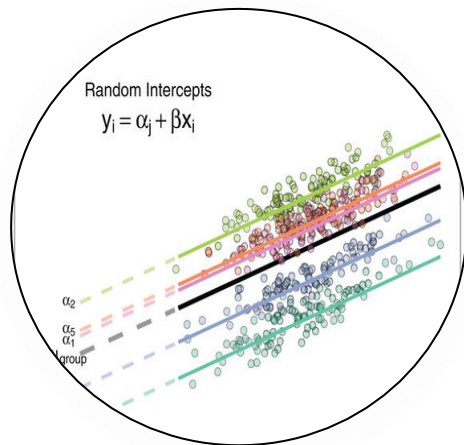


PhD course Mixed Linear Models

Session 3: General theory of mixed models with examples of variance components models with R

Gerrit Gort
Biometris, Wageningen University

June 10-11-12, 2025



Session 3

- General theory for Mixed Models
 - based on text from book “SAS System for Mixed Models”, Littell et al (2006)
- Examples and exercises using R:
 - paired and independent t-tests combined
 - random effects model, BLUP's and shrinkage
 - subsampling
 - block experiments, including balanced incomplete block designs and recovery of interblock information
 - split-plot experiments

Variance-covariance matrix

- Have good understanding of var-covar matrix of vector!
- Let \underline{y} be random vector with e.g. 4 elements:
- Let's call its variance-covariance matrix V :

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$

$$V = \text{Var}(\underline{y}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix}$$

variances on diagonal
covariances on off-diagonal places

$$\text{var}(y_i) = \sigma_i^2$$

$$\text{cov}(y_i, y_j) = \sigma_{ij}$$

V is symmetrical matrix: $\sigma_{ij} = \sigma_{ji}$

Calculation rule for variance-covariance matrices

- Let A be a matrix or a row vector of constants.

What is variance-covariance matrix of product $A \underline{y}$?

- $$\text{Var}(A \underline{y}) = A \text{Var}(\underline{y}) A^T = A \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix} A^T$$

- Example: difference of first 2 components of \underline{y} : $\underline{y}_1 - \underline{y}_2$.

Recall: $\text{var}(\underline{y}_1 - \underline{y}_2) = \text{var}(\underline{y}_1) + \text{var}(\underline{y}_2) - 2\text{cov}(\underline{y}_1, \underline{y}_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix}$$

Same!

$$\text{Var}((1 \ -1 \ 0 \ 0)\underline{y}) = (1 \ -1 \ 0 \ 0) \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} = (1 \ -1 \ 0 \ 0) \begin{pmatrix} \sigma_1^2 - \sigma_{12} \\ \sigma_{12} - \sigma_2^2 \\ \sigma_{13} - \sigma_{23} \\ \sigma_{14} - \sigma_{24} \end{pmatrix} = (\sigma_1^2 - \sigma_{12}) - (\sigma_{12} - \sigma_2^2)$$

Correlation matrix

- Correlation is scaled covariance (scaled by standard deviations):

$$\rho(\underline{y}_i, \underline{y}_j) = \text{cov}(\underline{y}_i, \underline{y}_j) / \sqrt{\text{var}(\underline{y}_i)} \sqrt{\text{var}(\underline{y}_j)} = \sigma_{ij} / \sqrt{\sigma_i^2} \sqrt{\sigma_j^2} = \sigma_{ij} / \sigma_i \sigma_j$$

- You can think of correlation as covariance between standardized variables (i.e. with standard deviation 1)
- In matrix form :

$$\text{Corr}(\underline{y}) = \begin{pmatrix} \sigma_1^2 / \sigma_1 \sigma_1 & \sigma_{12} / \sigma_1 \sigma_2 & \sigma_{13} / \sigma_1 \sigma_3 & \sigma_{14} / \sigma_1 \sigma_4 \\ \sigma_{12} / \sigma_1 \sigma_2 & \sigma_2^2 / \sigma_2 \sigma_2 & \sigma_{23} / \sigma_2 \sigma_3 & \sigma_{24} / \sigma_2 \sigma_4 \\ \sigma_{13} / \sigma_1 \sigma_3 & \sigma_{23} / \sigma_2 \sigma_3 & \sigma_3^2 / \sigma_3 \sigma_3 & \sigma_{34} / \sigma_3 \sigma_4 \\ \sigma_{14} / \sigma_1 \sigma_4 & \sigma_{24} / \sigma_2 \sigma_4 & \sigma_{34} / \sigma_3 \sigma_4 & \sigma_4^2 / \sigma_4 \sigma_4 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{pmatrix}$$

Mixed Linear Models in matrix notation

■ Matrix notation

● Standard linear model

$$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_n \end{bmatrix}$$

$$\underline{y} = X\beta + \underline{e}$$

$$\text{Var}(\underline{y}) = \text{Var}(\underline{e}) = \sigma^2 I_n$$

\underline{y} and \underline{e} normally distributed

$$E(\underline{e}) = 0, \quad E(\underline{y}) = X\beta$$

● Mixed linear model

$$\begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1r} \\ z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \underline{u}_1 \\ \vdots \\ \underline{u}_r \end{bmatrix} + \begin{bmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_n \end{bmatrix}$$

$$\underline{y} = X\beta + Z\underline{u} + \underline{e}$$

$$\text{Var}(\underline{u}) = G, \quad \text{Var}(\underline{e}) = R,$$

$$\text{Var} \begin{bmatrix} \underline{u} \\ \underline{e} \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \text{ (uncorrelated)}$$

$$\rightarrow \text{Var}(\underline{y}) = V = ZGZ^T + R$$

\underline{y} , \underline{u} and \underline{e} normally distributed

$$E(\underline{u}) = 0, \quad E(\underline{e}) = 0, \quad E(\underline{y}) = X\beta$$

Question: Matrix notation for standard Linear Model

- Suppose we have small regression problem: 6 plots with yields 12, 15, 18, 20, 19, 24, corresponding to nitrogen applications of 1, 1, 2, 2, 3, 3 units. Assume that the yields are uncorrelated, and have constant variance σ^2 .

We want to fit a simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \underline{e}_i$

- Write down the model in matrix notation.

Also give $V = \text{Var}(y)$.

- Suppose we have an ANOVA problem, 6 plots with same yields as before, but 3 fertilizer treatments: manure, nitrogen, combination. Assume that the yields are uncorrelated, and have constant variance σ^2 .

We want to fit a one-way ANOVA model: $y_{ij} = \mu + \alpha_i + \underline{e}_{ij}$

- Write down the model in matrix notation.

