MODELOS MATEMÁTICOS E APLICAÇÕES 2023/2024

EXERCISES

LINEAR MIXED MODELS

Elsa Gonçalves, ISA/UL, 2024

Exercise 1

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

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

b) 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"; apply the command *summary* to the two objects created above and identify the REML estimates for the variance components.

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

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

(iv) Explore commands *ranef* and *fitted* of packages "nlme" and "lme4".

c) The ANOVA table for a random model with one factor of random effects (Factor A), balanced, with $\boldsymbol{G}$ and $\boldsymbol{R}$ diagonal matrices, is described as follows:

	D.F.	S.Q.	QM	E[QM]
Factor A	$a-1$	$SQA = \sum^{a} b\,(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$QMA = \frac{SQA}{a-1}$	${\sigma_u}^2 + {\sigma_e}^2$
Resíduals	$a(b-1)$	$SQRE = \sum^{a}\sum^{b}(Y_{ij} - \bar{Y}_{i.})^2$	$QMRE = \frac{SQRE}{a(b-1)}$	${\sigma_e}^2$
TOTAL	$ab-1$	$SQT = \sum^{a}\sum^{b}(Y_{ij} - \bar{Y}_{..})^2$		

i) What are the estimators for the variance components (procedure based on expected mean squares from the analysis of variance)?

ii) For the analysis of this data set, the results obtained were:

	Df	Sum Sq	Mean Sq
clone	195	228.8	1.1736
Residuals	784	176.0	0.2245

What are the variance components estimates (use the result obtained in ci)). Compare the results with those obtained in item b(i).

d) In fact, the Touriga Nacional field trial described above 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 hypothesis tests for the variance components of the model.

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

Exercise 2

In a study with traditional tomato varieties, fruit weight (g/fruit) was evaluated on 41 varieties in a field trial planted according to a randomized complete block design (4 blocks). The researcher considered the effects of blocks and varieties as random. The main objective of this study was to evaluate the variability of the fruit weight in tomato varieties. The fruit weight is available in *data.frame* tomate. The results obtained in R are described as follows:

```
>tomate<-read.table("D:\\ELSA-T\\Aulas\\Modelosmatemáticos-
UCdoutoramento\\dados\\tomate.txt", header=T)
> tomate.lmer1<-lmer(pesofruto~1+(1|variedade)+(1|bloco), data=tomate)
> tomate.lmer1
Linear mixed model fit by REML ['lmerMod']
Formula: pesofruto ~ 1 + (1 | variedade) + (1 | bloco)
   Data: tomate
REML criterion at convergence: 1852.448
Random effects:
 Groups      Name        Std.Dev.
 variedade (Intercept)  72.988
 bloco     (Intercept)   9.628
 Residual                53.398
Number of obs: 164, groups: variedade, 41; bloco, 4
Fixed Effects:
(Intercept)
      214.9
> logLik(tomate.lmer1)
'log Lik.' -926.2243
> tomate.lmer2<-lmer(pesofruto~1+(1|bloco), data=tomate)
> logLik(tomate.lmer2)
'log Lik.' -967.8129.
```

a) Describe the appropriate model for this study.

b) Carry out the hypothesis tests that answer the objective of the study.

Exercise 3

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

a) Describe the appropriate model for this study.

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

```
> terrenolme1<-lme(rend~variedade, random=~1|terreno, data=terrenos)
> summary(terrenolme1)
Linear mixed-effects model fit by REML
 Data: terrenos
      AIC      BIC    logLik
  55.42708 66.65429 -21.71354
Random effects:
 Formula: ~1 | terreno
       (Intercept)  Residual
StdDev:   0.1604919 0.3123811
Fixed effects: rend ~ variedade
                    Value       Std.Error     DF     t-value      p-value
(Intercept)       1.5560000   0.09740463  36    15.974600    0.0000
variedadeB     -0.0238462   0.12252595   36    -0.194621    0.8468
variedadeC    -0.3890769    0.12252595   36    -3.175465    0.0031
variedadeD    -0.3778462   0.12252595   36     -3.083805    0.0039
> vcov(terrenolme1)
                    (Intercept)   variedadeB   variedadeC   variedadeD
(Intercept)       0.009487662 -0.007506305 -0.007506305 -0.007506305
variedadeB     -0.007506305  0.015012610  0.007506305  0.007506305
variedadeC     -0.007506305  0.007506305  0.015012610  0.007506305
variedadeD     -0.007506305  0.007506305  0.007506305  0.015012610
```

b)

i) 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).

ii) Is the mean yield of variety B equal to the mean yield of variety A (for $\alpha = 0.05$)?

c) Use the commands and comment the results.

```
> terrenos.lme1<-lme(rend~variedade, random=~1|terreno, data=terrenos)
> plot(terrenos.lme1)
> qqnorm(terrenos.lme1, ~resid(.))
> qqnorm(terrenos.lme1, ~ranef(.))
```

Exercise 4

The data set *Machines* (Pinheiro e Bates, 2000) is available in both *libraries nlme* and *lme4* of R. 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 productivity score taking into account the number and quality of components produced.

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

b) What are the restricted maximum likelihood estimates for the 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?

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) The ANOVA table for a linear mixed model considering one factor of fixed effects (factor A) and one factor of random effects (factor B), balanced, with interaction, and **G** and **R** diagonal matrices, is described as follows:

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

The appropriate F-statistic is a quotient of QM that is chosen such that the expected value of the numerator and the expected value of the denominator differ only in the fixed effects of the factor being tested. For this example, specify the F-statistic for the test of fixed effects of factor A (Machine).

