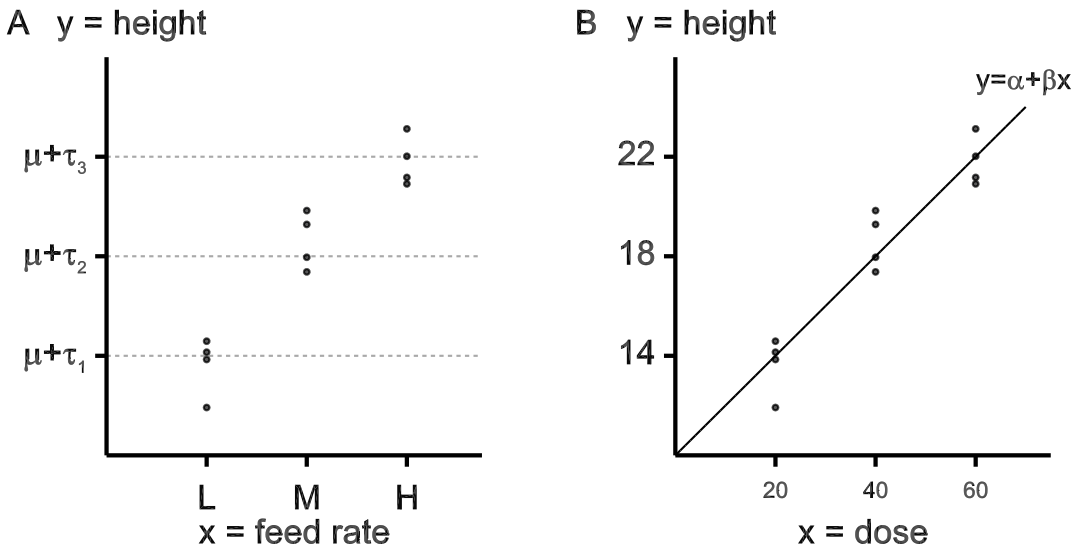
Example

Consider an experiment to investigate plant growth (height) in response to three doses of liquid feed (Low=20ml, Medium=40ml or High=60ml).

The explanatory variable can be considered as:

A) qualitative variable (factor): to look at mean response at each feed rate

B) quantitative variable (variate): to look at linear response to dose



Example (2)

For data ordered by replicates within groups:

A) qualitative variable

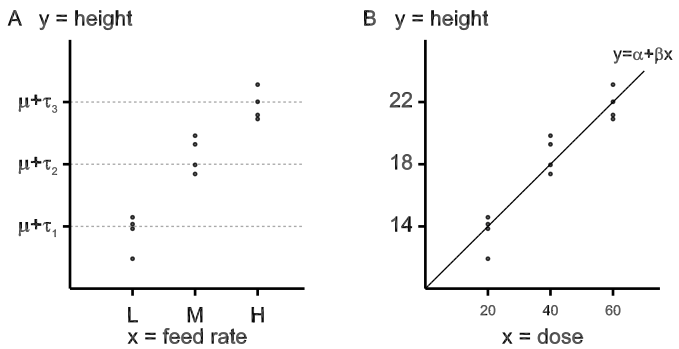
$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix};$$

B) quantitative variable

$$\boldsymbol{X} = \begin{bmatrix} 1 & 20 \\ 1 & 20 \\ 1 & 20 \\ 1 & 20 \\ 1 & 40 \\ 1 & 40 \\ 1 & 40 \\ 1 & 40 \\ 1 & 60 \\ 1 & 60 \\ 1 & 60 \\ 1 & 60 \end{bmatrix}$$

Example (3)

- Model with qualititative explanatory variable (factor):
  - more parameters - one for each group
  - no interpolation
  - more general model - fewer assumptions
- Model with quantitative explanatory variable (variate):
  - parsimonious (few parameters)
  - can interpolate to intermediate points of explanatory variable
  - goodness of fit depends on linearity of response



Estimation

The likelihood function for (1) is given by

$$\begin{aligned} L(\boldsymbol{\tau}, \sigma^2; \boldsymbol{y}) &= \prod_{i=1}^{n} f(y_i; \boldsymbol{x}'_{[i]}\boldsymbol{\tau}, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}'_{[i]}\boldsymbol{\tau}\right)^2\right) \end{aligned}$$

where

- $\boldsymbol{x}'_{[i]}$ is the ith row of $\boldsymbol{X}$
- $f(y; \mu, \sigma^2)$ is the probability density function for a normal random variable y with mean μ and variance σ^2

The log-likelihood function is then

$$\ell(\boldsymbol{\tau}, \sigma^2; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}'_{[i]}\boldsymbol{\tau}\right)^2$$

and the maximum likelihood estimates of $\boldsymbol{\tau}$ and σ^2 (denoted $\hat{\boldsymbol{\tau}}_{\text{ML}}$ and $\hat{\sigma}^2_{\text{ML}}$) are those values that maximise L or, equivalently, ℓ.

Estimation (2)

Note that the ordinary least squares estimator of $\boldsymbol{\tau}$ is the value that minimises the sum of squares of the residuals, i.e. the difference between the data (y_i) and its estimated value $(\boldsymbol{x}'_{[i]}\hat{\boldsymbol{\tau}}_{\mathrm{ML}})$:

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}'_{[i]}\hat{\boldsymbol{\tau}} \right)^2$$

This clearly yields the same estimate as the maximum likelihood method.

It follows from the Gauss-Markov theorem that this estimator of $\boldsymbol{\tau}$ is the BLUE, i.e. the

- Best (minimum variance)
- Linear
- Unbiased
- Estimator.

Estimation (3)

Maximisation of the log-likelihood function is achieved by differentiating with respect to each parameter and setting the derivative equal to zero, to give estimating equations for $\hat{\boldsymbol{\tau}}_{\text{ML}}$:

$$\frac{1}{\hat{\sigma}^2_{\text{ML}}} \sum_{i=1}^{n} \boldsymbol{x}_{[i]} \left(y_i - \boldsymbol{x}'_{[i]} \hat{\boldsymbol{\tau}}_{\text{ML}} \right) = \mathbf{0}$$

$$\Rightarrow \quad \sum_{i=1}^{n} \boldsymbol{x}_{[i]} \boldsymbol{x}'_{[i]} \hat{\boldsymbol{\tau}}_{\text{ML}} = \sum_{i=1}^{n} \boldsymbol{x}_{[i]} y_i$$

and for $\hat{\sigma}^2_{\text{ML}}$

$$-\frac{n}{2\hat{\sigma}^2_{\text{ML}}} + \frac{1}{2\hat{\sigma}^4_{\text{ML}}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}'_{[i]} \hat{\boldsymbol{\tau}}_{\text{ML}} \right)^2 = \mathbf{0}$$

$$\Rightarrow \quad \hat{\sigma}^2_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}'_{[i]} \hat{\boldsymbol{\tau}}_{\text{ML}} \right)^2$$

using

$$\frac{\partial \boldsymbol{a}' \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

Estimation (4)

In matrix form, the log-likelihood function can be written as

$$\ell(\boldsymbol{\tau}, \sigma^2; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau})$$

with resulting estimation equations for for $\hat{\boldsymbol{\tau}}_{\text{ML}}$:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}} = \boldsymbol{X}'\boldsymbol{y}$$

and for $\hat{\sigma}^2_{\text{ML}}$

$$\hat{\sigma}^2_{\text{ML}} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})/n$$

It is straightforward to show that these forms are equivalent to those on the previous slide.

Estimation of $\hat{\tau}_{\text{ml}}$

For $\boldsymbol{X}$ of full column rank:

$$\hat{\boldsymbol{\tau}}_{\text{ML}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

For $\boldsymbol{X}$ not of full column rank:

$$\hat{\boldsymbol{\tau}}_{\text{ML}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{A}^{-}$ is a generalized inverse of $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} = \boldsymbol{A}$.

- The estimator $\hat{\boldsymbol{\tau}}_{\text{ML}}$ is then not unique and depends on the generalized inverse used.
- The fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$ are unique, since $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'$ is invariant to the generalized inverse used (Searle *et al* 1992, Appendix M4).

Estimation of $\hat{\tau}_{ml}$ (2)

$\boldsymbol{X}$ may be rank deficient for several reasons:

1. presence of constant with one or more qualititative explanatory variables in the model
2. collinearity between quantitative explanatory variables
3. structural aliasing - eg. no data present for one (or more) groups for a qualitative variable

Reasons (2) and (3) (in general) cannot be predicted in advance of forming $\boldsymbol{X}$.

Reason (1) can be predicted from the structure of the model and explicit constraints can often be imposed to keep $\boldsymbol{X}$ full rank.

Common constraints used in statistical packages:

- sum-to-zero constraints
- corner-point constraints

Estimation of $\hat{\tau}_{\text{ml}}$ (3)

Consider the model containing a constant and single qualitative variable with g groups, each with r replicates, with data ordered by replicates within groups.

$$\boldsymbol{y} = c\mathbf{1} + (\boldsymbol{I}_g \otimes \mathbf{1}_r)\boldsymbol{a} + \boldsymbol{e}$$

for $\boldsymbol{a} = (a_1 \dots a_g)'$, so $\boldsymbol{\tau} = (c \;\; a_1 \dots a_g)'$

Then $\boldsymbol{X} = [\mathbf{1}_n \mid \boldsymbol{I}_g \otimes \mathbf{1}_r]$ is not full rank.