Also give $V = \text{Var}(y)$

- Same ANOVA model, written differently (without overparametrization)

- Model in matrix notation?

$$y_{ij} = \mu_i + \underline{e}_{ij}$$

Mixed Linear Models: software

- R: packages nlme, lme4, glmmTMB and others; add-on package asreml (not open source)
- SAS: PROC MIXED
- Genstat: through menu or programming language
- SPSS: through menu or programming language

- nlme() and lme4() will be our working horse this course

Mixed Linear Model

$$\underline{y} = X\beta + Z\underline{u} + \underline{e}$$

- Part $X\beta$ known from Linear Models (regressors are columns of X)
- \underline{u} = vector unknown random-effects (could be random (deviations of) intercepts, random slopes)
- Z = known design matrix of \underline{u} , as the X-matrix (like in Linear Models)
- Model $\text{Var}(\underline{y}) = V$ by setting up
 - random-effects design matrix Z
 - specifying covariance structures for
 - G
 - R
- “Simple” random effects:
 - Z contains dummy variables
 - G contains variance component in diagonal structure: $G = \sigma_u^2 I_r$
 - R is diagonal as well: $R = \sigma_e^2 I_n$

Mixed Linear Model

$$\text{Var}(\underline{y}) = V = ZGZ^T + R$$

■ $V = \text{Var}(\underline{y}) =$
 $= \text{Var}(X\beta + Z\underline{u} + \underline{e})$
 $= \text{Var}(Z\underline{u} + \underline{e})$
 $= \text{Var}(Z\underline{u}) + \text{Var}(\underline{e})$
 $= Z \text{var}(\underline{u}) Z^T + \text{Var}(\underline{e})$
 $= ZGZ^T + R$

[because $X\beta$ is constant]

[because \underline{u} and \underline{e} are uncorrelated]

[use calculation rule $\text{Var}(A\underline{y})$]

Example from plant breeding

- Three doubled haploid lines (DHs) of barley were evaluated for height, where each DH was represented by $k=10$ plants.
- Interest focusses on the amount of genetic variation between the lines in comparison to the amount of within line variation.
- In principle the variation within lines should represent only environmental variation, because all the plants within a line have the same genotype.

Data, scalar model and matrix model specification

cultivar	plant	height
1	1	88.0
1	2	88.0
1	3	94.8
1	4	90.0
1	5	93.0
1	6	89.0
1	7	86.0
1	8	92.9
1	9	89.0
1	10	93.0
2	1	85.9
2	2	88.6
2	3	90.0
2	4	87.1
2	5	85.6
2	6	86.0
2	7	91.0
2	8	89.6
2	9	93.0
2	10	87.5
3	1	94.2
3	2	91.5
3	3	92.0
3	4	96.5
3	5	95.6
3	6	93.8
3	7	92.5
3	8	93.2
3	9	96.2
3	10	92.5

[illegible]

$$y_{ij} = \mu + \underline{a}_i + \underline{e}_{ij}$$

$$\underline{a}_i \sim \mathcal{N}(0, \sigma_a^2), \underline{e}_{ij} \sim \mathcal{N}(0, \sigma_e^2)$$

$$y = X\beta + Z\underline{u} + \underline{\epsilon}$$

$$\text{Var}(y) = ZGZ^T + R$$

Variance-covariance matrix for barley example

$$V(\underline{y}) = ZGZ^T + R \text{ with } G = \sigma_a^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \sigma_a^2 I_3 \text{ and } R = \sigma_e^2 I_{30}$$

Check!

$$ZGZ^T = \sigma_a^2$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \sigma_a^2 \begin{bmatrix} J_{10} & 0_{10} & 0_{10} \\ 0_{10} & J_{10} & 0_{10} \\ 0_{10} & 0_{10} & J_{10} \end{bmatrix}$$

block
diagonal
(10×10)

```

1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1
    
```

Var-covar matrix for barley example (cont.)

$$V = ZGZ^T + R = \begin{bmatrix} \sigma_a^2 J_{10} + \sigma_e^2 I_{10} & 0 & 0 \\ 0 & \sigma_a^2 J_{10} + \sigma_e^2 I_{10} & 0 \\ 0 & 0 & \sigma_a^2 J_{10} + \sigma_e^2 I_{10} \end{bmatrix}$$

$$\begin{bmatrix} \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 \end{bmatrix}$$

Hence: $\text{var}(y_{ij}) = \sigma_a^2 + \sigma_e^2$
for $i=i^*$ (same genotype) and $j \neq j^*$:
 $\text{cov}(y_{ij}, y_{i^*j^*}) = \sigma_a^2$
 $\rho(y_{ij}, y_{i^*j^*}) = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$
0 otherwise

"Compound symmetry"
structure

Intraclass correlation
coefficient

Other example: growth curve with compound symmetry

- Suppose s individuals, 3 repeated growth measurements on time points 1,2,3 per individual
- Fit overall linear trend in time $y_{ij} = \beta_0 + \beta_1 t_{ij} + \underline{a}_i + \underline{e}_{ij}$
- Suppose common correlation among observations from single individual, being the same for all individuals (but is this a realistic assumption?) Var-covar matrix is block diagonal with s blocks, each with compound symmetrical structure. Two parameters: $\text{var}(\underline{a}_i) = \sigma_a^2$ and $\text{var}(\underline{e}_{ij}) = \sigma_e^2$:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

$$\text{Var}(\underline{y}) = V = \begin{bmatrix} \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 & \dots & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \dots & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & 0 & \dots & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 \\ 0 & 0 & 0 & \dots & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 \end{bmatrix}$$

Specification of model with compound symmetry (R)

- Package nlme, function lme() :

`lme(fixed = y ~ time, random = ~ 1 | indiv)`

- specify fixed part of model in same way as in linear models, specifying the design matrix X
- specify random part of model: random effects (intercepts) for individuals (specifying design matrix Z ; individuals form grouping factor)

$$\underline{y} = X\underline{\beta} + Z\underline{u} + \underline{e}$$

$$\text{Var}(\underline{y}) = ZGZ^T + R$$

- Package lme4, function lmer() :

`lmer(y ~ time + (1 | indiv))`

- From R-help: formula object describing both the fixed-effects and random-effects part of the model, with the response on the left of a ~operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars ("|") separating expressions for design matrices from grouping factors.

$$Z = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \\ & & & & & & 1 \\ & & & & & & & 1 \end{bmatrix} \quad G = \begin{bmatrix} \sigma_a^2 & & & & & \\ & \sigma_a^2 & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & \sigma_a^2 & \end{bmatrix}$$

Specification of model with compound symmetry (cont.)

- Another way to specify the same model: eliminate Z and G from the model, and specify matrix R directly:

- Use function `gls()` of `nlme` package, and specify compound symmetric correlation structure.
- Two parameters: σ_e^2 and correlation ρ

```
gls(y ~ time, correlation =  
    corCompSymm(form = ~ 1 | indiv))
```

$$\underline{y} = X\beta + \underline{e}$$

$$\text{Var}(\underline{y}) = R$$

$$R = \sigma_e^2 \begin{bmatrix} 1 & \rho & \rho & \cdots & 0 & 0 & 0 \\ \rho & 1 & \rho & \cdots & 0 & 0 & 0 \\ \rho & \rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \rho & \rho \\ 0 & 0 & 0 & \cdots & \rho & 1 & \rho \\ 0 & 0 & 0 & \cdots & \rho & \rho & 1 \end{bmatrix}$$

Exercise 3: paired or indep-samples t-test or both (part 2)

- Again fit the mixed model to include all available information, both from the “paired” part and “independent samples” part.
- In the R-script it is done in three different ways. Study the different ways:
 - Specifying the part Zu (using functions `lme` and `lmer`)
 - Bypassing the Zu part, specifying correlation matrix for R instead
- Pay attention to the different functions to extract the information from the fitted models.
- Write down the fitted model:
 - In scalar form
 - In matrix form:
 - X matrix, parameters β
 - Z matrix, random effects \underline{u}
 - Var-covar matrices G, R
 - Var-covar matrix V
- What is your conclusion w.r.t. feed treatment?

