Modelos Matemáticos e Aplicações

Módulo 2: Modelos lineares mistos

Linear Mixed Models

Some particular cases and respective application

2018-2019

Elsa Gonçalves
ISA/UL

Case 1

Random model (one factor of random effects), balanced, with G and R diagonal matrices

($G=\sigma_u^2 \, I_q$, $R=\sigma_e^2 \, I_n$)

One factor of random effects, balanced

$$Y_{ij} = \mu + u_i + e_{ij}$$

for $i = 1, \ldots, a, j = 1, \ldots, b$, $n = ab$.

Y_{ij} is the jth observation in the ith level of factor A;

μ is a general mean (population);

u_i is the effect of the level i of the factor A (random effects);

e_{ij} is the random error associated to the observation Y_{ij}.

- $u_i, i.i.d., \mathcal{N}\left(0, {\sigma^2}_u\right), \forall i$
- $e_{ij}, i.i.d., \mathcal{N}\left(0, {\sigma^2}_e\right), \forall ij$

The sums of squares are defined as in the case of fixed effects:

$$SQT = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(Y_{ij} - \bar{Y}_{..}\right)^2$$

$$SQA = \sum_{i=1}^{a} b\,(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

$$SQT = SQA + SQRE$$

- **Estimators for variance components:** procedure based on expected mean squares from the analysis of variance (ANOVA)

$$\mathrm{E}[SQA] = E\left[\sum_{i=1}^{a} b\,(\bar{Y}_{i.} - \bar{Y}_{..})^2\right] = (a-1)\left(b\sigma_u{}^2 + \sigma_e{}^2\right)$$

$$\mathrm{E}[QMA] = \frac{E[SQA]}{(a-1)} = b\sigma_u{}^2 + \sigma_e{}^2$$

$$E[SQRE] = E\left[\sum_{i=1}^{a}\sum_{j=1}^{b}\left(Y_{ij} - \bar{Y}_{i.}\right)^2\right] = a(b-1)\sigma_e{}^2$$

$$E[QMRE] = \frac{E[SQRE]}{a(b-1)} = \sigma_e{}^2$$

Equating sums of squares in their expected values, gives:

$$SQA = (a-1)\left(b\hat{\sigma}_u{}^2 + \hat{\sigma}_e{}^2\right)$$

$$SQRE = a(b-1)\hat{\sigma}_e{}^2$$

The estimators are:

$$\hat{\sigma}_e{}^2 = \frac{SQRE}{a(b-1)} = QMRE \qquad \hat{\sigma}_u{}^2 = \left(\frac{SQA}{a-1} - \hat{\sigma}_e{}^2\right)\Big/ b = \frac{QMA - QMRE}{b}$$

- **Maximum likelihood estimation**

For a model with one factor of random effects and balanced, the log-likelihood is given by:

$$l = \ln L = -\frac{1}{2} n \ln 2\pi - \frac{1}{2} a\left[\ln\left(\sigma_e^2 + b\sigma_u^2\right)\right] - \frac{1}{2} a(b-1)\ln \sigma_e^2 - \frac{\sum_i \sum_j \left(y_{ij} - \mu\right)^2}{2\sigma_e^2} + \frac{b^2 \sigma_u^2 \sum_i \left(\bar{y}_{i.} - \mu\right)^2}{2\sigma_e^2\left(\sigma_e^2 + b\sigma_u^2\right)}.$$

With some manipulation and rearranged so as to display SQA e SQRE (the sums of squares of ANOVA) and equating to zero the partial derivatives of lnL with respect to μ, ${\sigma_e}^2$ and ${\sigma_u}^2$, the following solutions are obtained:

$$\dot{\mu} = \bar{y}_{..}$$

$${\dot{\sigma}_e}^2 = QMRE$$

$${\dot{\sigma}_u}^2 = \frac{\left(1 - \frac{1}{a}\right) QMA - QMRE}{b}$$

These are the solutions to the maximum likelihood equations. But they are not necessarily the maximum likelihood estimators. It is necessary to verify if the matrix of second derivatives (Hessian matrix) is definite negative when the parameters in the Hessian are replaced by the solutions used. And ML estimators must be in the parameter space:

$$-\infty < \mu < +\infty, 0 < {\sigma_e}^2 < \infty, 0 \leq {\sigma_u}^2 < \infty$$

The maximum likelihood estimators for variance components are:

$$\hat{\sigma}_u^{\,2} = \frac{\left(1-\frac{1}{a}\right)QMA-QMRE}{b}, \quad \text{if} \quad \left(1-\frac{1}{a}\right)QMA \geq QMRE,$$

$$\hat{\sigma}_u^{\,2} = 0, \qquad \text{otherwise}$$

$$\hat{\sigma}_e^{\,2} = QMRE, \quad \text{if} \left(1-\frac{1}{a}\right)QMA \geq QMRE,$$

$$\hat{\sigma}_e^{\,2} = \frac{SQT}{ab}, \qquad \text{otherwise}$$

- **Restrited maximum likelihood estimation for variance components**

For a model with one factor of random effects and balanced, the restrited log-likelihood (l_R) is given by:

$$l_R = -\frac{1}{2}(ab-1)\ln 2\pi - \frac{1}{2}\ln ab - \frac{1}{2}a(b-1)\ln \sigma_e^2 - \frac{1}{2}(a-1)\ln \lambda - \frac{SQRE}{2\sigma_e^2} - \frac{SQA}{2\lambda}.$$

with $\lambda = \sigma_e^2 + b\sigma_u^2$

Equating to zero the partial derivatives of l_R with respect to $\sigma_e{}^2$ and $\sigma_u{}^2$, the following solutions are obtained:

$$\dot{\sigma}_e{}^2 = \frac{SQRE}{a(b-1)} = QMRE$$

$$\dot{\sigma}_u{}^2 = \frac{QMA - QMRE}{b}$$

The restricted maximum likelihood estimators for variance components are:

$$\begin{cases} \hat{\sigma}_u^{\,2} = \dfrac{QMA - QMRE}{b}, & \text{se } QMA \geq QMRE, \\ \hat{\sigma}_u^{\,2} = 0, & \text{caso contrário} \end{cases}$$

$$\begin{cases} \hat{\sigma}_e^{\,2} = QMRE, & \text{se } QMA \geq QMRE, \\ \hat{\sigma}_e^{\,2} = \dfrac{SQT}{ab-1}, & \text{caso contrário} \end{cases}$$

Asymptotic covariance matrix for REML estimators

$$\operatorname{var}\begin{bmatrix}\hat{\sigma}_e^{\,2}\\ \hat{\sigma}_u^{\,2}\end{bmatrix} \approx \begin{bmatrix}\dfrac{2\sigma_e^{\,4}}{a(b-1)} & \dfrac{-2\sigma_e^{\,4}}{ab(b-1)}\\ & \dfrac{2\sigma_e^{\,4}}{b^2}\left[\dfrac{(\sigma_e^{\,2}+b\sigma_u^{\,2})^2/\sigma_e^{\,4}}{a}+\dfrac{1}{a(b-1)}\right]\end{bmatrix}$$

ANOVA TABLE: random model with one factor of random effects (Factor A), balanced with $G=\sigma_u^2 I_q$, $R=\sigma_e^2 I_n$

$$Y_{ij} = \mu + u_i + e_{ij}$$

for $i = 1, \dots, a, j = 1, \dots, b$, $n = ab$.