- Sum-to-zero constraints: $\sum_{i=1}^{g} a_i = 0$ so $a_g = -\sum_{i=1}^{g-1} a_i$.
  Hence

$$\boldsymbol{X}^* = \begin{bmatrix} \mathbf{1}_{r(g-1)} & \boldsymbol{I}_{g-1} \otimes \mathbf{1}_r \\ \mathbf{1}_r & -\boldsymbol{J}_{r,(g-1)} \end{bmatrix}$$

  is a full-rank version of $\boldsymbol{X}$ which defines the same model, where $\boldsymbol{J}_{a,b}$ is an $a \times b$ matrix with value 1 everywhere.

  In this parameterization, the constant estimates the grand (overall) mean and the $g-1$ effects associated with the groups 1:g-1 estimate deviations from the overall mean for each group.

Estimation of $\hat{\tau}_{\text{ml}}$ (4)

- Corner-point constraints (zero first level): $a_1 = 0$. Then we can use

$$\boldsymbol{X}^{**} = \begin{bmatrix} \mathbf{1}_r & \mathbf{0}_{r,(g-1)} \\ \mathbf{1}_{r(g-1)} & \boldsymbol{I}_{(g-1)} \otimes \mathbf{1}_r \end{bmatrix}$$

  is a different full-rank version of $\boldsymbol{X}$ which still defines the same model.

  In this parameterization, the constant estimates the treatment mean for the first group, and the $g-1$ effects associated with groups $2{:}g$ estimate deviations of each group with respect to the first group.

Sum-to-zero constraints are easier to interpret in simple models.

Corner-point constraints are often used in statistical packages because they are easy to implement in complex models and give sparser design matrices.

In both cases, interpretation of individual parameters becomes difficult in more complex models.

An alternative to constraints is to work with the full matrix $\boldsymbol{X}$ and deal only with estimable functions of parameters.

Estimability

Searle (1971) defines a linear combination $\boldsymbol{D\hat{\tau}}$ of parameter estimates $\hat{\boldsymbol{\tau}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$ to be estimable if

$$E(\boldsymbol{D\hat{\tau}}) = \boldsymbol{D\tau}.$$

This implies that the function $\boldsymbol{D\hat{\tau}}$ is invariant to the parameterization chosen.

It is straightforward to derive conditions to assess whether a function is estimable:

First, reorder the columns of $\boldsymbol{X}$ (and parameter vector $\boldsymbol{\tau}$) so that $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ where $\boldsymbol{X}_2$ is a maximal set of linearly independent columns with $\boldsymbol{\tau} = [\boldsymbol{\tau}_1\ \boldsymbol{\tau}_2]$ ordered conformally. Then

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{21}' \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \qquad \text{with} \qquad (\boldsymbol{X}'\boldsymbol{X})^{-} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{bmatrix}$$

where $\boldsymbol{A}_{22} = \boldsymbol{X}_2'\boldsymbol{X}_2$ is a square invertible matrix. This generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$ yields a (non-unique) ML estimate

$$\begin{aligned} \hat{\boldsymbol{\tau}}_0 &= \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1' & \boldsymbol{X}_2' \end{bmatrix} \boldsymbol{y} \\ &= \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{A}_{22}^{-1}\boldsymbol{X}_2'\boldsymbol{y} \end{pmatrix} \end{aligned}$$

Estimability (2)

Consider

$$
\begin{aligned}
E(\boldsymbol{D}\hat{\boldsymbol{\tau}}_0) &= [\boldsymbol{D}_1\ \boldsymbol{D}_2]\begin{pmatrix}\mathbf{0}\\ \boldsymbol{A}_{22}^{-1}\boldsymbol{X}_2' E(\boldsymbol{y})\end{pmatrix}\\
&= [\boldsymbol{D}_1\ \boldsymbol{D}_2]\begin{pmatrix}\mathbf{0}\\ \boldsymbol{A}_{22}^{-1}\boldsymbol{X}_2'\begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}\begin{pmatrix}\boldsymbol{\tau}_1\\ \boldsymbol{\tau}_2\end{pmatrix}\end{pmatrix}\\
&= [\boldsymbol{D}_1\ \boldsymbol{D}_2]\begin{pmatrix}\mathbf{0}\\ \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{\tau}_1+\boldsymbol{\tau}_2\end{pmatrix}\\
&= \boldsymbol{D}_2\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{\tau}_1+\boldsymbol{D}_2\boldsymbol{\tau}_2.
\end{aligned}
$$

Since $\boldsymbol{D}\boldsymbol{\tau}=\boldsymbol{D}_1\boldsymbol{\tau}_1+\boldsymbol{D}_2\boldsymbol{\tau}_2$, estimability is achieved when

$$\boldsymbol{D}_1-\boldsymbol{D}_2\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}=\mathbf{0}.$$

Estimability (3)

For $\boldsymbol{D} = \boldsymbol{X}$, we use $\boldsymbol{X}_1 = \boldsymbol{X}_2\boldsymbol{M}$ (from the definition of the partition) to see that

$$\begin{aligned}\boldsymbol{D}_1 - \boldsymbol{D}_2\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} &= \boldsymbol{X}_1 - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1 \\ &= \boldsymbol{X}_2\boldsymbol{M} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{M} \\ &= \boldsymbol{0}.\end{aligned}$$

so the fitted values are estimable, as shown earlier.

For $\boldsymbol{D} = \boldsymbol{I}$:

$$\begin{aligned}\boldsymbol{D}_1 - \boldsymbol{D}_2\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} &= \begin{pmatrix}\boldsymbol{I}\\ \boldsymbol{0}\end{pmatrix} - \begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{I}\end{pmatrix}(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{M} \\ &= \begin{pmatrix}\boldsymbol{I}\\ -\boldsymbol{M}\end{pmatrix} \neq \boldsymbol{0}.\end{aligned}$$

so the parameter vector is not estimable, as expected.

Estimation of $\hat{\sigma}^2_{\text{ml}}$

Note that $\hat{\sigma}^2_{\text{ML}}$ is a biased estimate of σ^2:

$$\begin{aligned} E(\hat{\sigma}^2_{\text{ML}}) &= E\left\{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})/n\right\} \\ &= \frac{1}{n} E\left\{\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^- \boldsymbol{X}')\boldsymbol{y}\right\} \\ &= \frac{1}{n}\sigma^2 tr(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^- \boldsymbol{X}') + \frac{1}{n}\boldsymbol{\tau}'\boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^- \boldsymbol{X}')\boldsymbol{X}\boldsymbol{\tau} \\ &= \frac{n - r_X}{n}\sigma^2 \end{aligned}$$

where $r_X = \text{rank}(\boldsymbol{X})$ and using

- $E(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}) = \text{trace}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$ for $\boldsymbol{y}$ with expectation $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{V}$
- $\text{trace}[\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^- \boldsymbol{A}'] = \text{rank}(\boldsymbol{A})$
- see eg. Searle *et al* (1992), Appendix S5.

Fitted values and residuals

- Fitted values: $\hat{\boldsymbol{y}}_{\text{ML}} = \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}_X\boldsymbol{y}$
- $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'$ is called the hat matrix - a projection into the column space of $\boldsymbol{X}$
- Residuals: $\hat{\boldsymbol{e}}_{\text{ML}} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{\text{ML}} = (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y} = \boldsymbol{P}_{X^\perp}\boldsymbol{y}$
- $\boldsymbol{P}_{X^\perp}$ is a projection out of the column space of $\boldsymbol{X}$

The fitted values and residuals are statistically independent:

$$\text{cov}(\hat{\boldsymbol{y}}_{\text{ML}}, \hat{\boldsymbol{e}}_{\text{ML}}) = \text{cov}(\boldsymbol{P}_X\boldsymbol{y}, \boldsymbol{P}_{X^\perp}\boldsymbol{y}) = \sigma^2\boldsymbol{P}_X\boldsymbol{P}_{X^\perp} = \boldsymbol{0}$$

Analysis of variance

The linear model is often assessed by ANOVA:

Source	SS	DF	MS	Ratio
Treatments	TSS=$\boldsymbol{y}'(\boldsymbol{P}_X - \frac{1}{n}\mathbf{1}\mathbf{1}')\boldsymbol{y}$	$r_X - 1$	TMS=$\frac{1}{r_X - 1}$TSS	TMS/RMS
Residual	RSS=$\boldsymbol{y}'\boldsymbol{P}_{X^\perp}\boldsymbol{y}$	$n - r_X$	RMS=$\frac{1}{n - r_X}$RSS	
Total	$\boldsymbol{y}'(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\boldsymbol{y}$			

Note that the overall mean is usually removed from the variation prior to the ANOVA, hence the adjustment using $\frac{1}{n}\mathbf{1}\mathbf{1}'$ in the treatment and total sums of squares.

- Result: for $y \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, then $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} \sim \chi^2(r_X, \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu})$ if $\boldsymbol{A}\boldsymbol{V}$ is idempotent.
- Result: for $y \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, then $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{y}'\boldsymbol{B}\boldsymbol{y}$ are independent if $\boldsymbol{A}\boldsymbol{V}\boldsymbol{B} = \mathbf{0}$.

Using these results we can show that, under the null hypothesis $\boldsymbol{X}\boldsymbol{\tau} = \mu\mathbf{1}$, the ratio TMS/RMS has an F-distribution on $(r_X - 1), (n - r_X)$ degrees of freedom.

