Experimental Design and Statistical Methods

Workshop

MIXED MODELS

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Items

- Mixed models
  - Fixed and random effects
  - Matrix notation
  - Covariance matrices
  - Analysis of a time trend

- Basic commands
  - as.numeric
  - groupedData
  - gsub
  - lme
  - order
  - xtabs
- Libraries
  - lattice
  - multcomp
  - reshape

Repeated measures

Experimental units are often **measured repeatedly** if the precision of single measurements is not adequate or if changes are expected over time.

Variability among measurements on the same experimental unit can be homogeneous, but may alternatively be expected to change through time.

An experimental unit measured repeatedly is often called a **subject**.

Repeated measurements on the same subject can originate **correlations** between the repeated measurements. Measurements on the same animal are not independent. It may be necessary to define the **appropriate covariance structure** for such measurements.

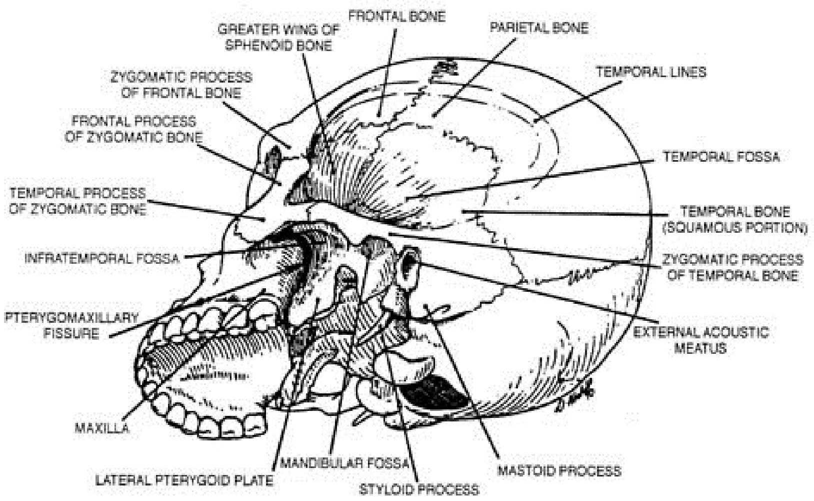
Since the experimental unit is an animal and not a single measurement on the animal, it is necessary to define the **appropriate experimental error** for testing hypothesis.

Models for analysing repeated measures can have the effects of period (time) defined as categorical or continuous independent variables.

Data

To develop this chapter we will use the classical example of Potthoff and Roy (1964).

Growth of 27 children, 10 girls and 17 boys, was monitored from 8 to 14 years, by measuring the distance in millimetres from the centre of the pituitary gland to the pterygomaxillary fissure, every two years.



http://www.emory.edu/ANATOMY/AnatomyManual/fossae.html

PERSON	SEX	Y8	Y10	Y12	Y14
1	f	21	20	21,5	23
2	f	21	21,5	24	25,5
3	f	20,5	24	24,5	26
4	f	23,5	24,5	25	26,5
5	f	21,5	23	22,5	23,5
6	f	20	21	21	22,5
7	f	21,5	22,5	23	25
8	f	23	23	23,5	24
9	f	20	21	22	21,5
10	f	16,5	19	19	19,5
11	f	24,5	25	28	28
12	m	26	25	29	31
13	m	21,5	22,5	23	26,5
14	m	23	22,5	24	27,5
15	m	25,5	27,5	26,5	27
16	m	20	23,5	22,5	26
17	m	24,5	25,5	27	28,5
18	m	22	22	24,5	26,5
19	m	24	21,5	24,5	25,5
20	m	23	20,5	31	26
21	m	27,5	28	31	31,5
22	m	23	23	23,5	25
23	m	21,5	23,5	24	28
24	m	17	24,5	26	29,5
25	m	22,5	25,5	25,5	26
26	m	23	24,5	26	30
27	m	22	21,5	23,5	25

Fixed and random factors

- **Fixed**
  - All levels of the classification or only the levels or interest are in the experiment
  - Levels were not chosen at random
  - A new experiment would bring the same set of variables as the old experiment
  - Variation is brought about by a random element of measurement σ_e^2
  - ➢ Sex, age classes, management system …

- **Random**
  - Levels are a random sample of a definable or conceptual population
  - Repeating the experiment would bring a new set of random variables
  - Variation is same of nature of random variables (measurement error)
  - ➢ Person, animal, litter effect …

The decision to call a factor fixed or random needs to be made by the researcher. It is a matter of experience and no close guidelines exist.