	G.L.	S.Q.	QM	E[QM]	F
Factor A	$a-1$	$SQA = \sum_{i=1}^{a} b\,(\bar{y}_{i.} - \bar{y}_{..})^2$	$QMA = \frac{SQA}{a-1}$	$b\sigma_u^2 + \sigma_e^2$	$\frac{QMA}{QMRE}$
Resíduals	$a(b-1)$	$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.})^2$	$QMRE = \frac{SQRE}{a(b-1)}$	σ_e^2	
TOTAL	$ab-1$	$SQT = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$			

Hypothesis test for variance component associated to factor A

- Hypotheses: $H_0{:}\sigma_u^2 = 0 \quad vs \quad H_1{:}\,\sigma_u^2 > 0$
- Test statistic: $F = \frac{QMA}{QMRE} \cap \mathcal{F}_{(a-1,a(b-1))}$, under H_0
- Significance level: α
- Rejection region: upper (right-hand) tail

Reject H_0 if $F_{calc} > f_{\alpha(a-1,a(b-1))}$

Note:

$$\frac{SQRE}{{\sigma_e}^2} \sim {\chi^2}_{a(b-1)}$$

$$\frac{SQA}{b{\sigma_u}^2 + {\sigma_e}^2} \sim {\chi^2}_{a-1}$$

Independent random variables,

$$\frac{\frac{QMA}{b{\sigma_u}^2 + {\sigma_e}^2}}{\frac{QMRE}{{\sigma_e}^2}} \sim \mathcal{F}_{a-1,ab-a}$$

Likelihood ratio tests for variance component $\sigma_u{}^2$

- Hypotheses: $H_0: \sigma_u{}^2 = 0 \quad vs \quad H_1: \sigma_u{}^2 > 0$
- The REML likelihood ratio statistic :

$$\Lambda = 2\left(l_{R_1} - l_{R_0}\right) \sim \chi^2_\nu$$

being l_{R_1} the REML log-likelihood of the more general model (full model) and l_{R_0} the REML log-likelihood of the reduced model (that is, the REML log-likelihood under the null hypothesis). Under regularity conditions and under the null hypothesis, the likelihood ratio statistic, has an approximate χ^2_ν distribution with the degrees of freedom (ν) equal to the difference in the number of parameters between the two models. However, when we test a variance component, under the null hypothesis the parameter falls on the boundary of the parameter space. Theoretically it can be shown that for a single variance component, the asymptotic distribution of the REMLRT is a mixture of χ^2 variates , where the mixing probabilities are 0.5, one with 0 degrees of freedom and the other with one degree of freedom. As a consequence we can perform the likelihood ratio test as if the standard conditions apply, and divide the resulting p-value by two.

- The REML likelihood ratio test is only valid if the fixed effects are the same for both model.
- Significance level: α
- Rejection region: upper (right-hand) tail

Reject H_0 if $\Lambda_{calc} > \chi^2{}_{\alpha(\nu)}$

Example: for a random model with one factor of random effects, balanced (factor with a levels, b observations per level), the empirical best linear unbiased predictor of u_i (for the level i) is:

$$EBLUP(u_i) = \frac{b\hat{\sigma}^2{}_u}{b\hat{\sigma}^2{}_u + \hat{\sigma}^2{}_e}(\bar{Y}_{i.} - \bar{Y}_{..})$$

Exercise 1

In a grapevine selection study to evaluate the genetic variability of the yield of the Touriga Nacional variety, a field trial was installed in Vila Nova de Fozcoa, with a random sample of genotypes (196 genotypes) of the variety. In the field, each genotype was randomly assigned in 5 plots (trial with 5 replicates). The yield (kg/plant) data obtained in 1994 are available in *data.frame* touriga.

a) Describe the adequate model to study the yield genetic variability of the variety.

b) Use the command *aov* of R to obtain the ANOVA table of the model previously described.

(i) What are the variance components estimates involved in the model described in item a?

(ii) With the available information, carry out an hypotesis test for yield genetic variance of the variety (use a significance level of 0.05).

(iii) Knowing that $\bar{Y}..= 1.196$ kg/plant and $\bar{Y}_{c0101.} = 1.6044$ kg/plant, what is the empirical best linear unbiased predictor of yield genotypic effect of the genotype c0101?

Exercise 1 (cont.)

c) Fit the model previously described, with the restricted maximum likelihood (REML) method.

(i) Use *lme* of the package“nlme”, and *lmer* of the package “lme4”;

(ii) Apply command *summary* to the two objects above created and identify the REML estimates for variance componentes. Compare the results with those obtained in item b(i).

(iii) What is the yield fitted value for clone c0101 in repetition 2?

(iv) Explore commands *ranef* and *fitted* of packages “nmle” and “lme4”.

Exercise 1 (cont.)

d) In fact, the Touriga Nacional field trial above described was planted according to a randomized complete block design (5 blocks).

(i) Fit a new model considering the block effect (assuming a random effects factor). Use package *lme4*.

(ii) Carry out hypotesis tests for variance componets of the model.

(iii) Compute AIC and BIC for both fitted models and select the best one according to those criteria.

Case 2

Linear mixed model: one factor of fixed effects, one factor of random effects, balanced, without interaction and with interaction, with G and R diagonal matrices (*G*=$\sigma_u{}^2\ \boldsymbol{I_q}$, *R*=$\sigma_e{}^2\ \boldsymbol{I_n}$)

Linear mixed model: one factor of fixed effects (factor A), one factor of random effects (factor B), balanced, without interaction

$$Y_{ijk} = \mu_1 + \beta_i + u_j + e_{ijk}$$

for $i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, c, n = abc$, with $\beta_1 = 0$.

Y_{ijk} is the observation in the ith level of factor A and jth level of factor B;

μ_1 is a general mean (population) in the level 1 of factor A;

β_i is the effect of the level i of the factor A ((the increased concerning to μ_1), fixed;

u_j is the effect of the level j of the factor B, random;

e_{ijk} is the random error associated to the observation Y_{ijk} .