Estimating G and R in Mixed Model

- Unknown parameters in β , G and R , unknown random effects u .
- If we would know G and R , we could use GLS (=Generalized Least Squares) to estimate β (instead of OLS=Ordinary Least Squares):

$$(\underline{y} - X\beta)^T V^{-1} (\underline{y} - X\beta)$$

- But V is unknown
- As $V = ZGZ^T + R$, estimates of (parameters in) G and R are needed
- G and R are estimated using maximum likelihood or restricted maximum likelihood

Estimating β and \underline{u}

- ML or REML deliver estimates \hat{G} and \hat{R}
- Remember normal equations from the ordinary linear model $\underline{y} = X\beta + \underline{e}$
$$X^T X \hat{\beta} = X^T y \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$
- In mixed models $\underline{y} = X\beta + Z\underline{u} + \underline{e}$ we have extended normal equations, so-called mixed model equations (Henderson, 1984):

$$\begin{bmatrix} X^T \hat{R}^{-1} X & X^T \hat{R}^{-1} Z \\ Z^T \hat{R}^{-1} X & Z^T \hat{R}^{-1} Z + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\underline{u}} \end{bmatrix} = \begin{bmatrix} X^T \hat{R}^{-1} y \\ Z^T \hat{R}^{-1} y \end{bmatrix}$$

- Result:
 - *estimation* of parameters: $\hat{\beta}$
 - *prediction* of random effects: $\hat{\underline{u}}$

BLUE and BLUP

- If G and R known: $\hat{\beta}$ is **BLUE**=best linear unbiased estimator
 \hat{u} is **BLUP**=best linear unbiased predictor

$$\text{Var} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = C = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-1}$$

- Plug in estimated values for G and R → empirical BLUE and BLUP
- True sampling variability of $\hat{\beta}$ and \hat{u} tends to be underestimated.
- Examples of BLUP's are e.g. breeding values for bulls in animal breeding.

Inference and test statistics

- For inference on covariance parameters: use likelihood ratio (LR) tests, by comparing nested Full and Reduced models, so fit 2 nested models separately.
- Inference on fixed- and random-effects parameters, consider linear combinations of form $L \begin{bmatrix} \beta \\ \underline{u} \end{bmatrix}$
- If L is single row: general t-statistic
 - In general only *approximately* t-distributed
 - df must be estimated as well
$$t = \frac{L \begin{bmatrix} \hat{\beta} \\ \hat{\underline{u}} \end{bmatrix}}{\sqrt{L \hat{C} L^T}}$$
- If rank(L)>1: general F-statistic
 - In general only *approximately* F-distribution
 - df1=rank(L), df2 must be estimated: different methods!
$$F = \frac{\begin{bmatrix} \hat{\beta} \\ \hat{\underline{u}} \end{bmatrix}^T L^T (L^T \hat{C} L)^{-1} L \begin{bmatrix} \hat{\beta} \\ \hat{\underline{u}} \end{bmatrix}}{\text{rank}(L)}$$
- For likelihood ratio tests of fixed effects, use ML, don't use REML!
- More in the part "Estimation and testing in a mixed model"

Compare fixed and random effects model – shrinkage

- Example: 3 barley lines, $k=10$ plants per line (see before)
- Fixed effects model (1-way ANOVA): $y_{ij} = \mu_i = \mu + \alpha_i + e_{ij}$
 - Interest in these 3 specific lines
 - Parameters: 3 mean parameters (μ_1, μ_2, μ_3 ; or $\mu, \alpha_1, \alpha_2, \alpha_3$, e.g. with $\sum \alpha_i = 0$, or $\alpha_1 = 0$)
1 variance parameter σ_e^2 (within line variance)
 - Estimator of $\mu + \alpha_i$ is simply the group mean: $\hat{\mu} + \hat{\alpha}_i = \bar{y}_{.i}$
- Random effects model: $y_{ij} = \mu + a_i + e_{ij}$
 - Lines are assumed to be sampled from population
 - Parameters: 1 mean parameter (μ)
2 variance components σ_a^2 and σ_e^2 (between and within line variance)
 - Predictor of $\mu + a_i$ is *NOT* the ordinary group mean, but *weighted average* of overall mean and group mean. This results in *shrinkage* towards the overall mean.

Shrinkage

- *Shrinkage* estimator / predictor:

$$\widehat{\mu + \underline{a}_i} = (1 - w)\bar{y}_{..} + w\bar{y}_{i.}$$

$$w = \frac{k\sigma_a^2}{k\sigma_a^2 + \sigma_e^2}$$

result is somewhere between overall mean $\bar{y}_{..}$ and the group mean $\bar{y}_{i.}$, “shrunk” towards overall mean.

- If $\sigma_a^2 = \sigma_e^2$ (between variance = within variance), then
 - E.g. $k=10$ (10 plants per line) $\rightarrow w = 10/11$

$$w = \frac{k}{k+1}$$

$$\widehat{\mu + \underline{a}_i} = \left(1 - \frac{10}{11}\right)\bar{y}_{..} + \frac{10}{11}\bar{y}_{i.} = \frac{1}{11}\bar{y}_{..} + \frac{10}{11}\bar{y}_{i.} \quad \text{so, close to the group mean, little shrinkage}$$

- If $\sigma_a^2 \gg \sigma_e^2$, then w will be close to 1, and there is hardly shrinkage towards the population mean.
- If $k\sigma_a^2 \ll \sigma_e^2$, then w will be close to 0, and shrinkage is almost complete.

Expected values and conditional expected values

- Mixed model: $\underline{y} = X\beta + Z\underline{u} + \underline{e}$
- As random effects \underline{u} and \underline{e} have expected values 0, we have $E(\underline{y}) = X\beta$ which is the *unconditional expectation* of \underline{y}
 - mean of \underline{y} averaged over all possible \underline{u} and \underline{e}
 - solution of mixed model equations results in BLUE = Best Linear Unbiased Estimator of $X\beta$, or estimable functions of it.
- Sometimes interest in *conditional expectation, given random effects \underline{u}* :
 $E(\underline{y} / \underline{u}) = X\beta + Z\underline{u}$
 - mean of \underline{y} for specific level of random effect actually observed
 - To “estimate” this conditional expectation, BLUP’s of random effects \underline{u} are needed.

Exercise 4 Barley example: fixed vs random effects

- a. Read data into software (30 observations: 3 lines, 10 plants per line)
- b. Fit fixed effects model (1-way ANOVA), and estimate means $\mu + \alpha_i$
 - Study output: ANOVA table, estimate of error variance σ_e^2 , parameter estimates, estimated means. Check that these are ordinary means.
 - F-test tests: $H_0: \alpha_1 = \alpha_2 = \alpha_3$. Conclusion?
- c. Fit random effects model:
 - Study output: log-likelihood (from REML), variance components and predictions of conditional means. Check that these indeed are shrunk towards the overall mean, according to the formula given earlier.
 - How strong is the shrinkage?
- d. If we take the ANOVA approach to estimate the variance components, what will be the expected mean squares for line and residual? Use `lm()` to get the mean squares of line and residual, and use these to reconstruct the estimates of variance components σ_a^2 and σ_e^2 as shown in the `lme()` output.