Mixed model (with both fixed and random effects)

$$y_{ijt} = \mu + age_t + sex_i + person_{j(i)} + \varepsilon_{ijt}$$

$$\underbrace{\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ . \\ . \\ y_{27,4} \end{bmatrix}}_{} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ . \\ . \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & . & 1 \end{bmatrix} \begin{bmatrix} Y_8 \\ Y_{10} \\ Y_{12} \\ Y_{14} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ . & . \\ . & . \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f \\ m \end{bmatrix}}_{} + \underbrace{\begin{bmatrix} 1 & 0 & . & . & 0 \\ 1 & 0 & . & . & 0 \\ 1 & 0 & . & . & 0 \\ 1 & 0 & . & . & 0 \\ 0 & 1 & . & . & 0 \\ 0 & 1 & . & . & 0 \\ 0 & 1 & . & . & 0 \\ 0 & 1 & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ . \\ . \\ p_{27} \end{bmatrix}}_{} + \underbrace{\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ . \\ . \\ \varepsilon_{27,4} \end{bmatrix}}_{}$$

$$\boxed{\mathbf{y}} = \underset{\text{fixed}}{\boxed{\mathbf{Xb}}} + \underset{\text{random}}{\boxed{\mathbf{Zu}}} + \boxed{\boldsymbol{\varepsilon}} \qquad V(\mathbf{y}) = \mathbf{ZGZ'} + \mathbf{R}$$

(Co)variance matrix

Suppose we define the matrices **G** and **R** as follows:

$$\mathbf{G} = \mathbf{I}\sigma_u^2 ; \quad \mathbf{R} = \mathbf{I}\sigma_e^2$$

Then

$$V(\mathbf{y}) = \mathbf{V} = \mathbf{ZZ}'\sigma_u^2 + \mathbf{I}\sigma_e^2 \Leftarrow$$

Person 1

$$\mathbf{ZZ}' = \begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & . & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & . & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & . & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & . & . & 0 & 0 & 0 & 0 \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & . & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & . & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & . & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & . & . & 1 & 1 & 1 & 1
\end{bmatrix}$$

Person 27

Which gives for each block matrix in **V**:

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

This structure is called **Compound Symmetry**.
Independence among subjects (persons) is assumed.

REML estimation in nlme

The method of estimation is Restricted Maximum Likelihood (REML). The REML log likelihood function of $y \sim N(\mathbf{Xb}, \mathbf{ZGZ}' + \mathbf{I}\sigma_e^2)$ is:

$$l_{RML} = -\tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{ML})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_{ML})$$
$$-\tfrac{n-p}{2}\log(2\pi) - \tfrac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|$$

And the solutions for fixed and random effects are:

$$\hat{\mathbf{b}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$
$$\hat{\mathbf{u}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

$\hat{\mathbf{b}}$is an EBLUE estimator and $\hat{\mathbf{u}}$is an EBLUP predictor.

(Badiella and Sánchez, 2011)

Transformation of the dataset

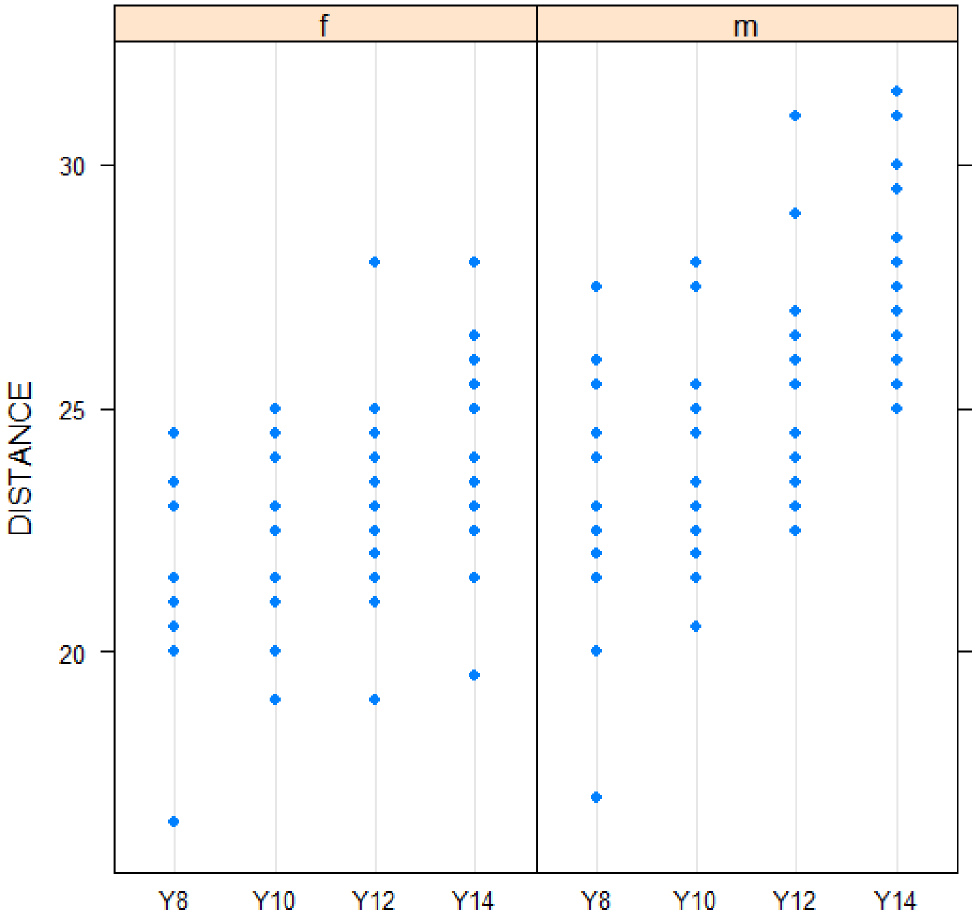
```
> require(reshape)
> GROWTH<-melt(GROWTH, id = c(“PERSON”, “SEX”))
> GROWTH<-rename(GROWTH, c(variable=“AGE”, value=“DISTANCE”))
> GROWTH[with(GROWTH, order(PERSON, AGE)),]
```