- $u_j, i.i.d., \mathcal{N}\left(0, {\sigma^2}_u\right), \forall j$
- $e_{ijk}, i.i.d., \mathcal{N}\left(0, {\sigma^2}_e\right), \forall ijk$

The sums of squares are defined as :

$$SQT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - \bar{Y}_{...}\right)^2$$

$$SQA = \sum_{i=1}^{a} bc\left(\bar{Y}_{i..} - \bar{Y}_{..,}\right)^2$$

$$SQB = \sum_{j=1}^{b} ac\left(\bar{Y}_{.j.} - \bar{Y}_{...,}\right)^2$$

$$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...,}\right)^2$$

$$SQT = SQA + SQB + SQRE$$

- **Estimators for variance components:** procedure based on expected mean squares from the analysis of variance (ANOVA)

$$\mathrm{E}[SQB] = (b-1)\left(ac\sigma_u{}^2 + \sigma_e{}^2\right)$$

$$\mathrm{E}[QMB] = \frac{E[SQB]}{(b-1)} = ac\sigma_u{}^2 + \sigma_e{}^2$$

$$E[SQRE] = n - (a+b-1)\sigma_e{}^2$$

$$E[QMRE] = \frac{E[SQRE]}{n-(a+b-1)} = \sigma_e{}^2$$

The estimators are:

$$\hat{\sigma}_e{}^2 = \frac{SQRE}{n-(a+b-1)} = QMRE \qquad \hat{\sigma}_u{}^2 = \frac{QMB - QMRE}{ac}$$

- **The maximum likelihood estimators for variance components are($\hat{\sigma}_u^{\,2} \geq 0$)**

$$\hat{\sigma}_e^{\,2} = \left[1 - \frac{a-1}{b(ac-1)}\right] QMRE,$$

$$\hat{\sigma}_u^{\,2} = \frac{SQB/b - \hat{\sigma}_e^{\,2}}{ac}$$

- **The restricted maximum likelihood estimators for variance components are ($\hat{\sigma}_u^{\,2} \geq 0$):**

$$\hat{\sigma}_e^{\,2} = \frac{SQRE}{n-(a+b-1)} = QMRE \qquad \hat{\sigma}_u^{\,2} = \frac{QMB - QMRE}{ac}$$

Asymptotic variance matrix for REML estimators

$$\text{var}\begin{bmatrix}\hat{\sigma}_e{}^2\\ \hat{\sigma}_u{}^2\end{bmatrix} \approx \frac{2\sigma_e{}^4}{b(ac-1)}\begin{bmatrix}1 & \frac{-1}{ac}\\ \frac{-1}{ac} & \left[\frac{1+(ac-1)(1+ac\sigma_u{}^2/\sigma_e{}^2)^2}{a^2c^2}\right]\end{bmatrix}$$

ANOVA TABLE: linear mixed model, one factor of fixed effects (Factor A) and one factor of random effects (factor B), balanced, with $G=\sigma_u^2 I_q$, $R=\sigma_e^2 I_n$

$$Y_{ijk} = \mu_1 + \beta_i + u_j + e_{ijk}$$

for $i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, c, n = abc$, with $\beta_1 = 0$.

	G.L.	S.Q.	QM	E[QM]	F
Factor A	$a-1$	SQA	QMA	$\frac{bc}{a-1}\sum_{i=1}^{a}(\beta_i - \bar{\beta}_.)^2 + \sigma_e^2$	$\frac{QMA}{QMRE}$
Factor B	$b-1$	SQB	QMB	$\sigma_e^2 + ac\sigma_u^2$	$\frac{QMB}{QMRE}$
Residuals	$n-(a+b-1)$	$SQRE$	$QMRE$	σ_e^2	
TOTAL	$n-1$	SQT			

Hypothesis test for variance component associated to factor B

- Hypotheses : $H_0: \sigma_u^2 = 0 \quad vs \quad H_1: \sigma_u^2 > 0$
- Test statistic : $F = \frac{QMB}{QMRE} \cap F_{(b-1,n-(a+b-1))}$, under H_0
- Significance level : α
- Rejection region : upper (right-hand) tail

Rejeitar H_0 se $F_{calc} > f_{\alpha(b-1,n-(a+b-1)))}$

Or, a **Likelihood ratio test for the variance component.**

Hypothesis test for fixed effects

- Hypotheses $H_0: \beta_i = 0, \quad \forall_{i=2,\ldots a} \quad vs \quad H_1: \exists_{i=2,\ldots,a} : \beta_i \neq 0$
- Test statistic : $F = \frac{QMA}{QMRE} \cap F_{(a-1,n-(a+b-1)\ )}$, sob H_0
- Significance level : α
- Rejection region : upper (right-hand) tail

Rejeitar H_0 se $F_{calc} > f_{\alpha(a-1,n-(a+b-1)))}$

Note: the test for fixed effects is identical to what was described in the context of fixed effects ANOVA

Exercise 3

Consider the data *data.frame* terrenos. The objective of the study is to compare the yield among four wheat varieties. In addition, 13 sites with different soil conditions were identified. Consider that those sites are a random sample of the sites where the four varieties of wheat will be grown. The four varieties were assigned at random within sites, each variety once per site.

a) Fit the adequate model for this study (for example, using *package nlme*, function *lme*).

ei) Carry out the hypothesis test for fixed effects of the model. For the calculation of the test statistic recall the hypothesis tests for linear combinations of fixed effects of the linear mixed model given in the theoretical classes. Consider the estimated covariance matrix of the fixed effects estimators (vcov (terrenolme1)), define the matrix L, create the vector with the fixed effects estimates and, with the help of R, compute the test statistic. For your conclusions, use the significance level of 0.05. At the end, run anova (terrenolme1).

eii) Is there a decrease in yield of variety B compared to variety A (for $\alpha = 0.05$)?

Linear mixed model: one factor of fixed effects (factor A), one factor with random effects (factor B), balanced, with interaction

$$Y_{ijk} = \mu_1 + \beta_i + u_j + (\beta u)_{ij} + e_{ijk}$$

for $i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, c, n = abc$, with $\beta_1 = 0$.

Y_{ijk} is the kth observation in the ith level of factor A and jth level of factor B;

μ_1 is a general mean (population) in the level 1 of factor A;

β_i is the effect of the level i of the factor A (the increased concerning to μ_1), fixed;

u_j is the effect of the level j of the factor B, random;

$(\beta u)_{ij}$ is the interaction effect of the ith level of factor A with the jth level of factor B, random;

e_{ijk} is the random error associated to the observation Y_{ijk}.