Subsampling

- Sometimes it is necessary or convenient to randomly sample subunits of the experimental units to obtain the required data for a study.
- Example: experiment with pots, each containing 5 plants; treatments are randomized over pots, making pots the experimental units. However, measurements are made on the 5 individual plants within pots: *pseudo-replication*.
- In such cases *observational units* are taken from the larger *experimental units*.
- Subsamples introduce another random source of variation in addition to that among experimental units.
- The distinction between experimental units and observational units is important to get the right standard errors and hypothesis tests.

Example Subsampling

- Pesticide residues on plants were assessed, with two treatments A and B. Six batches of plants were available. Treatments were randomized over the batches, making the batches experimental units. Two plants (subsamples) were taken from each batch.

method	batch	subsample	pesticide
A	1	1	120
A	1	2	110
A	2	1	120
A	2	2	100
A	3	1	140
A	3	2	130
B	4	1	71
B	4	2	71
B	5	1	70
B	5	2	76
B	6	1	63
B	6	2	68

Model: $y_{ijk} = \mu + \tau_i + b_{ij} + e_{ijk} \quad i=1\dots t, j=1\dots r, k=1\dots n$

$$b_{ij} \sim N(0, \sigma_b^2), \quad e_{ijk} \sim N(0, \sigma_e^2)$$

$t=2$ treatments

$r=3$ experimental units (batches) per treatment

$n=2$ observational units (plants) per exp. unit

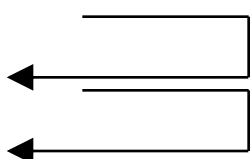
σ_b^2 is variance component *between* exp. units (batches)

σ_e^2 is variance component *within* exp. units (batches)

b_{ij} and e_{ijk} are assumed to be uncorrelated

Testing within ANOVA framework (balanced design)

- How to assess treatment differences?
- Use variation at level of *experimental units* ($MS(Exp.Error)$), not at level of the observational units!
- EMS table: for a valid F-test for treatment differences search for that term that has the same expected mean square under the null hypothesis of no differences, i.e., F for treatments = $MS(Treatments) / MS(Exp.Error)$
- Test for significance of the variance component for experimental error (between cluster variance), using $F = MS(Exp.Error) / MS(Obs.Error)$

Source of variation	df	E(MS)	F-test
treatments	$t-1$	$rn\tau_{tr}^2 + n\sigma_b^2 + \sigma_e^2$	
exp. unit error	$t(r-1)$	$n\sigma_b^2 + \sigma_e^2$	
obs. unit error	$tr(n-1)$	σ_e^2	

Variance and standard errors in subsampling scheme

- Variance and standard errors of treatment means and treatment difference in a subsampling scheme:

- variance treatment mean

$$\sigma_{y_{i..}}^2 = \frac{\sigma_b^2}{r} + \frac{\sigma_e^2}{r n} = \frac{n\sigma_b^2 + \sigma_e^2}{r n} = \frac{EMS(Exp.Error)}{r n}$$

- so, s.e.(mean) estimated by

$$\hat{\sigma}_{y_{i..}} = \sqrt{\frac{MS(Exp.Error)}{r n}}$$

- standard error of difference is

$$SED = SE(\bar{y}_{i..} - \bar{y}_{i^{*}..}) = \sqrt{2 \frac{MS(Exp.Error)}{r n}}$$

$$\text{var}(\bar{y}_{i..} - \bar{y}_{i^{*}..}) = \text{var}(\bar{y}_{i..}) + \text{var}(\bar{y}_{i^{*}..}) = 2 \text{var}(\bar{y}_{i..})$$

Choice of right model for random part

- Not acknowledging subsampling structure and pooling experimental and observational error will lead to erroneous conclusions in testing (df too large, se's wrong).
- The reason is that the wrong error structure is used: observations made on different plants for the same exp. unit (batch j within treatment i) are correlated:

$$\text{cov}(y_{ijk}, y_{ijk*}) = \text{Cov}(\underline{b}_{ij} + \underline{e}_{ijk}, \underline{b}_{ij} + \underline{e}_{ijk*}) = \text{Cov}(\underline{b}_{ij}, \underline{b}_{ij}) = \text{var}(\underline{b}_{ij}) = \sigma_b^2 \neq 0$$

wrong

right

```
lm(y ~ method)
```

```
lm(y ~ method + batch:method)
```

```
aov(y ~ method + Error(batch:method))
```

```
lme(y ~ method, random= ~ 1 | batch)
```

```
lmer(y ~ method + (1 | batch))
```

Naive output by `lm()`: what is wrong?

$$F = MS_{\text{method}} / MS_{\text{ObsErr}} = 7550.1 / 55.1 = 137.1$$

```
> lm.sub <- lm(pesticide ~ method + method:batch, data=subsampling)
```

```
> anova(lm.sub) # wrong F-test for method!
```

Analysis of Variance Table

Response: pesticide

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
method	1	7550.1	7550.1	137.0666	2.342e-05
method:batch	4	760.3	190.1	3.4508	0.08597
Residuals	6	330.5	55.1		

Wrong, must be:

$$F = MS_{\text{method}} / MS_{\text{batch}} = 7550.1 / 190.1 = 39.7$$

$$MS_{\text{batch}} = MS_{\text{ExpErr}}$$

$$MS_{\text{ObsErr}}$$

Wrong, standard error must be:

```
> emmeans(lm.sub, ~ method) # gives warning
```

method	emmean	SE	df	asympt.LCL	asympt.UCL
a	120.000	3.0299	6	112.586	127.41400
b	69.833	3.0299	6	62.419	77.24734

$$\sqrt{MS_{\text{ExpErr}} / (3 \times 2)} = \sqrt{190.08 / 6} = 5.63$$

```
> emmeans(lm.sub, ~ batch:method) # this is ok
```

batch	method	emmean	SE	df	lower.CL	upper.CL
1	a	115.0	5.248015	6	102.15857	127.84143
2	a	110.0	5.248015	6	97.15857	122.84143
3	a	135.0	5.248015	6	122.15857	147.84143
4	b	71.0	5.248015	6	58.15857	83.84143
5	b	73.0	5.248015	6	60.15857	85.84143
6	b	65.5	5.248015	6	52.65857	78.34143

$$\sqrt{MS_{\text{ObsErr}} / 2} = \sqrt{55.08 / 2}$$

OK

Proper analysis lme()

```
> mm.sub1 <- lme(fixed = pesticide ~ method, random = ~ 1 | batch)
> summary(mm.sub1)
```

...
Random effects:

Formula: ~1 | batch

(Intercept) Residual

StdDev: 8.215835 7.421816

 $\hat{\sigma}_b$
 $\hat{\sigma}_e$

Fixed effects: pesticide ~ method

	Value	Std.Error	DF	t-value	p-value
(Intercept)	120.00000	5.628547	6	21.319890	0.0000
methodb	-50.16667	7.959967	4	-6.302371	0.0032

s.e.d. OK

s.e.m. OK

s.e.d. OK

```
> anova(mm.sub1) # proper F-test
```

	numDF	denDF	F-value	p-value
(Intercept)	1	6	568.7513	<.0001
method	1	4	39.7199	0.0032

F-test for
method OK

```
> emmeans(mm.sub1, pairwise ~ method)
```

\$emmeans

method	emmean	SE	df	asympt.LCL	asympt.UCL
a	120.00000	5.628547	NA	108.96692	131.03308
b	69.83333	5.628547	NA	58.80025	80.86642