The data were presented as usually they are registered in an Excel spreadsheet (multivariate form). However, for the analysis of mixed models, the data must be organized in a univariate array, as shown in this table, where only data of the 3 first persons are shown.

```
   PERSON SEX AGE DISTANCE
1       1   f  Y8     21.0
28      1   f Y10     20.0
55      1   f Y12     21.5
82      1   f Y14     23.0
2       2   f  Y8     21.0
29      2   f Y10     21.5
56      2   f Y12     24.0
83      2   f Y14     25.5
3       3   f  Y8     20.5
30      3   f Y10     24.0
57      3   f Y12     24.5
84      3   f Y14     26.0
        ..................
```

A first look to the data

```
> library(lattice)
> dotplot(DISTANCE~AGE|SEX,GROWTH)
```



A first look to the means

```
> xtabs(DISTANCE~AGE+SEX,GROWTH)/xtabs(~AGE+SEX,GROWTH)
```

```
     SEX
AGE          f        m
  Y8  21.18182 22.87500
  Y10 22.22727 23.81250
  Y12 23.09091 25.71875
  Y14 24.09091 27.46875
```

There is an increase of the response variable (DISTANCE) with age in both sexes.

Age and sex as factors, with interaction (1)

```
> require(nlme)
> summary(MODF1<-lme(DISTANCE~AGE*SEX, GROWTH, random=~1|PERSON))
```

require is an alternative to **library**. The model includes fixed variables (**AGE** and **SEX**) and their interaction, as well as a random factor, **PERSON** affecting the intercept.

```
Linear mixed-effects model fit by REML
 Data: GROWTH
       AIC      BIC    logLik
  443.4085 469.4602 -211.7043

Random effects:
 Formula: ~1 | PERSON
        (Intercept) Residual
StdDev:    1.812564  1.40536
```

$\downarrow$ σ_u $\downarrow$ σ_e

Age and sex as factors, with interaction (2)

```
Fixed effects: DISTANCE ~ AGE * SEX
                Value Std.Error DF   t-value p-value
(Intercept) 21.181818 0.6915349 75 30.630149  0.0000
AGEY10       1.045455 0.5992477 75  1.744612  0.0851
AGEY12       1.909091 0.5992477 75  3.185813  0.0021
AGEY14       2.909091 0.5992477 75  4.854572  0.0000
SEXm         1.693182 0.8983302 25  1.884810  0.0711
AGEY10:SEXm -0.107955 0.7784456 75 -0.138680  0.8901
AGEY12:SEXm  0.934659 0.7784456 75  1.200674  0.2337
AGEY14:SEXm  1.684659 0.7784456 75  2.164132  0.0336
```

The estimates of the effects of the levels of AGE and SEX are statistically significant (but for AGEY10), whereas only one of the interaction effects is significant.