- $u_j, i.i.d., \mathcal{N}\left(0, \sigma^2{}_u\right), \forall j$
- $(\beta u)_{ij}, i.i.d., \mathcal{N}\left(0, \sigma^2{}_{\beta u}\right), \forall ij$
- $e_{ijk}, i.i.d., \mathcal{N}\left(0, \sigma^2{}_e\right), \forall ijk$

The sums of squares are defined as :

$$SQT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - \bar{Y}_{...}\right)^2$$

$$SQA = \sum_{i=1}^{a} bc\,(\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$SQB = \sum_{j=1}^{b} ac\left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2$$

$$SQAB = \sum_{i=1}^{a}\sum_{j=1}^{b} c\left(Y_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}\right)^2$$

$$SQRE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - \bar{Y}_{ij.}\right)^2$$

$$SQT = SQA + SQB + SQAB + SQRE$$

- **Estimators for variance components:** procedure based on expected mean squares from the analysis of variance (ANOVA)

$$\mathrm{E}[SQB] = (b-1)\left(ac\sigma_u{}^2 + c\sigma_{\beta u}{}^2 + \sigma_e{}^2\right)$$

$$\mathrm{E}[QMB] = \frac{E[SQB]}{(b-1)} = ac\sigma_u{}^2 + c\sigma_{\beta u}{}^2 + \sigma_e{}^2$$

$$\mathrm{E}[SQAB] = (a-1)(b-1)\left(c\sigma_{\beta u}{}^2 + \sigma_e{}^2\right)$$

$$\mathrm{E}[QMAB] = \frac{E[SQAB]}{(a-1)(b-1)} = c\sigma_{\beta u}{}^2 + \sigma_e{}^2$$

$$E[SQRE] = ab(c-1)\sigma_e{}^2$$

$$E[QMRE] = \frac{E[SQRE]}{ab(c-1)} = \sigma_e{}^2$$

These yield the estimators

$$\hat{\sigma}_e{}^2 = QMRE$$

$$\hat{\sigma}_{\beta u}{}^2 = \frac{QMAB - QMRE}{c}$$

$$\hat{\sigma}_u{}^2 = \frac{QMB - QMAB}{ac}$$

- **The maximum likelihood estimators for variance components are ($\hat{\sigma}_u^2 \geq 0, \hat{\sigma}_{\beta u}^2 \geq 0$)**

$$\hat{\sigma}_e^2 = QMRE$$

$$\hat{\sigma}_u^2 = \frac{(1-\frac{1}{b})(QMB-QMAB)}{ac}$$

$$\hat{\sigma}_{\beta u}^2 = \frac{(1-\frac{1}{b})QMAB - QMRE}{c}$$

- **The restricted maximum likelihood estimators for variance components are ($\hat{\sigma}_u^2 \geq 0, \hat{\sigma}_{\beta u}^2 \geq 0$)**

$$\hat{\sigma}_e^2 = QMRE$$

$$\hat{\sigma}_{\beta u}^2 = \frac{QMAB - QMRE}{c}$$

$$\hat{\sigma}_u^2 = \frac{QMB - QMAB}{ac}$$

Asymptotic variance matrix for REML estimators

$$\operatorname{var}\begin{bmatrix} \hat{\sigma}_e^{\,2} \\ \hat{\sigma}_u^{\,2} \\ \hat{\sigma}_{\beta u}^{\,2} \end{bmatrix} \approx \frac{2}{b}\begin{bmatrix} \dfrac{\sigma_e^{\,4}}{a(c-1)} & 0 & \dfrac{-\sigma_e^{\,4}}{ac(c-1)} \\ & \dfrac{(\sigma_e^{\,2}+c\sigma_{\beta u}^{\,2})^2}{a-1}+\dfrac{(\sigma_e^{\,2}+c\sigma_{\beta u}^{\,2}+ac\sigma_u^{\,2})^2}{a^2c^2} & \dfrac{-(\sigma_e^{\,2}+c\sigma_{\beta u}^{\,2})^2}{ac^2(a-1)} \\ & & \dfrac{1}{c^2}\left[\dfrac{(\sigma_e^{\,2}+c\sigma_\beta^{\,2})^2}{a-1}+\dfrac{\sigma_e^{\,4}}{a(c-1)}\right] \end{bmatrix}$$

ANOVA TABLE: linear mixed model, one factor of fixed effects (factor A) and one factor of random effects (factor B), balanced, with interaction

$$Y_{ijk} = \mu_1 + \beta_i + u_j + (\beta u)_{ij} + e_{ijk}$$

for $i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, c, n = abc$, with $\beta_1 = 0$.

- $u_j, i.i.d., \mathcal{N}\left(0, {\sigma^2}_u\right), \forall j; (\beta u)_{ij}, i.i.d., \mathcal{N}\left(0, {\sigma^2}_{\beta u}\right), \forall ij; e_{ijk}, i.i.d., \mathcal{N}\left(0, {\sigma^2}_e\right), \forall ijk$

	G.L.	S.Q.	QM	E[QM]	F
Factor A	$a-1$	SQA	QMA	$\frac{bc}{a-1}\sum_{i=1}^{a}\left(\beta_i - \bar{\beta}_.\right)^2 + {\sigma_e}^2 + c{\sigma_{\beta u}}^2$	$\frac{QMA}{QMAB}$
Factor B	$b-1$	SQB	QMB	${\sigma_e}^2 + c{\sigma_{\beta u}}^2 + ca{\sigma_u}^2$	$\frac{QMB}{QMAB}$
Interaction	$(a-1)(b-1)$	$SQAB$	$QMAB$	${\sigma_e}^2 + c{\sigma_{\beta u}}^2$	$\frac{QMAB}{QMRE}$
Resíduals	$ab(c-1)$	$SQRE$	$QMRE$	${\sigma_e}^2$	
TOTAL	$n-1$	SQT			

Hypothesis test for variance component associated to interaction

- Hypotheses: $H_0: \sigma^2_{\beta u} = 0 \quad vs \quad H_1 \sigma^2_{\beta u} > 0$
- Test statistic: $F = \frac{QMAB}{QMRE} \cap \mathcal{F}_{\left((a-1)(b-1),ab(c-1)\right)}$, under H_0
- Significance level: α
- Rejection region: upper (right-hand) tail

Reject H_0 if $F_{calc} > f_{\alpha\left((a-1)(b-1),ab(c-1)\right)}$

Or, a **Likelihood ratio test.**

Hypothesis test for variance component associated to factor B

- Hypotheses: $H_0: \sigma_u^2 = 0 \quad vs \quad H_1 \sigma_u^2 > 0$
- Test statistic: $F = \frac{QMB}{QMAB} \cap \mathcal{F}_{(b-1,(a-1)(b-1))}$, under H_0
- Significance level: α
- Rejection region: upper (right-hand) tail