ANOVA (2)

The sums of squares can alternatively be written as:

- TSS$=\hat{\boldsymbol{\tau}}'_{\text{ML}}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}} - \frac{1}{n}\boldsymbol{y}'\mathbf{1}\mathbf{1}'\boldsymbol{y}$ - an estimate of variation due to the fitted model (excluding overall mean)
- RSS $= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})$ - an estimate of residual - or background - variation

The variance ratio assesses whether variation due to the fitted model could plausibly be accounted for by background variation.

The residual mean square can also be used as an unbiased estimate of σ^2, ie

$$\hat{\sigma}^2_{\text{A}} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})/(n - r_X)$$

This estimate is usually used in preference to the maximum likelihood estimate because it gives 'more realistic' estimates of error to use in SEs/SEDs/CIs.

Linear models with several terms

- In general, a linear model may contain several (b) terms, each related to quantitative or qualitative explanatory variables.
- The design matrix is then partitioned into b terms as $\boldsymbol{X} = [\boldsymbol{X}_1 \ \ldots \boldsymbol{X}_b]$ where $\boldsymbol{X}_i$ is the $(n \times p_i)$ design matrix for term i, with $\sum_i p_i = p$.
- A term corresponding to a single explanatory variable is a 'main effect'.
- Combinations of explanatory variables are called interactions. The interaction of term A with design matrix $\boldsymbol{X}_A = [\boldsymbol{x}_{Aij}]$ for $i = 1 \ldots n, j = 1 \ldots p_A$ and term B with design matrix $\boldsymbol{X}_B$ has a design matrix with ith row

$$\left[\ \boldsymbol{x}_{Ai1}\boldsymbol{X}_B \ \ldots \ \boldsymbol{x}_{Aip_A}\boldsymbol{X}_B \ \right].$$

- Different types of explanatory variables combine as follows:
  - Two qualitative variables (factors) with p_A and p_B groups: forms a new set of $p_A p_B$ groups and fit a mean for each
  - Qualitative + quantitative variable: fit a linear trend for the quantitative variable within each group of the factor separately
  - Two quantitative variables: form a new variable by element-wise multiplication and fit a linear trend for the new variable

Symbolic representation of models

Within statistical packages, each variable is given a name, and the variable names used to define the model.

We will use the following notation for model formulae:

- A = qualitative variable (factor)
- x = quantitative variable (variate)
- A.B = interaction of terms A and B
- A*B = A + B + A.B = main effects and interaction
- A/B = A + A.B = term B nested within term A
- lin(A) = variate made from the levels (groups) of A
- fac(x) = factor made from the distinct values of x

Designed experiments

The motivation for the REML method originally arose from comparison of (unbiased) ANOVA and (biased) ML estimates of variance parameters in designed experiments.

Designed experiments use blocking to account for heterogeneity:

- similar units are grouped together to form blocks
- treatments are compared within blocks (as far as possible)
- variation due to blocks can be separated from background variation (residual error) and increase precision of estimates
- examples of blocking factors:
  - areas of similar fertility in a field trial
  - shelves in controlled environment cabinet
  - time of day - for processing plant samples
  - observer/experimenter - especially for subjective judgements
- it is usually assumed that there is no interaction between blocks and treatments

A mixed model

Designed experiments therefore usually have a set of terms relating to treatments ($\boldsymbol{X\tau}$) and a set of terms relating to blocks ($\boldsymbol{Zu}$):

- blocks are usually assumed to comprise a random sample for a (often notional) wider population.
- it is then natural to assume that block effects are realizations of a random variable - random effects

In the simple case, with one blocking factor, the model is written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{e} \tag{2}$$

where

- $\boldsymbol{y} = (y_1 \dots y_n)'$ is a vector of data
- $\boldsymbol{\tau}$ represent p fixed, unknown treatment effects with $(n \times p)$ design matrix $\boldsymbol{X}$
- $\boldsymbol{u}$ represent q random block effects with $(n \times q)$ design matrix $\boldsymbol{Z}$ and $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_b^2 \boldsymbol{I})$ - independent effects with equal variance.
- $\boldsymbol{e} = (e_1 \dots e_n)'$ is a vector of residual errors (noise) with $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ - independent effects with equal variance.

A mixed model (2)

Mixed model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

In the general case, both the fixed and random effects can be partitioned into terms associated with explanatory variables (in this context usually qualitative):

- $\boldsymbol{X} = [\, \boldsymbol{X}_1 \; \boldsymbol{X}_2 \ldots \boldsymbol{X}_b \,]$
- where $\boldsymbol{X}_i$ is an $n \times p_i$ design matrix for the ith fixed term, $\sum_i p_i = p$
- $\boldsymbol{Z} = [\, \boldsymbol{Z}_1 \; \boldsymbol{Z}_2 \ldots \boldsymbol{Z}_c \,]$
- where $\boldsymbol{Z}_j$ is an $n \times q_j$ design matrix for the jth random term, $\sum_j q_j = q$
- $\boldsymbol{\tau}, \boldsymbol{u}$ are partitioned conformally
  - $\boldsymbol{\tau} = (\boldsymbol{\tau}_1' \; \ldots \boldsymbol{\tau}_b)'$
  - $\boldsymbol{u} = (\boldsymbol{u}_1' \; \ldots \boldsymbol{u}_c)'$
  - with $\boldsymbol{u}_i \sim N(\boldsymbol{0}_{q_i}, \sigma_i^2 \boldsymbol{I}_{q_i})$ and $\text{cov}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \boldsymbol{0}$
  - independent random effects, with common variances within terms - simple variance components model

A mixed model (3)

Mixed model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

The variance matrix of the data takes the form

$$\begin{aligned}
\text{var}(\boldsymbol{y}) &= \boldsymbol{Z}\text{var}(\boldsymbol{u})\boldsymbol{Z}' + \text{var}(\boldsymbol{e}) \\
&= \sum_{i=1}^{c} \sigma_i^2 \boldsymbol{Z}_i \boldsymbol{Z}_i' + \sigma^2 \boldsymbol{I}_n \\
&= \boldsymbol{V} \\
&= \sigma^2 \left(\sum_{i=1}^{c} \gamma_i \boldsymbol{Z}_i \boldsymbol{Z}_i' + \boldsymbol{I}_n \right) \\
&= \sigma^2 \boldsymbol{H}
\end{aligned}$$

for $\gamma_i = \sigma_i^2/\sigma^2$.

This gives two alternative parameterizations of the variance model, in terms of $\boldsymbol{\sigma} = (\sigma_1^2 \ldots \sigma_c^2 \ \sigma^2)'$ (sigma parameterization) or in terms of $(\boldsymbol{\gamma}', \sigma^2)'$ for $\boldsymbol{\gamma} = (\gamma_1 \ldots \gamma_c)'$ (gamma parameterization).

We will usually work with the gamma parameterization.

Randomized complete block design

Balanced designed experiments have properties which simplify the estimation process for both treatments and variance parameters.

The simplest design with blocking is the randomized complete block design (RCBD).

For an experiment with g treatment groups and r replicates (blocks), each block consists of g units and contains one instance of each treatment, with treatments allocated to units within blocks at random.

The RCBD model for treatment j in block i can be written as

$$y_{ij} = b_i + \mu_j + e_{ij}$$

for $i = 1 \dots r$, $j = 1 \dots g$, $n = rg$.

This parameterization uses a single parameter for each treatment group so that the design matrix $\boldsymbol{X}$ is full rank.

RCBD (2)

The form of the previous model masks the randomization that takes place in the design and it is usually better to make this explicit as a model for unit k in block i as

$$y_{ik} = b_i + \mu_{s(ik)} + e_{ik}$$

for $i = 1 \dots r$, $k = 1 \dots g$, where $s(ik)$ indicates the treatment applied to this unit.

The advantage of the latter form is that it preserves the distinction between treatments and units. For simplicity, we will use the former notation.

If we order our data by blocks within treatment, ie
$\boldsymbol{y} = (y_{11}\ y_{21}\ \dots y_{r1}\ y_{12} \dots y_{rg})'$, then the matrix form of the model uses

- $\boldsymbol{X} = \boldsymbol{I}_g \otimes \mathbf{1}_r$
- $\boldsymbol{Z} = \mathbf{1}_g \otimes \boldsymbol{I}_r$
- $\boldsymbol{\tau} = (\mu_1 \dots \mu_g)'$
- $\boldsymbol{u} = (b_1 \dots b_r)'$

RCBD (3)

The variance matrix for the data then takes the form:

$$\begin{aligned} \text{var}(\boldsymbol{y}) = \boldsymbol{V} &= \boldsymbol{Z}\text{var}(\boldsymbol{u})\boldsymbol{Z}' + \text{var}(\boldsymbol{e}) \\ &= \sigma_b^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma^2 \boldsymbol{I}_n \\ &= \sigma_b^2 (\mathbf{1}_g \otimes \boldsymbol{I}_r)(\mathbf{1}_g \otimes \boldsymbol{I}_r)' + \sigma^2 \boldsymbol{I}_n \\ &= \sigma_b^2 (\mathbf{1}_g \mathbf{1}_g' \otimes \boldsymbol{I}_r) + \sigma^2 \boldsymbol{I}_n \end{aligned}$$

with

- $\text{var}(y_{ij}) = \sigma_b^2 + \sigma^2$
- $\text{cov}(y_{ij}, y_{ik}) = \sigma_b^2$ for $j \neq k$
- $\text{cov}(y_{ij}, y_{lk}) = 0$ for $i \neq l$

So data from units within the same block have a correlation of $\frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$, and units from different blocks are uncorrelated.