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.

h) 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])
```

Exercise 5

Six clones of the grapevine variety Antão Vaz were evaluated regarding yield (kg/plant) in 11 different environments, trying to make them representative of the range of environments where the clones will be grown. In each environment, a field trial was installed in which the six clones were planted according to a complete randomized experimental design with 8 repetitions. The yield data is available in *data.frame* antaovaz.

a) Describe the appropriate model for this study.

b) Fit the model previously described in R with the lme4 package. Proceed accordingly to answer the following questions: is there interaction variability? Is there variability between environments? Are there main effects associated with clones?

Exercise 6

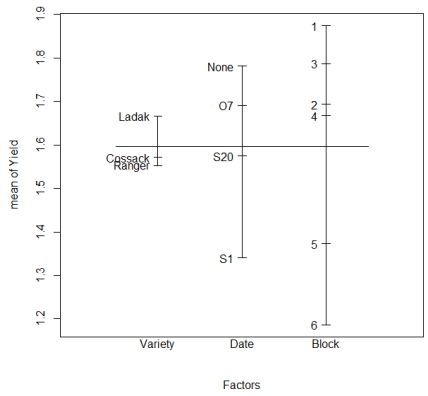
In *package* "*nlme*" of R, there is a data set named "*Alfalfa*".

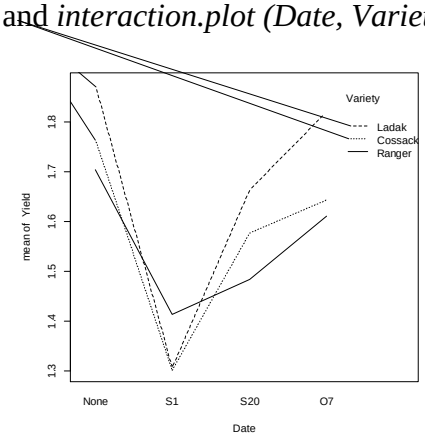
```
> head(Alfalfa)
Grouped Data: Yield ~ Date | Block/Variety
  Variety Date Block Yield
1   Ladak None     1  2.17
2   Ladak   S1     1  1.58
3   Ladak  S20     1  2.29
4   Ladak   O7     1  2.23
5   Ladak None     2  1.88
6   Ladak   S1     2  1.26
        …
```

This data is 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: the largest experimental units)*, 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.

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





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

d) The ANOVA table for a linear mixed model considering two factors with fixed effects (factors A and B) and random blocks, balanced, and **G** and **R** diagonal matrices, is described as follows:

	G.L.	S.Q.	QM	E[QM]
Factor A	$a-1$	SQA	QMA	$c{\sigma_{\alpha u}}^2 + {\sigma_e}^2 + bc\frac{\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}_.)^2}{a-1}$
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	${\sigma_e}^2 + ab\frac{\sum_{k=1}^{c}\left(\beta_k - \bar{\beta}_.\right)^2}{c-1}$
Interaction FactorA×FactorB	$(a-1)(c-1)$	$SQAB$	$QMAB$	${\sigma_e}^2 + b\frac{\sum_{i=1}^{a}\sum_{k=1}^{c}\left(\alpha\beta_{ik} - \overline{\alpha\beta}_{..}\right)^2}{(a-1)(c-1)}$
Residuals (Within plot error)	$a(b-1)(c-1)$	$SQRE$	$QMRE$	${\sigma_e}^2$

For this example, specify the F-statistic for the test of fixed effects of factors A and B.

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

Note:

(1) **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	$c{\sigma_{\alpha u}}^2 + b\frac{a}{a-1}{\sigma_{\alpha\beta}}^2 + {\sigma_e}^2 + bc\frac{\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}_.)^2}{a-1}$	$\frac{QMA+QMRE}{\text{QMWError}+QMAB}$**
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	$ab{\sigma_\beta}^2 + {\sigma_e}^2$	
Interaction FactorA×FactorB	$(a-1)(c-1)$	$QMAB$	$b\frac{a}{a-1}{\sigma_{\alpha\beta}}^2 + {\sigma_e}^2$	
Residuals	$a(b-1)(c-1)$	$QMRE$	${\sigma_e}^2$	

*The appropriate F-statistic is a quotient of QM that is chosen such that the expected value of the numerator and the expected value of the denominator differ only in the fixed effects of the factor

being tested. ** In this case, approximate degrees of freedom. For example, *Satterthwaite* method:

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

7. In *package* "*nlme*" of R, there is a data set named "*Oats*").

```
> head(Oats)
Grouped Data: yield ~ nitro | Block
  Block     Variety nitro yield
1     I     Victory   0.0   111
2     I     Victory   0.2   130
3     I     Victory   0.4   157
4     I     Victory   0.6   174
5     I Golden Rain   0.0   117
6     I Golden Rain   0.2   114
```

This data is described in Yates (1935) as an example of a *split-plot design*.

The objective is to study whether the yield (bushels/acre) of oats is affected by variety and nitrogen concentrations. There are two factors under study: factor oat variety, with 3 levels (Golden Rain, Marvellous and Victory) and factor nitrogen concentration (cwt/acre), with 4 levels (0.0, 0.2, 0.4 and 0.6). The experimental units were arranged in 6 blocks, each block with 3 large plots, each one divided into 4 plots. Oat varieties were randomly assigned to the large plots and nitrogen concentrations were randomly assigned to the split plots. Variety and nitrogen concentration were admitted as fixed effects factors and block as random effects factor.

a) Describe the appropriate model for this study.

b) Fit the model described in item a) in R using *lmer* of package "*lme4"*. Consider the nitrogen concentration as a factor (in R use factor(nitro) in the model description).

c) For a significance level of 0.05, what are the main conclusions of this study?

Exercise 8

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 tools are currently widely used in mixed models context.

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.

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, the dependence among the within-group errors can be modelled. 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) Is the model selected in item c) significantly better than the model fitted in item b?