Reject H_0 if $F_{calc} > f_{\alpha(b-1,(a-1)(b-1))}$

Or, a **Likelihood ratio test.**

Tests of hypotheses , fixed effects of factor A

- Hypotheses: $H_0: \beta_i = 0, \quad \forall_{i=2,\ldots a} \quad vs \quad H_1: \exists_{i=2,\ldots,a}$ tal que $\beta_i \neq 0$
- Test statistic: $F = \frac{QMA}{QMAB} \cap F_{(a-1,(a-1)(b-1))}$, sob H_0
- Significance level: α
- Rejection region: upper (right-hand) tail: unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha(a-1,(a-1)(b-1))}$

Some considerations

- For random or complex mixed models there are no exact statistical tests for certain model effects (the numerator and denominator of the F statistics are linear combinations of mean squares). In these cases, approximate F tests are performed. One of the classic methods most used for this approach is the method of Satterthwaite (1941). However, other methods are implemented in more complex mixed models frequently reported in the literature and commonly used in several packages, for example, the methods of Giesbrecht and Burns (1985) and Kenward and Roger (1997). (next slide, additional information)

❑ Example: Satterthwaite Degrees of freedom Approximation

Satterthwaite showed that given the ratio

$$\frac{X_{num}^2/\nu_1}{X_2^*/\upsilon_2^*}$$

where $X_{num}^2 \cap \chi_{\nu_1}^2$ and X_2^* is a linear combination of chi-square random variable all independent of X_{num}^2, the $X_2^* \cap \chi_{\upsilon_2^*}^2$, where

$$\upsilon_2^* \cong \frac{\left(\sum_m l_m X_m^2\right)^2}{\sum_m (l_m X_m^2)/df_m},$$

X_m^2 denotes the $\chi_{df_m}^2$ random variables, l_m denote the constants in the linear combination, df_m the degrees of freedom for the respective X_m^2.

Note

- There are no exact confidence intervals for the variance components associated with the random effects of the model (the distribution of the estimator of variance components is a linear combination of chi-square random variables, remember these estimators for the classic cases in slides 5, 21, 30).

Exercise 4

In *library nlme* and *lme4* of R is available the data set *Machines* (Pinheiro e Bates, 2000). The objective of the experiment is to compare three brands of machines used in an industrial process. Six workers were chosen randomly among the employees of a factory to operate each machine three times. The response variable is an overall productuvity score taking into account the number and quality of components produced.

a) Describe the appropriate model for this study. Fit the model using R, function *lmer* of package *lme4*. Use the commands plot.design (Machines) and interaction.plot (Machine,Worker,score) and comment.

b) What are the restrict maximum likelihood estimates for variance components of the model?

c) Would the values of the variance components estimates obtained by the maximum likelihood method be higher or lower than the estimates given in the previous item ?

Exercise 4 (cont.)

d) Carry out the hypothesis test for worker×machine interaction. Use a significance level of 0.01.

e) Carry out the hypothesis test for the variability associated to worker. Use a significance level of 0.01.

f) What are the values of the fixed effects estimates of the model? Explain the meaning of those estimates.

g) Carry out an appropriate hypothesis test to assess if there are any major effects associated with machine brands. Use a significance level of 0.01.

Exercise 4 (cont.)

Note: use the commands and comment the results.

```
plot(machines1r)
residuos<-resid(machines1r)
qqnorm(residuos)
eblupsworker<-ranef(machines1r)$Worker
qqnorm(eblupsworker[,1])
eblupsinteraccao<-ranef(machines1r)$`Worker:Machine`
qqnorm(eblupsinteraccao[,1])
```

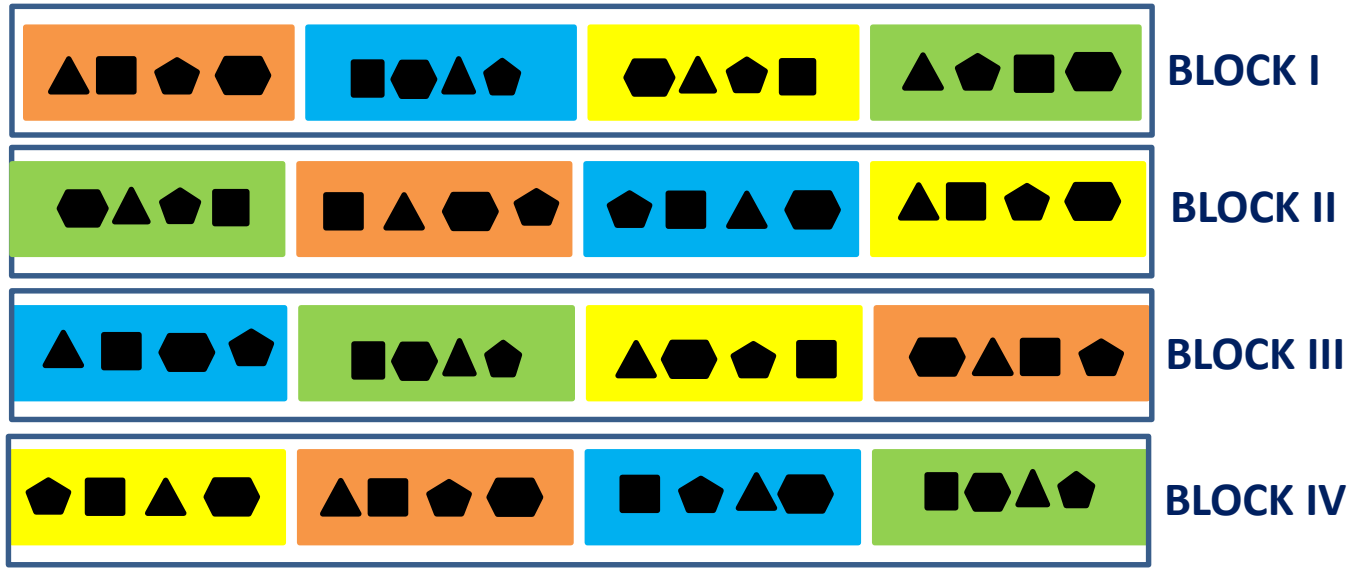
Case 3

Linear mixed models for analysis of split plot experiments

The split-plot design on a RCB

• Main treatments (levels of factor A) are assigned at random within blocks, each treatment once per block; they are divided further into additional independent units (subplots) to which another set of treatments (levels of factor B) are randomly assigned.
• The number of blocks is the number of replications.
• Any main treatment can be adjacent to any other treatment, but not to the same treatment within the block.

Example:
Different colors represent different main treatments (levels of factor A) ; each row represents a block. There are 4 blocks (I-IV) each of 4 main treatments (colors) divided into 4 additional independent units (subplots) to which another set of treatments (levels of factor B, symbols) are randomly assigned.