Strata

Within balanced designed experiments, the term strata is used to describe different levels within the blocking structure, such that items within the same level have the same variance.

Bailey (2008) gives a good overview of designed experiments and defines

- a stratum is an eigenspace of $\text{var}(\boldsymbol{y}) = \boldsymbol{V}$

In the RCBD, it is straightforward to verify that columns of $\boldsymbol{Z}$ form a set of eigenvectors for $\boldsymbol{V}$ with common eigenvalue $g\sigma_b^2 + \sigma^2$ and in fact these columns form an eigenspace (a maximal set of independent eigenvectors with a common eigenvalue).

Projection into this eigenspace can be achieved by the matrix $\boldsymbol{P}_Z = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' = \frac{1}{g}\boldsymbol{Z}\boldsymbol{Z}'$.

Similarly, projection into the orthogonal space can be achieved using $\boldsymbol{P}_{Z^\perp} = \boldsymbol{I}_n - \boldsymbol{P}_Z$.

We can therefore project onto the eigenspaces using orthogonal projection matrices that add up to the identity matrix.

It is conventional (arising from randomization theory) to partition out the 1-dimensional subspace corresponding to the mean, using the set $\boldsymbol{P}_0 = \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n'$ with $\boldsymbol{P}_1 = \boldsymbol{P}_Z - \boldsymbol{P}_0$ and $\boldsymbol{P}_2 = \boldsymbol{I} - \boldsymbol{P}_Z$.

Strata (2)

Note that we can write the variance matrix in terms of these projection matrices by equating coefficients, so

$$\begin{aligned} \boldsymbol{V} &= \xi_0 \boldsymbol{P}_0 + \xi_1 \boldsymbol{P}_1 + \xi_2 \boldsymbol{P}_2 \\ &= \xi_0 \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' + \xi_1 (\frac{1}{g} \boldsymbol{Z}\boldsymbol{Z}' - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n') + \xi_2 (\boldsymbol{I} - \frac{1}{g} \boldsymbol{Z}\boldsymbol{Z}') \\ &= \sigma_b^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma^2 \boldsymbol{I} \end{aligned}$$

where

- $\xi_0 = \xi_1 = g\sigma_b^2 + \sigma^2$
- $\xi_2 = \sigma^2$

Then $\boldsymbol{V}^{-1} = \frac{1}{\xi_0} \boldsymbol{P}_0 + \frac{1}{\xi_1} \boldsymbol{P}_1 + \frac{1}{\xi_2} \boldsymbol{P}_2$.

Given a set of orthogonal projection matrices $\boldsymbol{P}_i$, $i = 0, 1, 2$, we can write $\boldsymbol{P}_i = \boldsymbol{U}_i \boldsymbol{U}_i'$ where

- $\boldsymbol{U}_i$ is an $n \times \text{rank}(\boldsymbol{P}_i)$ matrix of full column rank
- $\boldsymbol{U}_i' \boldsymbol{U}_j = \boldsymbol{0}$ for $i \neq j$
- $\boldsymbol{U}_i' \boldsymbol{U}_i = \boldsymbol{I}$

Strata (3)

Then we can consider a transformation of the data as

$$\begin{pmatrix} \boldsymbol{U}_0'\boldsymbol{y} \\ \boldsymbol{U}_1'\boldsymbol{y} \\ \boldsymbol{U}_2'\boldsymbol{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{U}_0'\boldsymbol{X}\boldsymbol{\tau} \\ \boldsymbol{U}_1'\boldsymbol{X}\boldsymbol{\tau} \\ \boldsymbol{U}_2'\boldsymbol{X}\boldsymbol{\tau} \end{pmatrix}, \begin{pmatrix} \xi_0\boldsymbol{I}_1 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \xi_1\boldsymbol{I}_{r(\boldsymbol{P}_1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \xi_2\boldsymbol{I}_{r(\boldsymbol{P}_2)} \end{pmatrix} \right)$$

using

- $\text{var}(\boldsymbol{U}_i'\boldsymbol{y}) = \boldsymbol{U}_i'\boldsymbol{V}\boldsymbol{U}_i = \xi_i\boldsymbol{I}_{r(\boldsymbol{P}_i)}$
- $\text{cov}(\boldsymbol{U}_i'\boldsymbol{y}, \boldsymbol{U}_j'\boldsymbol{y}) = \boldsymbol{U}_i'\boldsymbol{V}\boldsymbol{U}_j = \boldsymbol{0}$

The transformed data has log-likelihood function:

$$\ell = \sum_{i=0,1,2} \left\{ -\frac{r(\boldsymbol{P}_i)}{2}\log(2\pi\xi_i) - \frac{1}{2\xi_i}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})'\boldsymbol{P}_i(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau}) \right\}$$

for $r(\boldsymbol{P}_i) = \text{rank}(\boldsymbol{P}_i)$.

Strata (4)

This form of the log-likelihood function yields estimating equations:

$$\sum_{i=0,1,2} \frac{1}{\xi_i} \left\{ \boldsymbol{X}' \boldsymbol{P}_i \boldsymbol{X} \hat{\boldsymbol{\tau}}_{\text{ML}} - \boldsymbol{X}' \boldsymbol{P}_i \boldsymbol{y} \right\} = \boldsymbol{0}$$

$$-\frac{r(\boldsymbol{P}_i)}{2\xi_i} + \frac{1}{2\xi_i^2} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})' \boldsymbol{P}_i (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}}) = \boldsymbol{0}$$

hence immediately

$$\hat{\xi}_{i\text{ML}} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}})' \boldsymbol{P}_i (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}}) / r(\boldsymbol{P}_i)$$

but this requires evaluation of $\hat{\boldsymbol{\tau}}_{\text{ML}}$, which does not appear to be calculated immediately from this form.

However, the structure of balanced designs can be used to rearrange these equations to get estimates of $\hat{\boldsymbol{\tau}}_{\text{ML}}$ directly.

Strata (5)

We partition the set of treatment means into an overall mean, main effects and interactions using another set of orthogonal projections, $\boldsymbol{T}_j$, so $\boldsymbol{\tau} = \sum_{j=1}^{l} \boldsymbol{T}_j \boldsymbol{\tau}$ where

- $\sum_{j=1}^{l} \boldsymbol{T}_j = \boldsymbol{I}$
- $\boldsymbol{T}_i \boldsymbol{T}_j = \boldsymbol{0}$ for $i \neq j$
- $\boldsymbol{T}_i \boldsymbol{T}_i = \boldsymbol{T}_i$ for $i = 1 \ldots l$.

This gives a set of interpretable effects in the case of balanced factorial treatment structures.

A design then has the property of *general balance* if

$$\boldsymbol{X}' \boldsymbol{P}_i \boldsymbol{X} = \sum_{j=1}^{l} \lambda_{ij} \boldsymbol{T}_j$$

and λ_{ij} is then the effective replication of treatment j in stratum i.

A design is orthogonal if, for each treatment j, $\lambda_{ij} \neq 0$ for only one value of i (stratum).

Strata (6)

In the case of the RCBD with a single treatment factor, $l = 2$ and projections correspond to the overall mean:

$$\boldsymbol{T}_1 = \frac{1}{g}\mathbf{1}_g\mathbf{1}_g'$$

and deviations of treatment groups from the mean:

$$\boldsymbol{T}_2 = \boldsymbol{I}_g - \boldsymbol{T}_1 = \boldsymbol{I}_g - \frac{1}{g}\mathbf{1}_g\mathbf{1}_g'$$

Hence

$$\boldsymbol{T}_1\boldsymbol{\tau} = \mu\mathbf{1}_g$$

for $\mu = \frac{1}{g}\sum_{i=1}^{g}\mu_i$, and

$$\boldsymbol{T}_2\boldsymbol{\tau} = \begin{pmatrix} \mu_1 - \mu \\ \vdots \\ \mu_g - \mu \end{pmatrix}$$

Then:

$$\lambda_{01} = r \qquad \lambda_{11} = 0 \qquad \lambda_{21} = 0$$
$$\lambda_{02} = 0 \qquad \lambda_{12} = 0 \qquad \lambda_{22} = r$$

Strata (7)

Consider estimation of $\boldsymbol{T}_k\boldsymbol{\tau}$ from each stratum separately, for a balanced design, using estimating equation:

$$\boldsymbol{X}\boldsymbol{P}_i\boldsymbol{X}\hat{\boldsymbol{\tau}}_{\text{ML}} = \boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y}$$

$$\Rightarrow \quad \sum_{j=1}^{l} \lambda_{ij}\boldsymbol{T}_j\hat{\boldsymbol{\tau}}_{\text{ML}} = \boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y}$$

(pre-multiply by $\boldsymbol{T}_k$)

$$\Rightarrow \quad \lambda_{ik}\boldsymbol{T}_k\hat{\boldsymbol{\tau}}_{\text{ML}} = \boldsymbol{T}_k\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y}$$

$$\Rightarrow \quad \boldsymbol{T}_k\hat{\boldsymbol{\tau}}_{\text{ML}} = \frac{1}{\lambda_{ik}}\boldsymbol{T}_k\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y}$$

then use

$$\hat{\boldsymbol{\tau}}_{\text{ML}} = \sum_{j=1}^{l} \boldsymbol{T}_j\hat{\boldsymbol{\tau}}_{\text{ML}}$$

For an orthogonal design, for treatment k, λ_{ik} is non-zero in only one strata, so this process gives a single unique estimate of treatment $\boldsymbol{T}_k\boldsymbol{\tau}$.