Age and sex as factors, with interaction (3)

```
 Correlation:
             (Intr) AGEY10 AGEY12 AGEY14 SEXm   AGEY10: AGEY12:
AGEY10      -0.433
AGEY12      -0.433  0.500
AGEY14      -0.433  0.500  0.500
SEXm        -0.770  0.334  0.334  0.334
AGEY10:SEXm  0.334 -0.770 -0.385 -0.385 -0.433
AGEY12:SEXm  0.334 -0.385 -0.770 -0.385 -0.433  0.500
AGEY14:SEXm  0.334 -0.385 -0.385 -0.770 -0.433  0.500   0.500
```

This correlation matrix is computed from the variance-covariance matrix of the estimates of fixed effects: $Var(\hat{\mathbf{b}}) = \hat{s}^2(\mathbf{X}'\mathbf{X})^{-1}$.

Absolute values above 0.90 indicate over-parameterization of the model. It is not the case of our model.

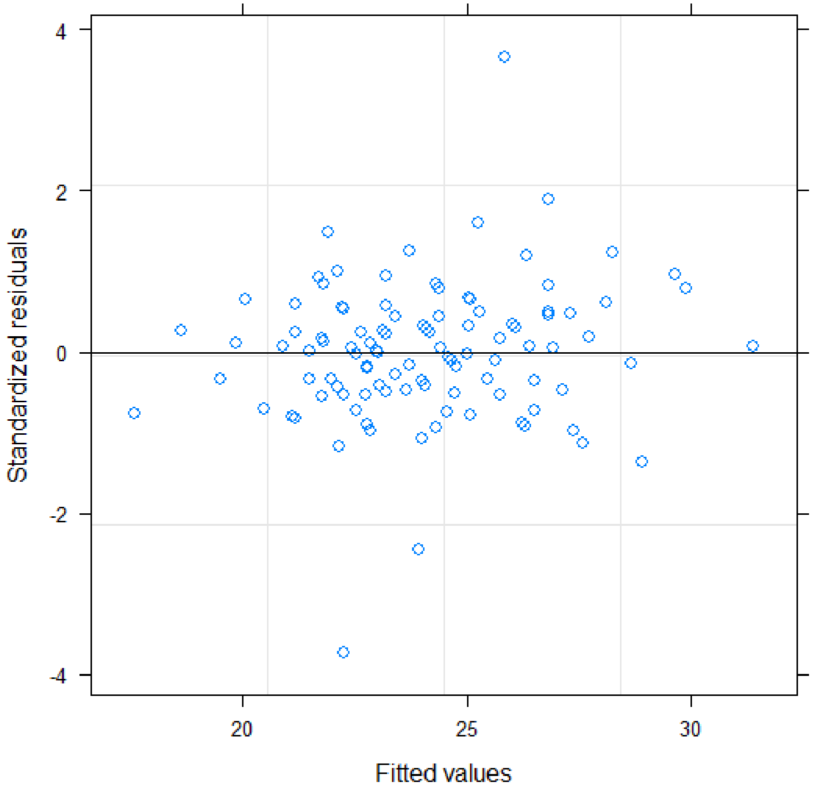
Age and sex as factors, with interaction (4)

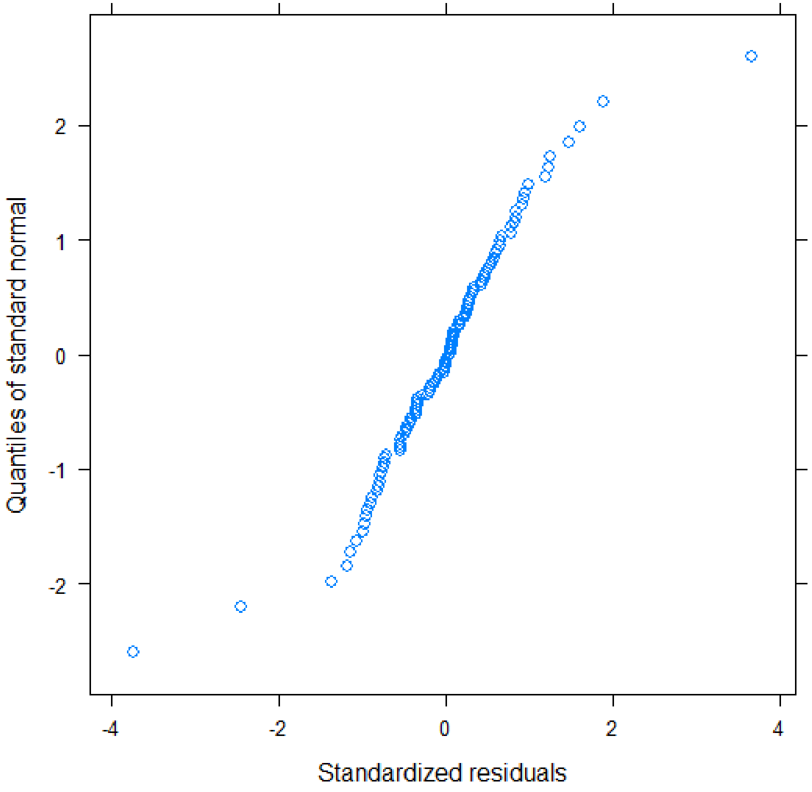
```
> anova(MODF1)
            numDF denDF  F-value p-value
(Intercept)     1    75 4123.156  <.0001
AGE             3    75   40.032  <.0001
SEX             1    25    9.292  0.0054
AGE:SEX         3    75    2.362  0.0781
```