Considering two factors with fixed effects (factors A and B) and random blocks. The model can be described as:

$$Y_{ijk} = \mu_{11} + \alpha_i + u_j + (\alpha u)_{ij} + \beta_k + (\alpha\beta)_{ik} + (\beta u)_{kj} + e_{ijk}$$

with $i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, c, n = abc,$
and $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1k} = 0, \forall_k,$ $(\alpha\beta)_{i1} = 0, \forall_i.$
Where:

Y_{ijk}, is the observation from ith level of factor *A* (*whole-plot i*), block *j*, and k^{th} level of factor *B* (*sub-plot* or *split-plot k*);

μ_{11}, is the general mean (population) in the level 1 of factor *A* with level 1 of factor *B*;

α_i, is the effect of the level *i* of the factor *A* (increase), assigned to whole-plot (fixed);

u_j, is the effect of block *j* (random);

$(\alpha u)_{ij}$, is the interaction effect of the ith level of factor *A* with block *j*, named as *whole-plot error* (random);

β_k, is the effect of the level *k* of the factor *B* (increase), assigned to *sub-plot* (fixed);

$(\alpha\beta)_{ik}$, is the interaction effect of the ith level of factor *A* with the kth level of factor *B* (increase) (fixed);

$(\beta u)_{kj}$, is the interaction effect of the kth level of factor *B* with block *j* (random);

e_{ijk}, is the random error associated to the observation Y_{ijk}.

In the common approach, the effect $(\beta u)_{jk}$ is set to zero (thus, $(\beta u)_{jk}$ is incorporated in e_{ijk}). The random error includes $(\beta u)_{jk}$ and $(\alpha\beta u)_{ijk}$, and is called as *within plot error*.

Therefore, the common assumptions are:

$u_j, i.i.d., \mathcal{N}\left(0, \sigma^2{}_u\right), \forall j; (\alpha u)_{ij}, i.i.d., \mathcal{N}\left(0, \sigma^2{}_{\alpha u}\right), \forall ij;$

$e_{ijk}, i.i.d., \mathcal{N}\left(0, \sigma^2{}_e\right), \forall ijk; Cov\left(u_j, (\alpha u)_{ij}\right) = 0; Cov\left(u_j, e_{ijk}\right) = 0;$

$Cov\left((\alpha u)_{ij}, e_{ijk}\right) = 0.$

ANOVA TABLE, considering a balance design:

	G.L.	S.Q.	QM	E[QM]	F
Factor A	$a-1$	SQA	QMA	$c\sigma_{\alpha u}{}^2 + \sigma_e{}^2 + bc\frac{\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}_.)^2}{a-1}$	$\frac{QMA}{\text{QMWError}}$
Block	$b-1$	SQBL	QMBL	$ac\sigma_u{}^2\ c\sigma_{\alpha u}{}^2 + \sigma_e{}^2$	
Interaction FactorA×Block (*Whole-plot error*)	$(a-1)(b-1)$	SQWError	QMWError	$c\sigma_{\alpha u}{}^2 + \sigma_e{}^2$	
Factor B	$c-1$	SQB	QMB	$\sigma_e{}^2 + ab\frac{\sum_{k=1}^{c}(\beta_k - \bar{\beta}_.)^2}{c-1}$	$\frac{QMB}{QMRE}$
Interaction FactorA×FactorB	$(a-1)(c-1)$	$SQAB$	$QMAB$	$\sigma_e{}^2 + b\frac{\sum_{i=1}^{a}\sum_{k=1}^{c}(\alpha\beta_{ik} - \overline{\alpha\beta}_{..})^2}{(a-1)(c-1)}$	$\frac{QMAB}{QMRE}$
Residuals (*Within plot error*)	$a(b-1)(c-1)$	$SQRE$	$QMRE$	$\sigma_e{}^2$	

Exercise 6

In *package* "*nlme*" do R, there is a data set named as "*Alfalfa*".

```
> head(Alfalfa)
Grouped Data: Yield ~ Date | Block/Variety
  Variety Date Block Yield
1   Ladak None     1  2.17
2   Ladak   S1     1  1.58
3   Ladak  S20     1  2.29
4   Ladak   O7     1  2.23
5   Ladak None     2  1.88
6   Ladak   S1     2  1.26
      …
```