This gives an easy and computationally efficient algorithm for estimation of treatment effects that does not require inversion of $\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}$ and leads directly to an ANOVA decomposition in terms of sums of squares.

Multi-stratum ANOVA

A multi-stratum ANOVA partitions the sums of squares into strata defined by orthogonal projections $\boldsymbol{P}_i$ and partitions the treatments according to the orthogonal projections $\boldsymbol{T}_j$, with SS for treatment $\boldsymbol{T}_j\boldsymbol{\tau}$ appearing in strata where $\lambda_{ij} > 0$.

Stratum total sums of squares take the form $\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{y}$, with expected value:

$$\begin{aligned} E\left(\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{y}\right) &= \xi_i \mathsf{trace}(\boldsymbol{P}_i) + \sum_{j=1}^{l} \lambda_{ij}\boldsymbol{\tau}'\boldsymbol{T}_j\boldsymbol{\tau} \\ &= \xi_i \mathsf{rank}(\boldsymbol{P}_i) + \sum_{j=1}^{l} \lambda_{ij}\boldsymbol{\tau}'\boldsymbol{T}_j\boldsymbol{\tau} \end{aligned}$$

The sum of squares for treatment term $\boldsymbol{T}_k\boldsymbol{\tau}$ in stratum i is

$$\frac{1}{\lambda_{ik}}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T}_k\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{T}_k\hat{\boldsymbol{\tau}}_{\text{ML}})'\boldsymbol{P}_i(\boldsymbol{X}\boldsymbol{T}_k\hat{\boldsymbol{\tau}}_{\text{ML}})$$

with expected value

$$\begin{aligned} E\left(\frac{1}{\lambda_{ik}}\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T}_k\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y}\right) &= \frac{\xi_i}{\lambda_{ik}}\mathsf{tr}(\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T}_k\boldsymbol{X}'\boldsymbol{P}_i) + \frac{1}{\lambda_{ik}}\boldsymbol{\tau}'\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T}_k\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{\tau} \\ &= \xi_i \mathsf{rank}(\boldsymbol{T}_k) + \lambda_{ik}\boldsymbol{\tau}'\boldsymbol{T}_k\boldsymbol{\tau} \end{aligned}$$

Multi-stratum ANOVA (2)

ANOVA for the RCBD

Stratum Source	SS	DF	EMS
Mean stratum	$\boldsymbol{y}'\boldsymbol{P_0}\boldsymbol{y}$	1	$\xi_0 + n\mu'\mu$
Mean	$\boldsymbol{y}'\boldsymbol{P_0}\boldsymbol{y}$	1	$\xi_0 + n\mu'\mu$
Residual	-	0	-
Block stratum	$\boldsymbol{y}'\boldsymbol{P_1}\boldsymbol{y}$	r-1	$(r-1)\xi_1$
Residual	$\boldsymbol{y}'\boldsymbol{P_1}\boldsymbol{y}$	r-1	$(r-1)\xi_1$
Block.Units stratum	$\boldsymbol{y}'\boldsymbol{P_2}\boldsymbol{y}$	r(g-1)	$r(g-1)\xi_2 + r\boldsymbol{\tau}'\boldsymbol{T}_2\boldsymbol{\tau}$
Treatments	TSS	$g-1$	$(g-1)\xi_2 + r\boldsymbol{\tau}'\boldsymbol{T}_2\boldsymbol{\tau}$
Residual	$\boldsymbol{y}'\boldsymbol{P_2}\boldsymbol{y}$-TSS	$(r-1)(g-1)$	$(r-1)(g-1)\xi_2$
Total	$\boldsymbol{y}'\boldsymbol{y}$		

- TSS$=\frac{1}{r}\boldsymbol{y}'\boldsymbol{P}_2\boldsymbol{X}\boldsymbol{T_2}\boldsymbol{X}'\boldsymbol{P}_2\boldsymbol{y}$
- $\boldsymbol{\tau}'\boldsymbol{T}_2\boldsymbol{\tau} = (\boldsymbol{\mu} - \mu\mathbf{1})'(\boldsymbol{\mu} - \mu\mathbf{1})$
- The DF of a sum of squares is the rank of the associated projection matrix.
- Under the hypothesis $\boldsymbol{\tau} = \mu\mathbf{1}$ the variance ratio (treatment MS/residual MS) in the units stratum has an F distribution with $(g-1), (r-1)(g-1)$ DF.

Multi-stratum ANOVA (3)

The ANOVA estimates of the stratum variances are equal to the stratum residual mean squares:

- $\hat{\xi}_{1\mathrm{A}} = \frac{RSS_1}{r-1} = \frac{\boldsymbol{y}'\boldsymbol{P}_1\boldsymbol{y}}{r-1}$
- $\hat{\xi}_{2\mathrm{A}} = \frac{RSS_2}{(r-1)(g-1)} = \frac{\boldsymbol{y}'\boldsymbol{P}_2\boldsymbol{y}-TSS}{(r-1)(g-1)}$

In general the estimates take the form

$$\hat{\xi}_{i\mathrm{A}} = \mathrm{RSS}_i/[r(\boldsymbol{P}_i) - \sum_{j=1}^{l} I(\lambda_{ij} > 0) r(\boldsymbol{T}_j)].$$

Compared to the maximum likelihood estimates

$$\hat{\xi}_{i\mathrm{ML}} = \mathrm{RSS}_i/[r(\boldsymbol{P}_i)],$$

the ANOVA estimates adjust for the DF used in estimation of treatment effects to achieve unbiased estimates.

Unbiased estimates of variance components follow from:

- $\hat{\xi}_{1\mathrm{A}} = g\hat{\sigma}_b^2 + \hat{\sigma}^2, \qquad \hat{\xi}_{2\mathrm{A}} = \hat{\sigma}^2$

Unbalanced designs & REML

In unbalanced designs, this simple estimation process cannot be followed and more complex iterative algorithms are required, such as maximum likelihood estimation.

However, the downwards bias of ML estimates of stratum variances is undesirable as this means that estimates of treatment standard errors are also underestimated.

For this reason, Patterson & Thompson (1971) introduced the method of residual (or restricted) maximum likelihood (REML) which includes an adjustment for degrees of freedom used in estimating fixed effects from the general linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

as defined earlier (equation 2). Assume (make) $\boldsymbol{X}$ full rank.

Verbyla (1990) gives a clear interpretation of the REML method as follows: partition the likelihood into two independent parts: one ($\boldsymbol{y}_1 = \boldsymbol{L}_1'\boldsymbol{y}$) relating to the fixed effects and one part ($\boldsymbol{y}_2 = \boldsymbol{L}_2'\boldsymbol{y}$) relating to the residual contrasts (zero expectation) with

- $\boldsymbol{L}_1$ is an $n \times p$ matrix of full column rank
- $\boldsymbol{L}_2$ is an $n \times [n-p]$ matrix of full column rank
- $\boldsymbol{L}_1'\boldsymbol{X} = \boldsymbol{I}$, $\boldsymbol{L}_2'\boldsymbol{X} = \boldsymbol{0}$

REML

Then

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\tau} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_1 & \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_2 \\ \boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_1 & \boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2 \end{bmatrix} \right)$$

and we can use

$$\ell(\boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2; \boldsymbol{y}) = \ell(\boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2; \boldsymbol{y}_1|\boldsymbol{y}_2) + \ell(\boldsymbol{\gamma}, \sigma^2; \boldsymbol{y}_2).$$

The conditional distribution of $\boldsymbol{y}_1|\boldsymbol{y}_2$ has expected value

$$E(\boldsymbol{y}_1|\boldsymbol{y}_2) = \boldsymbol{\tau} + \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_2(\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{y}_2$$

and variance

$$\text{var}(\boldsymbol{y}_1|\boldsymbol{y}_2) = \sigma^2(\boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_1 - \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_2(\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_1).$$

Use result

- if $\boldsymbol{K}'\boldsymbol{X} = \mathbf{0}$ for $\boldsymbol{K}$ of full column rank and $\boldsymbol{H}$ is positive definite, then

$$\boldsymbol{K}(\boldsymbol{K}'\boldsymbol{H}\boldsymbol{K})^{-1}\boldsymbol{K}' = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{H}^{-1} = \boldsymbol{P}$$

to get

$$E(\boldsymbol{y}_1|\boldsymbol{y}_2) = \boldsymbol{\tau} + \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{P}\boldsymbol{y}\,; \quad \text{var}(\boldsymbol{y}_1|\boldsymbol{y}_2) = \sigma^2(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}.$$

REML (2)

Then

$$\begin{aligned}\ell(\boldsymbol{\tau},\boldsymbol{\gamma},\sigma^2;\boldsymbol{y}_1|\boldsymbol{y}_2) &= -\frac{1}{2}\left\{p\log(2\pi\sigma^2)+\log|(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}|\right.\\ &\quad\left.+(\boldsymbol{L}_1'(\boldsymbol{I}-\boldsymbol{H}\boldsymbol{P})\boldsymbol{y}-\boldsymbol{\tau})'\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{L}_1'(\boldsymbol{I}-\boldsymbol{H}\boldsymbol{P})\boldsymbol{y}-\boldsymbol{\tau})/\sigma^2\right\}\end{aligned}$$

and

$$\begin{aligned}\ell(\boldsymbol{\gamma},\sigma^2;\boldsymbol{y}_2) &= -\frac{1}{2}\left\{(n-p)\log(2\pi\sigma^2)+\log|\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2|+\boldsymbol{y}'\boldsymbol{L}_2(\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2')^{-1}\boldsymbol{L}_2\boldsymbol{y}\right\}\\ &= -\frac{1}{2}\left\{(n-p)\log(2\pi\sigma^2)+\log|\boldsymbol{L}'\boldsymbol{L}|+\log|\boldsymbol{H}|+\log|(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})|\right.\\ &\quad\left.+\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma^2\right\}\end{aligned}$$

Using the result on partitioned matrices

- for a non-singular matrix $\begin{vmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{vmatrix} = |\boldsymbol{D}||\boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{D}|$

to get

$$\log|\boldsymbol{L}'\boldsymbol{H}\boldsymbol{L}| = \log|\boldsymbol{L}'\boldsymbol{L}|+\log|\boldsymbol{H}| = \log|\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2|+\log|(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}|$$

Note that $\ell(\boldsymbol{\gamma},\sigma^2;\boldsymbol{y}_2)$ is not a function of $\boldsymbol{\tau}$.