\$contrasts

contrast	estimate	SE	df	t.ratio	p.value
a - b	50.16667	7.959967	4	6.302	<.0032

Proper analysis lmer()

```
> mm.sub2 <- lmer(pesticide ~ method + (1 | batch))
> summary(mm.sub2)
```

Linear mixed model fit by REML ['lmerMod']
Formula: pesticide ~ method + (1 | batch)

....
Random effects:

Groups	Name	Variance	Std.Dev.
batch	(Intercept)	67.50	8.216
	Residual	55.08	7.422

$\hat{\sigma}_b^2$
 $\hat{\sigma}_e^2$

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	120.000	5.629	21.320
methodb	-50.167	7.960	-6.302

...

```
> library(lmerTest)
> anova(mm.sub2, ddf="kenward-roger")
Analysis of Variance Table of type 3 with Kenw
Roger approximation for degrees of freedom
```

	Sum Sq	Mean Sq	NumDF	DenDF	F-value	Pr(>F)
method	2187.9	2187.9	1	4	39.72	0.00324

F-test for
method OK

```
> emmeans(mm.sub2, pairwise ~ method)
```

method	emmean	SE
a	120.00000	5.628548
b	69.83333	5.628548

s.e.m. OK

Confidence level used: 0.95

```
$contrasts
```

contrast	estimate	SE	df	t.ratio	p.value
a - b	50.16667	7.959969	4	6.302	0.0032

s.e.d. OK

Proper analysis aov()

```
> aov.sub <- aov(pesticide ~ method + Error(batch), data=subsampling)
> summary(aov.sub)
```

Error: batch

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
method	1	7550	7550	39.72	0.00324 **
Residuals	4	760	190		

F-test for
method OK

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: within

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	6	330.5	55.08		

$\hat{\sigma}_e^2$

```
> emmeans(aov.sub, pairwise ~ method)
Note: re-fitting model with sum-to-zero contrasts
```

\$emmeans

method	emmean	SE	df	lower.CL	upper.CL
a	120.0	5.63	4	104.4	135.6
b	69.8	5.63	4	54.2	85.5

s.e.m. OK

s.e.d. OK

Warning: EMMs are biased unless design is perfectly balanced
Confidence level used: 0.95

\$contrasts

contrast	estimate	SE	df	t.ratio	p.value
a - b	50.2	7.96	4	6.302	0.0032

Subsampling in factorial experiment (§7.6.3 S&P)

- Example: 2 crossed treatment factors

- amount of daylight (8, 12, 16 hrs)
- night temperature (low, high)

- per treatment combination 3 pots = replicates, so in total $3 \times 2 \times 3 = 18$ pots in completely randomized design

- per pot 4 mint plants (subsamples), so $4 \times 18 = 72$ observations

- pots are experimental units, not plants!

- so we have 18 experimental units, not 72!

pseudo-replication

- response = one week stem growth of mint plant = y

- Model

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \underline{p}_{ijk} + \underline{e}_{ijkl}$$

$$\underline{p}_{ijk} \sim N(0, \sigma_p^2), \quad \underline{e}_{ijkl} \sim N(0, \sigma_e^2)$$

with variance components

$$\text{var}(\underline{p}_{ijk}) = \sigma_p^2$$

variance component for pots

$$\text{var}(\underline{e}_{ijkl}) = \sigma_e^2$$

variance component for plants within pots

Book: Schabenberger O. & Pierce FJ (2002), *Contemporary Statistical Models for the Plant and Soil Sciences*; many examples using SAS

Exercise 5: Subsampling in factorial experiment

Follow the analysis in hand-out of S&P (§7.6.3), starting from pg 482:

- a. Check how data are read into R.
- b. Function `lm()` is meant for fixed effects models, but can be (ab)used for some some simple random effects models. The model specified in `lm()` (cf first PROC GLM pg 483 of book) allows for main effects of hour and night, their interaction, and for a pot effect (notice the nesting: `pot %in% (hour*night)`). The F-tests for main effects of hour and night, and the interaction, are wrong, since the mean square in the denominator is an estimate of the within pot variance, and not of the needed between pot variance. F-statistic denominator should be the mean square of `pot(hour*night)`.
- c. In this part we use the function `aov()` to get the right F-tests. (Cf second PROC GLM pg 483). Ignore the warning.

Exercise 5: Subsampling in factorial experiment (cont.)

- d. Now construct the F-statistics yourself by taking the mean square for pots out of anova table, and use as denominator of F-statistics for hour, night, and hour:night. (Cf third PROC GLM pg 483).
- e. Now a mixed model approach is taken. Function lmer() is used. Try to specify the model using lme() from nlme package yourself.
- f. What would be a simpler approach here? Try yourself.

Experimental design: blocking

- Statistical theory for efficient estimation (unbiased estimates, low variance) of treatment differences is based on principles of
 - **Randomization**: random assignment of treatments to experimental units provides a valid estimate of error, eliminates biases in the comparison of treatments, and justifies the statistical inference.
 - **Replication**: $\text{variance of mean} = \text{variance of single observation} / \text{number of replicates}$; more replicates yield more accurate means
 - **Blocking**: grouping of experimental units in such a way that disturbances are minimal within blocks and occur mainly between blocks. A good choice of blocks will reduce the error variance.
- Examples of block designs:
 - **Completely randomized design (CRD)**: no block structure
 - **Randomized complete blocks (RCBD)**: each block contains all treatments at least once
 - **Incomplete blocks (IB)**: number of treatments exceeds number of plots in a block
 - **Split plot (in blocks)**: two types of experimental units within blocks; main plots and sub plots

CRD and RCBD

- CRD = Completely randomized design:
 - Example: 4 treatments (a, b, c, d), replicated 3 times, making 12 plots
 - Each plot has the same probability of receiving a particular treatment.
 - So, how is the randomization? Restrictions on randomization?

- RCBD = Randomized Complete Block Design
 - Example: 4 treatments in 3 blocks of size 4
 - Each treatment appears in each block.
 - How is the randomization done? Restrictions on randomization?

a	b	d
b	c	c
a	d	a
c	b	d

c	a	c
b	c	b
a	d	d
d	b	a

RCBD: example (Kuehl, Ex. 8.1, p264)

Book: Kuehl RO (2000), *Design of Experiments – Statistical Principles of Design and Analysis*

■ Example:

- A researcher wants to compare 6 different application schemes of N-fertilizer (rate and timing differences) in wheat.
- Experiment was conducted on an irrigated field with a water gradient along one direction.
- To control for moisture differences between the plots, blocks of 6 plots were formed.

BLOCK	TREATMNT	NITROGEN
1	1	34.98
1	2	40.89
1	3	42.07
1	4	37.18
1	5	37.99
1	6	34.89
2	1	41.22
2	2	46.69
2	3	49.42
2	4	45.85
2	5	41.99
2	6	50.15
3	1	36.94
3	2	46.65
3	3	52.68
3	4	40.23
3	5	37.61
3	6	44.57
4	1	39.97
4	2	41.9
4	3	42.91
4	4	39.2
4	5	40.45
4	6	43.29

RCBD: blocks fixed or random?