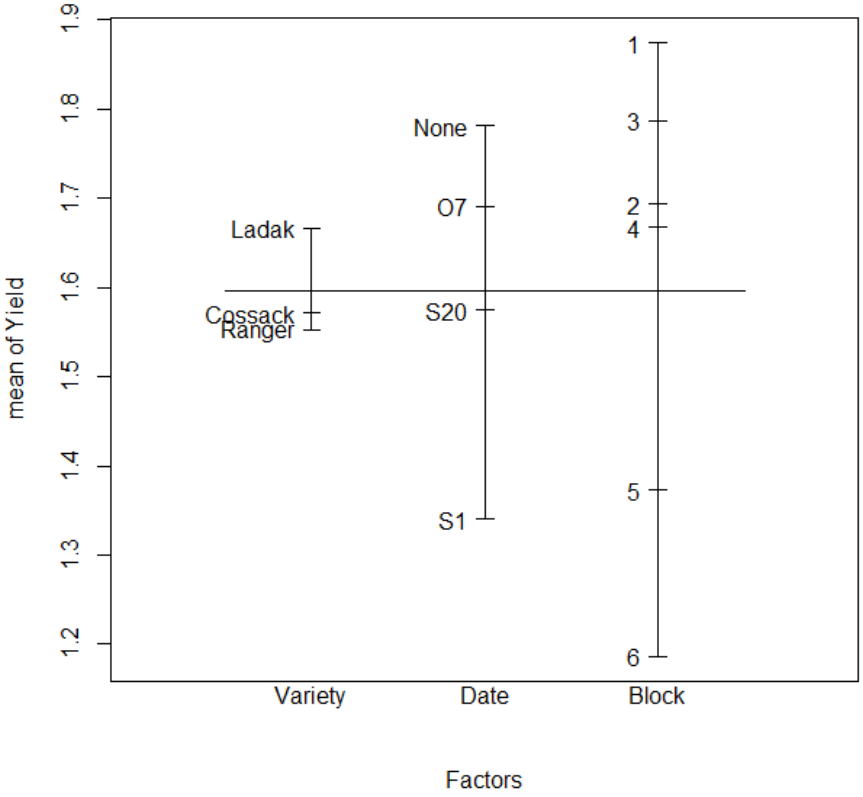
Exercise 6

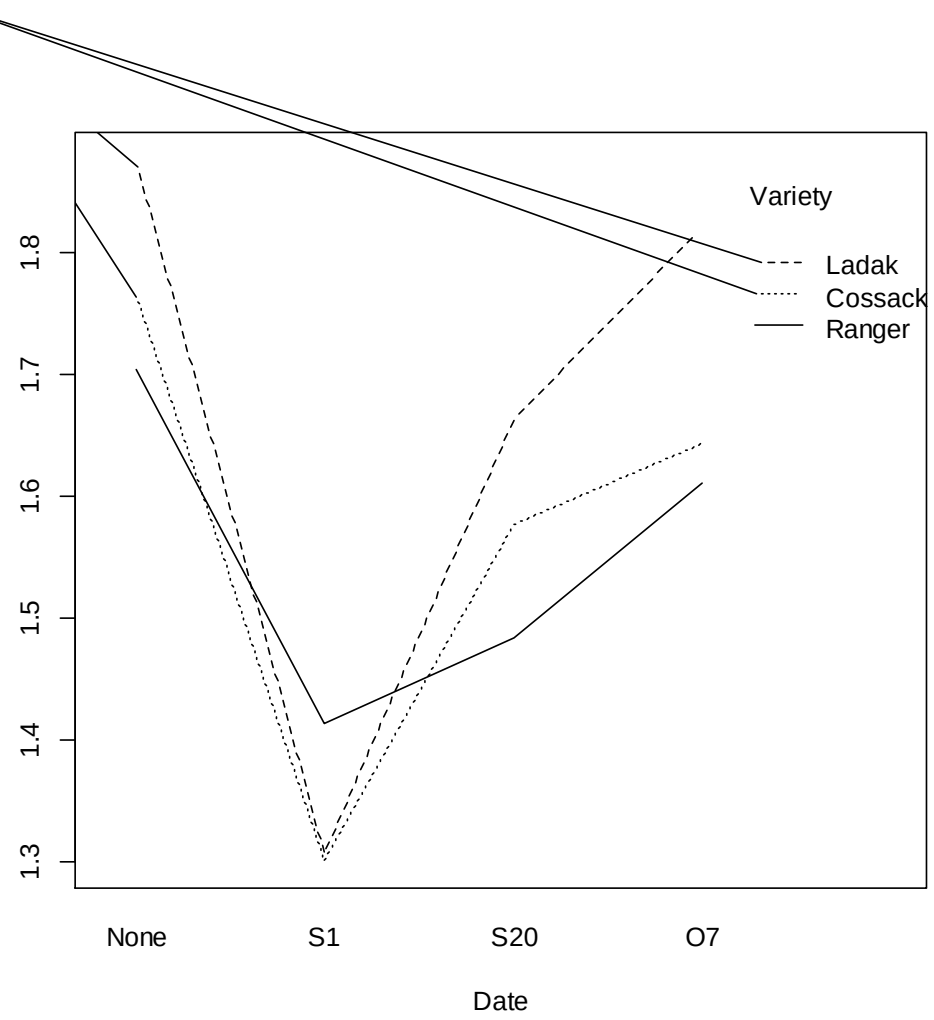
These data are described in Snedecor & Cochran (1980) as an example of a *split-plot design* (Pinheiro and Bates, 2000). The objective is to study if the yield (T/acre) of alfalfa (*Medicago sativa*) is afected by variety and date of third cutting. Therefore, there are two factors: variety of alfalfa, with 3 levels (*Cossac*, *Ladak* e *Ranger*) and date of third cutting, with 4 levels (*none*–sem corte, *S1*–Sep1; *S20* – Sep20; and *O7* – Oct7). The treatment structure used in the experiment was a 3×4 full factorial. The experimental units were arranged into 6 blocks, each block was divided into 3 plots (*whole plots; whole plot, largest experimental unit)*, where the varieties of alfalfa were randomly assigned; and each whole plot was divided into four subplots (split plots), where the dates of third cutting were randomly assigned.

a) Describe the appropriate model for this study.

Exercise 6 (cont.)

b) Plot the data using *plot.design (Alfalfa)* and *interaction.plot (Date, Variety, Yield)*. Comment.





Exercise 6 (cont.)

c) Fit the model described in item a) in R using *lmer* of package "*lme4*".

d) Carry out the hypothesis tests that answer the objectives of the study.

e) Compare the previous results with those obtained with the command
"*aov(Yield~Date*Variety+Error(Block*Variety), data=Alfalfa)*".

Some considerations

(1) Factors A and B with random effects;

(2) Factor A with fixed effects and factor B with random effects.

(1) ANOVA table: factors A and B with random effects**,** balanced:

	G.L.	QM	E[QM]	F
Factor A	$a-1$	QMA	$bc\sigma_{\alpha}{}^2 + c\sigma_{\alpha u}{}^2 + b\sigma_{\alpha\beta}{}^2 + \sigma_e{}^2$	$\frac{QMA+QMRE}{\text{QMWError}+QMAB}^*$
Block	$b-1$	$QMBL$	$ac\sigma_u{}^2\, c\sigma_{\alpha u}{}^2 + \sigma_e{}^2$	
Interaction FactorA×Block (*Whole-plot error*)	$(a-1)(b-1)$	$QMWError$	$c\sigma_{\alpha u}{}^2 + \sigma_e{}^2$	
Factor B	$c-1$	QMB	$ab\sigma_{\beta}{}^2 + b\sigma_{\alpha\beta}{}^2 + \sigma_e{}^2$	$\frac{QMB}{QMAB}$
Interaction FactorA×FactorB	$(a-1)(c-1)$	$QMAB$	$b\sigma_{\alpha\beta}{}^2 + \sigma_e{}^2$	$\frac{QMAB}{QMRE}$
Residuals	$a(b-1)(c-1)$	$QMRE$	$\sigma_e{}^2$	

*Approximate degrees of freedom. For example, *Satterthwaite* method:

$$\nu_1 = \frac{(QMA+QMRE)^2}{\frac{(QMA)^2}{a-1}+\frac{(QMRE)^2}{a(b-1)(c-1)}},\ \nu_2 = \frac{(\text{QMWError}+QMAB)^2}{\frac{(\text{QMWError})^2}{(a-1)(b-1)}+\frac{(QMAB)^2}{(a-1)(c-1)}}$$

(2) ANOVA table: factor A with fixed effects and factor B with random effects**,** balanced:

	G.L.	QM	E[QM]	F
Factor A	$a-1$	QMA	$c{\sigma_{\alpha u}}^2 + b\frac{a}{a-1}{\sigma_{\alpha\beta}}^2 + {\sigma_e}^2 + bc\frac{\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}_.)^2}{a-1}$	$\frac{QMA+QMRE}{\mathrm{QMWError}+QMAB}^*$
Block	$b-1$	$QMBL$	$ac{\sigma_u}^2\ c{\sigma_{\alpha u}}^2 + {\sigma_e}^2$	
Interaction FactorA×Block (*Whole-plot error*)	$(a-1)(b-1)$	$QMWError$	$c{\sigma_{\alpha u}}^2 + {\sigma_e}^2$	
Factor B	$c-1$	QMB	$ab{\sigma_\beta}^2 + {\sigma_e}^2$	$\frac{QMB}{QMRE}$
Interaction FactorA×FactorB	$(a-1)(c-1)$	$QMAB$	$b\frac{a}{a-1}{\sigma_{\alpha\beta}}^2 + {\sigma_e}^2$	$\frac{QMAB}{QMRE}$
Residuals	$a(b-1)(c-1)$	$QMRE$	${\sigma_e}^2$	

*Approximate degrees of freedom. For example, *Satterthwaite* method:

$$\nu_1 = \frac{(QMA+QMRE)^2}{\frac{(QMA)^2}{a-1}+\frac{(QMRE)^2}{a(b-1)(c-1)}},\ \nu_2 = \frac{(\mathrm{QMWError}+QMAB)^2}{\frac{(\mathrm{QMWError})^2}{(a-1)(b-1)}+\frac{(QMAB)^2}{(a-1)(c-1)}}$$

Case 4

The following is an example of the application of linear mixed models with categorical and numerical predictor variables (covariance analysis) and in which the observations are made in the same individual over time*

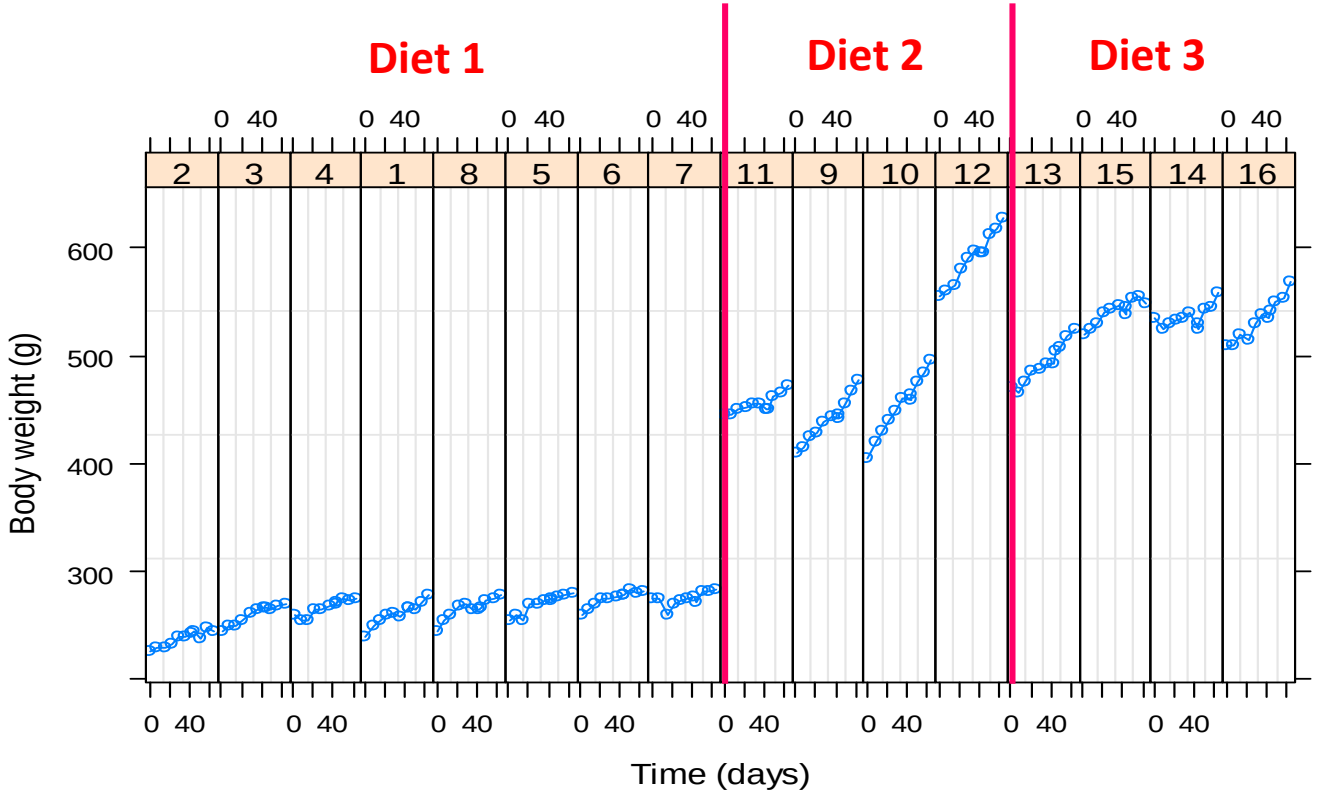
* The correlation matrices used for this type of analysis are used in time series and spatial statistics. For its understanding would be necessary theoretical bases on time series and spatial statistics, which is not part of this UC. Therefore, we will only exemplify its application, so that it is recorded that these instruments are currently widely used in mixed models context.

Exercise 8

Data set *BodyWeight* (Pinheiro e Bates, 2000) is available in R, and is related to the body weights of rats measured over 64 days. The body weights of the rats (in grams) are measured on day 1 and every seven days thereafter until day 24, with an extra measurement on day 44. There are 3 groups of rats, each on a different diet.

```
> head(BodyWeight)
Grouped Data: weight ~ Time | Rat
  weight Time Rat Diet
1    240    1   1    1
2    250    8   1    1
3    255   15   1    1
4    260   22   1    1
```

a) Plot the data using *plot(BodyWeight)* and comment.



It can be observed:

- differences among the three diet groups;
- there is evidence of a rat in diet group 2 with an unusually high initial body weight;
- the body weights appear to grow linearly with time, possibly with different intercepts and slopes for each diet.

b) In R use *lme* of package "*nlme4"* to fit the appropriate model for this study (consider intercept and slope random effects to account for rat-to-rat variation). Use the commands *summary, anova, ranef* and *fitted.* Explain how each fitted value is obtained.

c) The observations are made in the same individual over time. In this context it can be model the dependence among the within-group errors. The observations are not equally spaced in time, as an extra observation is taken at 44 days. In this case, we can use a spatial correlation structure for random errors. Several correlation structures are available in package *nlme*, for example, corEXp, corGaus, corSpher. Use the commands:

```
bodyw2.lme<-update(bodyw1.lme, corr=corExp(form=~Time))
bodyw3.lme<-update(bodyw1.lme, corr=corGaus(form=~Time))
bodyw4.lme<-update(bodyw1.lme, corr=corSpher(form=~Time)).
```

According to AIC and BIC criteria, what is the best correlation structure?

d) The model selected in item c) is significantly better than the model fitted in item b?