Estimation

The fixed effects must therefore be estimated from $\ell_1 = \ell(\boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2; \boldsymbol{y}_1 | \boldsymbol{y}_2)$:

$$\begin{aligned} \frac{\partial \ell_1}{\partial \tau} &= \boldsymbol{X}' \boldsymbol{H}^{-1} \boldsymbol{X} (\boldsymbol{L}_1' (\boldsymbol{I} - \boldsymbol{H}\boldsymbol{P}) \boldsymbol{y} - \boldsymbol{\tau}) / \sigma^2 \\ &\Rightarrow \quad \hat{\boldsymbol{\tau}} = \boldsymbol{L}_1' (\boldsymbol{I} - \boldsymbol{H}\boldsymbol{P}) \boldsymbol{y} \\ &\Rightarrow \quad \hat{\boldsymbol{\tau}} = (\boldsymbol{X}' \boldsymbol{H}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{H}^{-1} \boldsymbol{y} \end{aligned}$$

Note that $\hat{\boldsymbol{\tau}}$ depends on the estimated variance parameters through $\boldsymbol{H}$.

For inference, $\text{var}(\hat{\boldsymbol{\tau}}) = \sigma^2 (\boldsymbol{X}' \boldsymbol{H}^{-1} \boldsymbol{X})^{-1}$.

Note that this can also be obtained as the generalized least squares estimate of $\boldsymbol{\tau}$ given $\boldsymbol{H}$, hence $\hat{\boldsymbol{\tau}}$ is a BLUE.

As both $\boldsymbol{y}_1$ and $\boldsymbol{\tau}$ are of length p, there is no further information in $\boldsymbol{y}_1$ and the variance parameters $\boldsymbol{\gamma}$ and σ^2 must be estimated from $\ell_2 = \ell(\boldsymbol{\gamma}, \sigma^2; \boldsymbol{y}_2)$.

Estimation (2)

We write

$$\begin{aligned}\ell_2 &= -\frac{1}{2}\left\{c(\boldsymbol{X}) + (n-p)\log(\sigma^2) + \log|\boldsymbol{H}| + \log|(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})| \right. \\ &\quad \left. + \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma^2\right\}\end{aligned}$$

where $c(\boldsymbol{X}) = (n-p)\log(2\pi) + \log|\boldsymbol{L}'\boldsymbol{L}|$ is a function of $\boldsymbol{X}$ (through $\boldsymbol{L}$).

Then

$$\frac{\partial \ell_2}{\partial \sigma^2} = -\frac{1}{2}\left\{\frac{(n-p)}{\sigma^2} - \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma^4\right\}$$

This gives the score equation

$$U_{\sigma^2}(\boldsymbol{\gamma}, \sigma^2) = -\frac{1}{2}\left\{\frac{(n-p)}{\sigma^2} - \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma^4\right\}$$

Setting this equal to zero yields the estimate:

$$\hat{\sigma}^2 = \boldsymbol{y}'\hat{\boldsymbol{P}}\boldsymbol{y}/(n-p)$$

So, given an estimate of $\boldsymbol{\gamma}$ and hence $\boldsymbol{P}$, we can obtain an estimate of $\boldsymbol{\sigma}^2$ directly.

Estimation (3)

For the gamma parameters, we need to use results on matrix differentiation:

- for a general matrix $\boldsymbol{A}(\theta)$

$$\frac{\partial \log |\boldsymbol{A}|}{\partial \theta} = \text{tr}\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \theta}\right)$$

- also

$$\frac{\partial \boldsymbol{A}^{-1}}{\partial \theta} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \theta}\boldsymbol{A}^{-1}$$

- and (eventually) for matrix $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}\boldsymbol{H}^{-1}$, with $\boldsymbol{H} = \boldsymbol{H}(\theta)$,

$$\frac{\partial \boldsymbol{P}}{\partial \theta} = -\boldsymbol{P}\frac{\partial \boldsymbol{H}}{\partial \theta}\boldsymbol{P}$$

Estimation (4)

Hence

$$\begin{aligned}
\frac{\partial \ell_2}{\partial \gamma_i} &= -\frac{1}{2}\left\{ \text{tr}\left(\boldsymbol{H}^{-1}\frac{\partial \boldsymbol{H}}{\partial \gamma_i}\right) - \text{tr}\left((\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\frac{\partial \boldsymbol{H}}{\partial \gamma_i}\boldsymbol{H}^{-1}\boldsymbol{X}\right)\right. \\
&\qquad \left. -\frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\frac{\partial \boldsymbol{H}}{\partial \gamma_i}\boldsymbol{P}\boldsymbol{y}\right\} \\
&= -\frac{1}{2}\left\{ \text{tr}\left(\boldsymbol{P}\frac{\partial \boldsymbol{H}}{\partial \gamma_i}\right) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\frac{\partial \boldsymbol{H}}{\partial \gamma_i}\boldsymbol{P}\boldsymbol{y}\right\} \\
&= -\frac{1}{2}\left\{ \text{tr}\left(\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{Z}_i\right) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{y}\right\}
\end{aligned}$$

The estimating equations (score equations) for $\boldsymbol{\gamma}$ are therefore

$$U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}, \sigma^2) = \text{tr}\left(\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{Z}_i\right) - \frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{y}$$

These equations are usually a complex non-linear function of $\boldsymbol{\gamma}$ (through $\boldsymbol{H}$ and $\boldsymbol{P}$) and cannot be solved (by setting equal to zero) directly. Note: balanced designs are a special case.

An iterative algorithm is therefore required.

ML vs REML comparison

Before getting onto iterative algorithms, it is helpful to review the difference between the log-likelihood function ℓ used to calculate maximum likelihood estimates, and that (ℓ_2) used for REML:

$$\begin{aligned} -2\ell(\boldsymbol{\tau}, \boldsymbol{\gamma}, \sigma^2; \boldsymbol{y}) &= c + n\log\sigma^2 + \log|\boldsymbol{H}| + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau})'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau})/\sigma^2 \\ -2\ell_2 = -2\ell(\boldsymbol{\gamma}, \sigma^2; \boldsymbol{y}_2) &= c(\boldsymbol{X}) + (n-p)\log(\sigma^2) + \log|\boldsymbol{H}| + \log|(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})| \\ &\quad + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}})'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}})/\sigma^2 \end{aligned}$$

using $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}})'\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}) = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}$ for $\hat{\boldsymbol{\tau}} = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}$ with $\boldsymbol{X}\hat{\boldsymbol{\tau}} = (\boldsymbol{I} - \boldsymbol{H}\boldsymbol{P})\boldsymbol{y}$ and $\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}$.

The term $\log|(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})|$ makes the adjustment for degrees of freedom used in estimating treatment effects, so that REML estimates of variance components are less biased than ML estimates.

The other major differences are:

- ℓ_2 is not a function of the fixed effects $\boldsymbol{\tau}$
- the constant in ℓ_2 is a function of the fixed design matrix $\boldsymbol{X}$

Iterative algorithms

Two main alternatives:

- EM (expectation-maximisation) algorithm - see eg Searle *et al* (1992)
- Newton-Raphson type algorithms - we will look briefly at these

Newton-Raphson algorithm can be derived from a Taylor expansion of the score equations U for a set of parameters $\boldsymbol{\theta}$ about a value $\hat{\boldsymbol{\theta}}$:

$$U(\boldsymbol{\theta}) \quad = \quad U(\hat{\boldsymbol{\theta}}) + \left.\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$$

We are aiming to get $U(\boldsymbol{\theta}) = \mathbf{0}$.

Given an initial estimate $\hat{\boldsymbol{\theta}}_{(0)}$, say, we can get an improved estimate, $\hat{\boldsymbol{\theta}}_{(1)}$, by solving for $U(\hat{\boldsymbol{\theta}}_{(1)}) = \mathbf{0}$ as

$$\hat{\boldsymbol{\theta}}_{(1)} = \hat{\boldsymbol{\theta}}_{(0)} - \left[\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{(0)}} U(\hat{\boldsymbol{\theta}}_{(0)})$$

Newton-Raphson algorithm

In our context, we can alternate between two equations over several iterations:

$$\begin{aligned}\hat{\sigma}^2_{(i+1)} &= \boldsymbol{y}'\hat{\boldsymbol{P}}_{(i)}\boldsymbol{y}/(n-p)\\ \hat{\boldsymbol{\gamma}}_{(i+1)} &= \hat{\boldsymbol{\gamma}}_{(i)} - \left[\frac{\partial U(\boldsymbol{\gamma},\sigma^2)}{\partial \boldsymbol{\gamma}}\right]^{-1}_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{(i)},\sigma^2=\hat{\sigma}^2_{(i+1)}} U(\hat{\boldsymbol{\gamma}}_{(i)},\hat{\sigma}^2_{(i+1)})\end{aligned}$$

where the subscript (i) indicates evaluation using estimates from iteration i.