- Model for fixed block effects:

$$\underline{y}_{ij} = \mu + \tau_i + \beta_j + \underline{e}_{ij}$$

$$\underline{e}_{ij} \sim N(0, \sigma_e^2)$$

treatment effects ($i=1, \dots, t$); block effects ($j=1, \dots, r$)

- RCBD with fixed block effects: two-way factorial treatment design: factors block and treatment
- Usually: **additive model**, i.e. no interaction between blocks and treatments assumed
- Gives ANOVA table

Source	df	MS	$E(MS)$
Blocks	$r-1$	$MS(\text{Blocks})$	
Treatments	$t-1$	$MS(\text{Tr})$	$r\theta_{tr}^2 + \sigma_e^2$
Error	$(r-1)(t-1)$	$MS(\text{Error})$	σ_e^2

- Test for treatment main effect, use variance ratio $F = MS(\text{Tr})/MS(\text{Error})$, which under H_0 will be F distributed with $t-1$ and $(r-1)(t-1)$ df. "Within block analysis"
- $\hat{S\hat{E}D} = S\hat{E}_{(\underline{\bar{y}}_{i.}, \underline{\bar{y}}_{i^*})} = \sqrt{2MS(\text{Error})/r}$
- Besides: checking treatment-block interaction in fixed RCBD requires more than one experimental unit per treatment in each block. Then σ_e^2 represents variation between experimental units with the same treatment within a block.

RCBD: blocks fixed or random?

- Model for random block effects:

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}$$

$$b_j \sim N(0, \sigma_b^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

- Gives ANOVA table

Source	df	MS	EMS
Blocks	$r-1$	$MS(\text{Blocks})$	$t\sigma_b^2 + \sigma_e^2$
Treatments	$t-1$	$MS(\text{Tr})$	$r\sigma_{\tau}^2 + \sigma_e^2$
Error	$(r-1)(t-1)$	$MS(\text{Error})$	σ_e^2

- Test for treatment effects: $F = MS(\text{Tr}) / MS(\text{Error})$ under H_0 distributed as $F(t-1, (r-1)(t-1))$
- Means and standard errors:

$$\bar{y}_{1.} = \mu + \tau_1 + \bar{b}_{.} + \bar{e}_{1.}$$

$$\bar{y}_{1.} - \bar{y}_{2.} = \tau_1 - \tau_2 + \bar{e}_{1.} - \bar{e}_{2.}$$

$$\text{var}(\bar{y}_{1.}) = (\sigma_b^2 + \sigma_e^2) / r$$

$$\text{var}(\bar{y}_{1.} - \bar{y}_{2.}) = 2\sigma_e^2 / r$$

model values

estimated as

$$\hat{\sigma}_e^2 = MS(\text{Error})$$

$$\hat{\sigma}_b^2 = (MS(\text{Blocks}) - MS(\text{Error})) / t$$

$$\hat{\text{var}}(\bar{y}_{1.}) = (\hat{\sigma}_b^2 + \hat{\sigma}_e^2) / r =$$

$$[MS(\text{Blocks}) / rt] + [(t-1)MS(\text{Error}) / rt]$$

$$\hat{\text{var}}(\bar{y}_{1.} - \bar{y}_{2.}) = 2MS(\text{Error}) / r$$

Difference between fixed and random blocks: block variance enters in variance and SE of a mean. But SEDs stay the same!

Exercise 6: RCBD: fixed or random block effects

Example Kuhl on fertilizer application schemes in RCBD

- Fit the model with fixed block effects
- Fit the model with random block effects.

Check that:

- the standard error of treatment mean (s.e.m.) is different for the two approaches; in which case is the s.e.m. larger?
- The standard error of the difference of two means (s.e.d.) is the same for the two approaches.
- Write down the elements of the model in matrix notation.

$$\underline{y} = X\beta + Z\underline{u} + \underline{e}$$

$$\text{Var}(\underline{y}) = ZGZ^T + R$$

Incomplete Block Design

- With many treatments, inclusion of all treatments in a block can make the blocks too large, so that they become heterogeneous again. Or, in some cases, the block size is simply too small to contain all treatments.
- In these situation incomplete blocks designs may be used, retaining most of the good properties of complete block designs, but allowing larger numbers of treatments to be compared.
- For *efficient* analysis of incomplete block designs, mixed model approaches *must* be used.
- For incomplete block designs the ANOVA approach using expected mean squares becomes cumbersome as there are no simple rules anymore to derive EMS tables.
- Example: 4 treatments, replicated 3 times in blocks of size 3
- Each pair of treatments appears equally often (here: 2) times together in a block → Balanced Incomplete Block Design = BIBD
- Randomization procedure? Restrictions on randomization?

a	b	a	c
c	a	d	b
b	d	c	d

Recovery of inter block information

- If blocks are taken fixed, then all information about treatment differences stems from within the block → **intra block** analysis
- If blocks are taken random, then comparison between blocks also reveals information about treatment differences → **intra and inter block** analysis
- This last approach is known as **recovery of inter block information**.

Exercise 7: Recovery of inter block information

We follow roughly exercise 7.6.4 from Schabenberg & Pierce. It concerns data on corn yield from 13 corn hybrids.

a) Read the data into the program.

- Is the design a BIBD?
- How often is each treatment (=hybrid) replicated?
- How often is each treatment co-occurring with each other treatments within a block?

b) Fit the ANOVA model using fixed block effects.

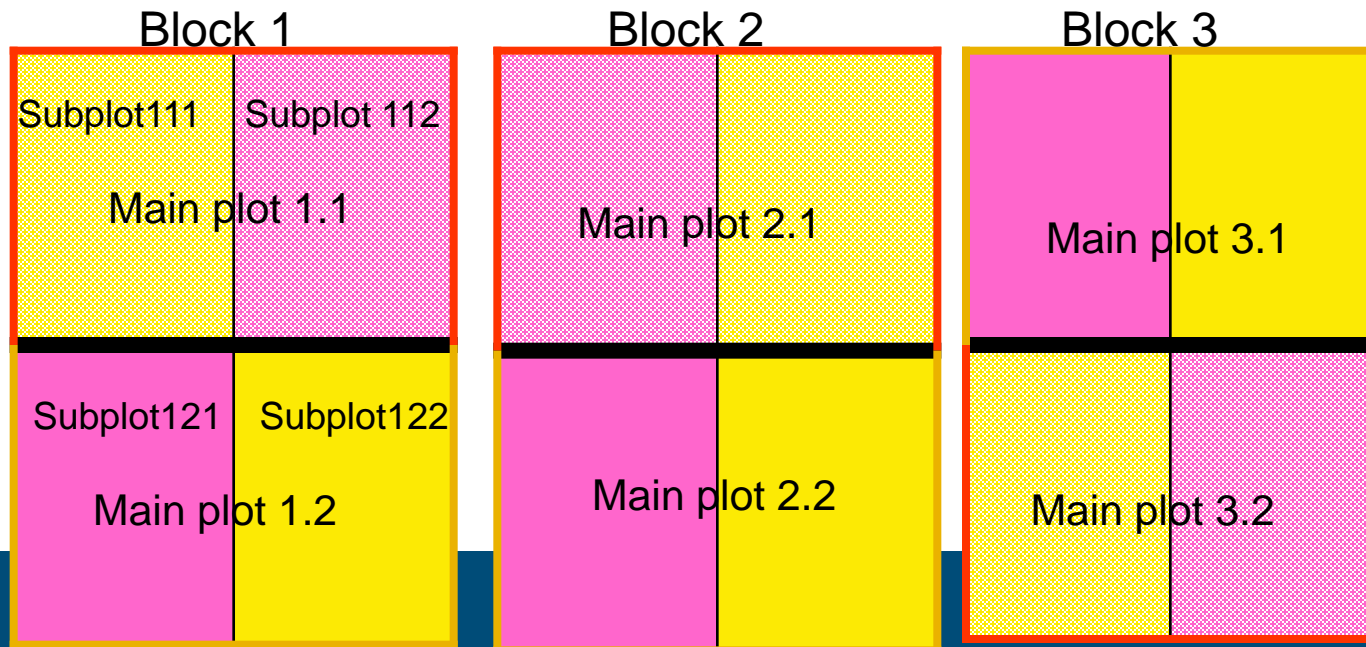
- Notice (and understand) the difference between emmeans (predicted means based upon the fitted model) and ordinary means.
- Pairwise comparisons between hybrids are obtained through `summary()` and `emmeans()` functions and by specifying user-defined contrasts using the `glht()` function.

c) Fit the ANOVA model using random block effects.