Note that

$$\begin{aligned}\frac{\partial U(\boldsymbol{\gamma},\sigma^2)}{\partial \boldsymbol{\gamma}} &= \frac{\partial^2 \ell_2(\boldsymbol{\gamma},\sigma^2;\boldsymbol{y})}{(\partial\boldsymbol{\gamma})(\partial\boldsymbol{\gamma})'}\\ &= -\mathcal{I}^o_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\gamma},\sigma^2)\end{aligned}$$

where $\mathcal{I}^o_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ is known as the observed information matrix for gamma.

An alternative version of this algorithm updates $\boldsymbol{\gamma}$ and σ^2 using the joint information matrix for the two sets of parameters:

$$\mathcal{I}^o = \begin{bmatrix}\mathcal{I}^o_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & \mathcal{I}^o_{\boldsymbol{\gamma}\sigma^2}\\ \mathcal{I}^o_{\sigma^2\boldsymbol{\gamma}} & \mathcal{I}^o_{\sigma^2\sigma^2}\end{bmatrix}$$

Information matrices

The observed information matrix contains:

$$\begin{aligned}
\mathcal{I}^o_{\gamma_i\gamma_j} &= -\frac{\partial^2 \ell_2}{\partial\gamma_i\partial\gamma_j} &= \frac{1}{2}\left\{-\text{tr}(\boldsymbol{PZ}_i\boldsymbol{Z}_i'\boldsymbol{PZ}_j\boldsymbol{Z}_j') + \frac{2}{\sigma^2}\boldsymbol{y}'\boldsymbol{PZ}_i\boldsymbol{Z}_i'\boldsymbol{PZ}_j\boldsymbol{Z}_j'\boldsymbol{Py}\right\}\\
\mathcal{I}^o_{\gamma_i\sigma^2} &= -\frac{\partial^2 \ell_2}{\partial\gamma_i\partial\sigma^2} &= \frac{1}{2}\left\{\frac{1}{\sigma^4}\boldsymbol{y}'\boldsymbol{PZ}_i\boldsymbol{Z}_i'\boldsymbol{Py}\right\}\\
\mathcal{I}^o_{\sigma^2\sigma^2} &= -\frac{\partial^2 \ell_2}{\partial\gamma_i\partial\gamma_j} &= \frac{1}{2}\left\{-\frac{n-p}{\sigma^4} + \frac{2}{\sigma^6}\boldsymbol{y}'\boldsymbol{Py}\right\}
\end{aligned}$$

The most common variation on the Newton-Raphson algorithm uses the expected information matrix, $\mathcal{I}^e$, (Fisher scoring algorithm) with

$$\begin{aligned}
\mathcal{I}^e_{\gamma_i\gamma_j} &= E\left(-\frac{\partial^2 \ell_2}{\partial\gamma_i\partial\gamma_j}\right) &= \frac{1}{2}\left\{\text{tr}(\boldsymbol{PZ}_i\boldsymbol{Z}_i'\boldsymbol{PZ}_j\boldsymbol{Z}_j')\right\}\\
\mathcal{I}^e_{\gamma_i\sigma^2} &= E\left(-\frac{\partial^2 \ell_2}{\partial\gamma_i\partial\sigma^2}\right) &= \frac{1}{2}\left\{\frac{1}{\sigma^2}\text{tr}(\boldsymbol{PZ}_i\boldsymbol{Z}_i')\right\}\\
\mathcal{I}^e_{\sigma^2\sigma^2} &= E\left(-\frac{\partial^2 \ell_2}{\partial\gamma_i\partial\gamma_j}\right) &= \frac{1}{2}\left\{\frac{n-p}{\sigma^4}\right\}
\end{aligned}$$

This algorithm tends (not always) to be more stable - less influenced by data values in second derivatives.

Information matrices (2)

Terms of the type $\mathrm{tr}(\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{Z}_j\boldsymbol{Z}_j')$ can be computationally expensive to calculate.

Gilmour *et al* (1995) suggested use of an alternative 'average' information matrix consisting of sum of squares terms only that can be more easily calculated.

This 'average information matrix' is motivated as an average of the observed and expected information matrices but in fact omits some inconvenient terms to give:

$$\begin{aligned}
\mathcal{I}^a_{\gamma_i\gamma_j} &= \frac{1}{2}\left\{\frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{Z}_j\boldsymbol{Z}_j'\boldsymbol{P}\boldsymbol{y}\right\} \\
\mathcal{I}^a_{\gamma_i\sigma^2} &= \frac{1}{2}\left\{\frac{1}{\sigma^4}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{P}\boldsymbol{y}\right\} \\
\mathcal{I}^a_{\sigma^2\sigma^2} &= \frac{1}{2}\left\{\frac{1}{\sigma^6}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}\right\}
\end{aligned}$$

With an efficient algorithm, as in Gilmour *et al* (1995), this gives much faster computing time.

Estimation of random effects: BLUPs

Since the random effects $\boldsymbol{u}$ are not parameters of the model, it is conventional to talk of prediction rather than estimation of these effects. The predictors are usually denoted by $\tilde{\boldsymbol{u}}$ rather than $\hat{\boldsymbol{u}}$.

The criteria used for predictors of $\boldsymbol{u}$ are:

- minimum mean squared error $E[(\tilde{\boldsymbol{u}} - \boldsymbol{u})'(\tilde{\boldsymbol{u}} - \boldsymbol{u})]$ ('best')
- linear (in $\boldsymbol{y}_2$ since no information is available from $\boldsymbol{y}_1$)
- unbiased in the sense $E(\tilde{\boldsymbol{u}}) = E(\boldsymbol{u})$.

Hence BLUPs (best linear unbiased predictors).

A linear estimator must be of the form

$$\tilde{\boldsymbol{u}} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{y}_2$$

for some known vector $\boldsymbol{a}$ and matrix $\boldsymbol{B}$.

Then

$$E(\tilde{\boldsymbol{u}}) = E(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{y}_2) = \boldsymbol{a}$$

and $E(\boldsymbol{u}) = \boldsymbol{0}$ so $\boldsymbol{a} = 0$.

BLUPs (2)

Following Searle *et al* (1992), let $\boldsymbol{u}_0 = E(\boldsymbol{u}|\boldsymbol{y}_2)$ and consider the MSE:

$$\begin{aligned} E[(\tilde{\boldsymbol{u}} - \boldsymbol{u})'(\tilde{\boldsymbol{u}} - \boldsymbol{u})] &= E[(\tilde{\boldsymbol{u}} - \boldsymbol{u}_0 + \boldsymbol{u}_0 - \boldsymbol{u})'(\tilde{\boldsymbol{u}} - \boldsymbol{u}_0 + \boldsymbol{u}_0 - \boldsymbol{u})] \\ &= E[(\tilde{\boldsymbol{u}} - \boldsymbol{u}_0)'(\tilde{\boldsymbol{u}} - \boldsymbol{u}_0)] + 2E[(\tilde{\boldsymbol{u}} - \boldsymbol{u}_0)'(\boldsymbol{u}_0 - \boldsymbol{u})] \\ &\quad + E[(\boldsymbol{u}_0 - \boldsymbol{u})'(\boldsymbol{u}_0 - \boldsymbol{u})] \end{aligned}$$

We wish to find $\tilde{\boldsymbol{u}}$ to minimise this expression:

- the first term is minimised for $\tilde{\boldsymbol{u}} = \boldsymbol{u}_0$
- the second term is zero, since $\tilde{\boldsymbol{u}} = \boldsymbol{B}\boldsymbol{y}_2$ is fixed given $\boldsymbol{y}_2$ and using

  $$E[(\tilde{\boldsymbol{u}} - \boldsymbol{u}_0)'(\boldsymbol{u}_0 - \boldsymbol{u})] = E_{y_2}\{E_u[(\tilde{\boldsymbol{u}} - \boldsymbol{u}_0)'(\boldsymbol{u}_0 - \boldsymbol{u})|\boldsymbol{y}_2]\} = \boldsymbol{0}$$

- the third term is constant

Hence the minimum mean squared error is achieved by

$$\begin{aligned} \tilde{\boldsymbol{u}} = E(\boldsymbol{u}|\boldsymbol{y}_2) &= \text{cov}(\boldsymbol{u}, \boldsymbol{y}_2)[\text{var}(\boldsymbol{y}_2)]^{-1}\boldsymbol{y}_2 \\ &= \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{L}_2(\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2\boldsymbol{y} \\ &= \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{G} = \text{var}(\boldsymbol{u})/\sigma^2 = \oplus\{\gamma_i \boldsymbol{I}_{q_i}\}$

BLUPs (3)

The BLUP $\tilde{\boldsymbol{u}}$ can also be expressed as

$$\tilde{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}})$$

by using $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}})$ and expanding $\boldsymbol{H}^{-1}$ as

$$\boldsymbol{H}^{-1} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'.$$

The BLUPs are unbiased in a population sense

$$E(\tilde{\boldsymbol{u}}) = E(\boldsymbol{u}) = 0$$

but conditionally biased towards zero (shrinkage)

$$E(\tilde{\boldsymbol{u}}|\boldsymbol{u}) = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{u}.$$

The variance of the predictors is

$$\text{var}(\tilde{\boldsymbol{u}}) = \sigma^2\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}$$

but variation is usually considered in terms of the prediction error variance

$$\text{var}(\tilde{\boldsymbol{u}} - \boldsymbol{u}) = \sigma^2(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G})$$

which measures variation in terms of distance from the unobserved true value.

Mixed model equations (MMEs)

Estimates of fixed and random effects can be achieved from solving the mixed model equations:

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{y} \end{pmatrix}$$

where $\boldsymbol{G} = \oplus\{\gamma_i \boldsymbol{I}_{q_i}\}$.

These equations yield

- $\hat{\boldsymbol{\tau}} = (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}$
- $\tilde{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}})$

as required.

Many statistical packages use the mixed model equations (or parts of) for this calculation as it involves inversion of a matrix size $p + q$ (often $\ll n$) but does not require $\boldsymbol{H}^{-1}$ $(n \times n)$ for be formed explicitly.

Mixed model equations (2)

The inverse of the coefficient matrix also gives access to the variance matrix for the estimated (predicted) fixed and random effects.

The inverse of the coefficient matrix takes the form:

$$\begin{bmatrix} \boldsymbol{C}^{11} & \boldsymbol{C}^{12} \\ \boldsymbol{C}^{21} & \boldsymbol{C}^{22} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & -(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Z}\boldsymbol{G} \\ -\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1} & \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G} \end{bmatrix}$$

It is immediately clear that $\sigma^2\boldsymbol{C}^{11} = \text{var}(\hat{\boldsymbol{\tau}})$ and $\sigma^2\boldsymbol{C}^{22} = \text{var}(\tilde{\boldsymbol{u}} - \boldsymbol{u})$.

It is straightforward to verify that $\sigma^2\boldsymbol{C}^{12} = \text{cov}(\hat{\boldsymbol{\tau}}, \tilde{\boldsymbol{u}} - \boldsymbol{u})$.

eBLUEs and eBLUPs

In practice, variance parameters used in calculation of $\hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}$ are unknown.

Estimates of variance parameters are plugged in to get empirical or eBLUEs for the fixed effects and eBLUPs for the random effects.

Prediction error variances (PEVs) for the effects are also estimated by plugging the variance parameter estimates.

The estimated PEV then ignores uncertainty in the variance parameter estimates.

Inference on variance model

As the REML log-likelihood function ℓ_2 is a function of the variance parameters, we can use likelihood ratio tests.

Simple case

- test of $H_0 : \sigma_i^2 = 0$ against $H_a : \sigma_i^2 \neq 0$
- test statistic $-2(\ell_{20} - \ell_{2a})$ where ℓ_{20} (ℓ_{2a}) is the maximum value of ℓ_2 under the null (alternative) hypothesis
- usual asymptotic result: $-2(\ell_{20} - \ell_{2a}) \sim \chi_1^2$
- in general DF of χ^2-distribution for test increases with the number of variance parameters tested

Complications arise if the null hypothesis fixes one or more parameters on the boundary of the parameter space, eg. for $H_0 : \sigma_b^2 = 0$ against $H_a : \sigma_i^2 \geq 0$.

Constraints on variance components

The variance parameter σ_i^2 was defined by

$$\text{var}(\boldsymbol{u}_i) = \sigma_i^2 \boldsymbol{I}_{q_i}.$$

This definition has the implicit assumption $\sigma_i^2 \geq 0$.

However we can also specify the model as

$$E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\tau}, \quad \text{var}(\boldsymbol{y}) = \sigma^2 \boldsymbol{H}$$

for $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{I}_n$.

This specification requires only that $\boldsymbol{H}$ is positive-definite, without any explicit individual constraints on σ_i^2, $i = 1 \ldots c$.

Constraints on variance components (2)

For example, consider the RCBD with

- $\text{var}(y_{ij}) = \sigma_b^2 + \sigma^2$
- $\text{cov}(y_{ij}, y_{ik}) = \sigma_b^2$ for $j \neq k$
- $\text{cov}(y_{ij}, y_{lk}) = 0$ for $i \neq l$

The rationale for the RCBD is that units in the same block should be more similar than units in different blocks, which implies $\sigma_b^2 > 0$.

This assumption is based on prior expectations

- which may be wrong
- or the experimental procedure may be changed

In these cases, it is reasonable to allow for the case where correlation within blocks is smaller than correlation between blocks, ie $\sigma_b^2 < 0$.

In general, appropriate constraints depend on context.

Inference on variance model (2)

Consider a RLRT of $H_0 : \sigma_b^2 = 0$ against $H_a : \sigma_b^2 \geq 0$

- Under certain conditions, test statistic distribution is a 50:50 mixture of χ_0^2 distribution : χ_1^2 distribution.
- Crainiceanu & Ruppert (2004) showed that conditions used by Stram & Lee (1994) to develop this result do not necessarily hold in the general mixed model setting
- The asymptotic result requires that either
  - the data be independent and identically distributed
  - the data can be partitioned into independent subsets & number of subsets increases with the size of the dataset
- Crainiceanu & Ruppert (2004) assert that the asymptotic approximation is poor if either the independence condition does not hold, or if the number of independent subsets is small.
- No solution for general case! Can use parametric bootstrap to get empirical p-value.

Inference on fixed model

BLUEs of fixed effects have the distribution

$$\hat{\boldsymbol{\tau}} \sim N(\, \boldsymbol{\tau},\, \sigma^2(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\,) = N(\, \boldsymbol{\tau},\, \boldsymbol{V}_{\hat{\boldsymbol{\tau}}}\,)$$

To test $H_0 : \boldsymbol{\tau} = 0$ against general alternative $H_a : \boldsymbol{\tau} \neq 0$:

- cannot use LRT because ℓ_2 is not a function of $\boldsymbol{\tau}$ and ℓ_1/ℓ_2 partition depends on $\boldsymbol{X}$
- Wald test for $\boldsymbol{D\tau}$ uses statistic $\hat{\boldsymbol{\tau}}'\boldsymbol{D}'\boldsymbol{V}_{\hat{\boldsymbol{\tau}}}^{-1}\boldsymbol{D}\hat{\boldsymbol{\tau}}$
- in the case that $\boldsymbol{X}$ is not of full rank, issues of estimability arise and can be solved as for the linear model
- Wald test has asymptotic χ^2 distribution with DF equal to rank($\boldsymbol{D}$)
- Wald test is analogous to using χ^2 test in ANOVA rather than F-test
- asymptotic distribution is poor if stratum residual DF is small, especially if term DF is large
- calculation of denominator DF for approximate F-tests presented by Kenward & Roger (1997)

Bayesian interpretation

- Random effects from a REML analysis can be considered as empirical Bayes estimates from a prior

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}(\gamma))$$

- The parameters of the prior are estimated from the data rather than provided with their own prior distributions
- The BLUP is then the mean of the posterior distribution

$$\boldsymbol{u}|\boldsymbol{L}_2'\boldsymbol{y}$$

- Shrinkage in BLUPs can then be interpreted as the contribution of the prior to the posterior estimate: if there is much information about $\boldsymbol{u}$ in the data, then the contribution of the prior (shrinkage) is small

References / Further reading

Bailey, R.A. (2008) *Design of Comparative Experiments* Cambridge University Press, Cambridge.

Crainiceanu, C. M., & Ruppert, D. (2004) Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B*, **66**, 165-185.

Cullis, B.R, Smith, A.B. & Verbyla, A.R. (2005) Mixed models: a one day tour. Unpublished course notes.

Gilmour AR, Thompson R & Cullis BR (1995) AI, an efficient algorithm for REML estimation in linear mixed models. *Biometrics*, **51**, 1440-1450.

Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective.* Springer, New York.

Kenward, M. G. & Roger, J. H. (1997) The precision of fixed effects estimates from restricted maximum likelihood. *Biometrics*, **53**, 983-997.

Nelder, J. A. (1965) The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proceedings of the Royal Society, Series A*, **283**, 147–162.

Patterson, H.D. & Thompson, R. (1971) Recovery of interblock information when block sizes are unequal. *Biometrika* **31**, 100–109.

Pinheiro, J. & Bates, D. M. (2000) *Mixed Effects Models in S and S-Plus*, Springer-Verlag, New York.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992) *Variance Components*, J. W. Wiley, New York.

Stram, D. O. and Lee, J. W. (1994) Variance components testing in the longitudinal mixed effects setting. *Biometrics*, **50**, 1171-1177.

Verbeke G & Molenberghs G (2000) *Linear Mixed Models for Longitudinal Data*, Springer, New York.

Verbyla, A. P. (1990) A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, **32**, 227-230.

Problems

1. Derive the effective replications $\{\lambda_{ij}\}$ for the RCBD (slides 30-38)
2. Show that $\frac{\partial \boldsymbol{P}}{\partial \theta} = -\boldsymbol{P}\frac{\partial \boldsymbol{H}}{\partial \theta}\boldsymbol{P}$ for $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}\boldsymbol{H}^{-1}$, with $\boldsymbol{H} = \boldsymbol{H}(\theta)$ (slide 48)
3. Derive the form of the inverse coefficient matrix from the mixed model equations (slide 59)