- Notice that s.e.d.'s (e.g. of difference hybrid 1 vs 2, and hybrid 2 vs 3) are smaller for the model with random block effects (compared to the model with fixed block effects). In other words, we are “recovering inter block information”.

Split-plot designs

- Split-plot design = factorial treatment design ($A \times B$) for which factor A is randomized over large plots (whole plots or main plots), whereas factor B is randomized over smaller plots within the main plots (subplots).
- Therefore: 2 types of experimental units
- Example: 3 blocks, 2 main plots per block, 2 subplots per main plot;
2 main plot treatments A1 and A2, 2 sub plot treatments B1 and B2



Mixed model for split-plot design

- Two treatment factors A and B
 - Whole plots, usually (but not always) arranged in randomized block design
 - Treatment factor A randomized over main plots
 - Treatment factor B randomized over subplots within main plots
 - Random effects of block and main plots

- How do matrices X and Z , vectors β and \underline{u} , and covariance matrices G and R look like?

Split-plot mixed model formulation

- $y_{ijk} = \mu + \rho_k + \alpha_i + \underline{d_{ik}} + \beta_j + \alpha\beta_{ij} + \underline{e_{ijk}}$
 - μ general mean
 - ρ_k block effect (random or fixed)
 - α_i main plot treatment effect
 - $\underline{d_{ik}}$ main plot error
 - β_j subplot treatment effect
 - $\alpha\beta_{ij}$ interaction of main plot and subplot treatment
 - $\underline{e_{ijk}}$ subplot error
- Main plot and subplot errors are assumed to be independent (randomization), normally distributed, with means 0 and variances σ_d^2 and σ_e^2 .
- Observations within whole plots are all assumed to be equally correlated (randomization).

EMS Table for Split Plot in Blocks (ANOVA framework)

- For balanced case expected mean squares can be obtained, from which correct F-tests may be formed.
- r blocks, a main plots per block (and a factor A levels) per block, b subplots per main plot (and b factor B levels)

Source	df	MS	EMS	F - tests
<i>blocks</i>	$r - 1$	$MS(blocks)$		
<i>A</i>	$a - 1$	$MS(A)$	$rb\theta_a^2 + b\sigma_d^2 + \sigma_e^2$	_____
<i>error1</i>	$(a - 1)(r - 1)$	$MS(main\ plots)$	$b\sigma_d^2 + \sigma_e^2$	← _____
<i>B</i>	$b - 1$	$MS(B)$	$ra\theta_b^2 + \sigma_e^2$	_____
<i>AB</i>	$(a - 1)(b - 1)$	$MS(AB)$	$r\theta_{ab}^2 + \sigma_e^2$	_____
<i>error2</i>	$a(r - 1)(b - 1)$	$MS(subplots)$	σ_e^2	← _____

Standard Error of Differences (SEDs)

- Between 2 main plot treatments

$$\sqrt{2MS(\text{main plot})/rb}$$

- Between 2 subplot treatments

$$\sqrt{2MS(\text{subplot})/ra}$$

- Between 2 main plot treatments at the same or different levels of subplot treatment (A1B1 vs A2B1 or A1B2 vs A2B1)

$$\sqrt{2((b-1)MS(\text{sub plot}) + MS(\text{main plot}))/rb}$$

- Between 2 subplot treatments at the same level of main plot treatment (A1B1 vs A1B2)

$$\sqrt{2MS(\text{subplot})/r}$$

- Interaction contrast ((A1B1–A1B2) – (A2B1–A2B2))

$$\sqrt{4MS(\text{subplot})/r}$$

Split plot in R: aov(), lme(), lmer()

```
> aov(y ~ a + b + a:b + Error(block / a))  
> lme(y ~ a + b + a:b, random= ~ 1 | block / a)  
> lmer(y ~ a + b + a:b + (1 | block / a) )
```

Exercise 8: Split plot (example from Kuehl, pg 487)

A split plot experiment was conducted on sorghum with two treatment factors, plant population density and hybrid. The aim was to assess the optimal combination of density and hybrid. The whole plots were used for the four levels of plant population density: 10, 15, 25, and 40 plants per meter per row. There were three hybrids randomly allocated to the subplots of each plot. The experiment was conducted in a randomized complete block design with four replications. The data that follow are the weights of the seeds per plant in grams.

	DENSITY	10	15	25	40
HYBRID	BLOCK				
1	1	40.7	24.2	16.1	11.2
	2	37.8	44.4	17.6	12.7
	3	32.9	27.8	19.9	14.5
	4	43.1	34.1	20.1	15.4
2	1	39.4	31.3	17.9	14.8
	2	47.8	34.5	30.5	17.3
	3	44.4	25.6	22.5	17.7
	4	49.0	50.4	25.2	18.7
3	1	68.7	26.2	20.5	18.9
	2	56.2	48.1	28.2	26.2
	3	44.8	41.1	30.0	19.2
	4	59.3	46.0	24.7	22.0

- a) Give a possible field lay-out for the experiment, in which you have (as good as possible) randomly allocated the blocks, main plot treatments and subplot treatments.

Exercise 8: Split plot (cont.)

- b) Write the linear model in scalar and matrix form and explain each of the terms. For the matrices you can indicate schematically how \mathbf{X} , $\boldsymbol{\beta}$, \mathbf{Z} , \mathbf{u} , \mathbf{e} , \mathbf{ZGZ}' , and \mathbf{R} will look like.
- c) Fit the appropriate mixed model.
- d) Assuming that an interaction exists between density and hybrid, what could be the interpretation of such an interaction?
- e) Test for interaction between density and hybrid at $\alpha=0.05$, and give your conclusion in words about the consequences of this test for the recommendation of the optimal combination of hybrid and density.
- f) Test for the main effects of density and hybrid at $\alpha=0.05$ and give your conclusions.
- g) Based on the results of the tests for main effects and interaction, what type of means and standard errors would you report as final results from your analysis?
- h) Produce the tables of means you suggested under g) with the appropriate standard errors of differences.
- i) Calculate the required standard errors of differences yourself using the estimates for the variance components produced by the mixed model analysis (see powerpoint for formulae).

Exercise 9: Split plot without blocks

Example taken from lecture notes “Variantiecomponenten” by Verdooren, page 252. Research was done on the effects of 3 types of diets on the early growth of the liver in 4 strains of mice. One suspects an interaction between diet and strain. From each of the strains a number of pregnant mouse females were chosen such that two litters with 6 offspring per litter were eventually available per strain for the experiment. From each litter 2 mice were assigned to each of the 3 diets. Measured was the liver weight as percentage of body weight.

Run the program “exercise9.r” and answer the following questions.

- a) Write down the statistical model in scalar and matrix form (schematically), including the assumptions for the error term(s), give schematic representations of the variance-covariance matrices V , G and R .

Exercise 9: Split plot without blocks (cont.)

- b) Investigate whether interaction between diet and strain can be found. Take care to use the right error term.
- c) Are main effects of strains significant? Take care to choose the right error term.
- d) Idem for diets? Take care to use the right error term.
- e) Is the average of diets 1 and 3 together different from that of diet 2 (averaged over strains)?

Exercise 10: analysis of data from a split-plot design

In this exercise we take a look at the data which were presented during the lecture. Yield of oats is related to two factors: *type of fertilizer* (with 3 levels) and *variety* (with 4 levels).

Factor fertilizer is randomized over nine fields, making field the experimental unit for factor fertilizer. Each fertilizer type is replicated 3 times (3 fields).

Per field 16 plots are available. Factor variety is randomized over plots within fields, making plot the experimental unit for factor variety. Per field each variety is replicated 4 times, giving in total $4 \times 9 = 36$ replicates per variety. In total there are $9 \text{ (fields)} \times 16 \text{ (plots per field)} = 144$ observations.

Open R-script exercise10.r in Rstudio.

- a) Read the data (in file exercise10.csv) into R and check that the dataset has 144 observations and 5 variables.
- b) Make sure that variables for fields, type of fertilizer and variety are treated as factors.

Exercise 10: analysis of data from a split-plot design, continued

- c) First do an Anova analysis without fields, including fertilizer type, variety and their interaction. In the R script you see that function `aov()` is used. Function `aov()` is based upon `lm()` which is meant for ordinary linear models, but it works a bit different. E.g. the function `summary()` applied to an object obtained from `aov()` gives an ANOVA table now. Show this ANOVA table. Explain what is incorrect in this analysis. What are the consequences for the P-values shown in the ANOVA table?
- d) Redo the split-plot analysis using `aov()` by including field in the model in 3 ways:
- Model 1: field as *random* term: model formula is `y ~ t + v + t:v + Error(field)`
 - Model 2: field as *fixed* term, entering the model *before* fertilizer `t : y ~ field + t + v + t:v`
 - Model 3: field as *fixed* term, entering the model *after* fertilizer `t : y ~ t + v + t:v + field`

Write down the F-values with degrees of freedom and P-values obtained from these three models in a cross table with the three models in the rows and the interaction `t:v` and main effects of `t` and `v` in the columns. Explain the differences that you see for `t`. Why are no results produced for `t` in model 2? Which model is needed for a correct analysis?

Exercise 10: analysis of data from a split-plot design, continued

- e) Calculate the nine mean yields per field (over the 4 varieties and replicates per variety), and do a one-way Anova using these means, testing for differences among fertilizer types. Compare results (F-test with P-value) with those obtained from the (correct) split-plot analysis. Regarding all results obtained above, it could be argued that the split-plot Anova combines two separate conventional ANOVA analyses. Which two analyses are these?

We continue the mixed model analysis, first with the Anova method (function `anova()`).

- f) Next we want to estimate and test the whole plot variance, which is the variance component for field. Convince yourself first that in the anova table obtained with `summary()` function the variance component for field (whole plot) is not available. We can however construct it from the bits and pieces in the anova table. Remember from class that $E(MS_{Field}) = \sigma_e^2 + 16\sigma_F^2$, so that $\sigma_F^2 = (E(MS_{Field}) - \sigma_e^2)/16$, to be estimated by $\hat{\sigma}_F^2 = (MS_{Field} - MSE)/16$. The F-statistic for testing $H_0: \sigma_F^2 = 0$ is $F = MS_{Field}/MSE$.
Now estimate σ_F^2 , compare it with the estimate of σ_E^2 and test $H_0: \sigma_F^2 = 0$. Conclusion?

Exercise 10: analysis of data from a split-plot design, continued

We continue the mixed model analysis, next with the REML method (function `lmer()` from the `lme4` package).

- g) The Anova method (`aov()`) for fitting mixed models is limited to balanced cases. A more general approach is the REML method, which was discussed in class.
- Load the packages `lme4` and `pbkr` needed for the REML method.
 - Rerun the split-plot analysis using the Anova method and run the split-plot analysis using the REML method and produce ANOVA tables for both. Notice that the degrees of freedom and F-statistics are identical in both analyses (for interaction and main effects). In the ANOVA table for the REML method no P-values are shown (package `lmerTest` is needed for that).
 - In the REML method the variance components can easily be obtained using function `VarCorr()`. Also locate the variance components in the output of the `summary()` function. Compare with the result you obtained in question 3f (using the Anova method).
 - Test $H_0: tv_{ij} = 0$ (no interaction) using the Kenward-Roger method. Conclusion?
 - If interaction is not significant, continue with tests for main effects. Test $H_0: t_i = 0$ (no fertilizer main effect) and $H_0: v_j = 0$ (no variety main effect). Two methods are shown: tests produced directly from the Full Model (using the Kenward-Roger method), and by removing the interaction from the model first (thereby changing the estimate of error variance).

Exercise 10: analysis of data from a split-plot design, continued

We arrive at the final part of the analysis: comparison of treatment means.

- h) Compare means using REML method. Function `emmeans()` can be used to obtain model-based (marginal) means with standard errors, confidence intervals and post-hoc tests for pairwise comparisons.
 - i. Fit again the mixed effects model with interaction using REML. Convince yourself (again) that the interaction of type of fertilizer with variety is unimportant. As before, remove this interaction from the model and fit the additive model (i.e. model without interaction).
 - ii. Have a look at the estimated variety means and their standard errors, the df and the confidence intervals. Also make a plot of the 95% confidence intervals. Can overlap between confidence intervals be interpreted as non-significance of the difference?
 - iii. Do the same for fertilizer means.
 - iv. Do post-hoc tests (pairwise comparisons) for factor variety and for factor type of fertilizer (without correction for multiple comparisons). Give the standard error of difference for variety and for fertilizer. Notice that one is *much larger* than the other? Which standard error of difference is largest? Why is this the case?
 - v. Make “compact letter displays” (cld) of the pairwise hypothesis testing results. Check that P-values from the pairwise comparisons and letter display results correspond.
 - vi. Use the tukey method for multiple comparisons for comparing varieties pairwise, both giving P-values and cld’s. Which pairwise comparisons, which were originally significant, are not significant anymore with the Tukey method?

Exercise 10: analysis of data from a split-plot design, continued

- vii. Check that the Anova method produces in this balanced situation identical means and standard error of means as the REML method.
- viii. Test the importance of the whole plot variance component (field) using a likelihood ratio test. Take the following steps:
 - Fit the additive models with and without the random effect (FM and RM)
 - Calculate as likelihood ratio test statistic twice the difference in log-likelihoods.
 - The P-value is the right-tail probability from the chi-square distribution with one d.f. *divided by two*.

[Note: function `rand()` from package `lmerTest` fails to divide by two.]

Compare the P-value with the result from the F-test in subquestion 3f.