This ANOVA table indicates that the levels of the fixed effects AGE and SEX are statistically significant, whereas the interaction AGE:SEX only approaches statistical significance.

Age and sex as factors, with interaction (5)

```
> #Validation plots
> plot(MODF1); qqnorm(MODF1)
```





The distribution of the standardized residuals is fairly random and no obvious deviations to normality is observed, but for 2 observations.

Age and sex as factors, with interaction (6)

```
> require(multcomp)
> summary(glht(MODF1, linfct=mcp(SEX="Tukey")))

Multiple Comparisons of Means: Tukey Contrasts

Linear Hypotheses:
           Estimate Std. Error z value Pr(>|z|)
m - f == 0   1.6932     0.8983   1.885   0.0595 .

> summary(glht(MODF1, linfct=mcp(AGE="Tukey")))

Linear Hypotheses:
               Estimate Std. Error z value Pr(>|z|)
Y10 - Y8 == 0    1.0455     0.5992   1.745  0.30056
Y12 - Y8 == 0    1.9091     0.5992   3.186  0.00795 **
Y14 - Y8 == 0    2.9091     0.5992   4.855  < 0.001 ***
Y12 - Y10 == 0   0.8636     0.5992   1.441  0.47355
Y14 - Y10 == 0   1.8636     0.5992   3.110  0.00988 **
Y14 - Y12 == 0   1.0000     0.5992   1.669  0.34033
```

The differences between sexes across ages approach significance and there are significant differences when the interval is at least 4 years.

Age and sex as factors, without interaction (1)

```
> summary(MODF2<-lme(DISTANCE~AGE+SEX,GROWTH,random=~1|PERSON))

Linear mixed-effects model fit by REML
 Data: GROWTH
       AIC      BIC    logLik
  447.7076 466.1507 -216.8538

Random effects:
 Formula: ~1 | PERSON
        (Intercept) Residual
StdDev:    1.805417 1.441689

Fixed effects: DISTANCE ~ AGE + SEX
                Value Std.Error DF  t-value p-value
(Intercept) 20.809764 0.6334777 78 32.85003  0.0000
AGEY10       0.981481 0.3923780 78  2.50137  0.0145
AGEY12       2.462963 0.3923780 78  6.27702  0.0000
AGEY14       3.907407 0.3923780 78  9.95827  0.0000
SEXm         2.321023 0.7614168 25  3.04829  0.0054
```

All effects (for AGE and SEX) were statistically significant.

Age and sex as factors, without interaction (2)

```
 Correlation:
       (Intr) AGEY10 AGEY12 AGEY14
AGEY10 -0.310
AGEY12 -0.310  0.500
AGEY14 -0.310  0.500  0.500
SEXm   -0.712  0.000  0.000  0.000

> anova(MODF2)
            numDF denDF  F-value p-value
(Intercept)     1    78 4123.156  <.0001
AGE             3    78   38.040  <.0001
SEX             1    25    9.292  0.0054
```

The values of the correlations are lower than in the previous model.
The ANOVA table confirms the significance of both effects, age and sex.

Age and sex as factors, without interaction (3)

```
> require(multcomp)
> summary(glht(MODF2, linfct=mcp(SEX="Tukey")))

Multiple Comparisons of Means: Tukey Contrasts

Linear Hypotheses:
           Estimate Std. Error z value Pr(>|z|)
m - f == 0   2.3210     0.7614   3.048   0.0023 **

> summary(glht(MODF2, linfct=mcp(AGE="Tukey")))

Linear Hypotheses:
               Estimate Std. Error z value Pr(>|z|)
Y10 - Y8 == 0    0.9815     0.3924   2.501   0.0598 .
Y12 - Y8 == 0    2.4630     0.3924   6.277   <0.001 ***
Y14 - Y8 == 0    3.9074     0.3924   9.958   <0.001 ***
Y12 - Y10 == 0   1.4815     0.3924   3.776   0.0011 **
Y14 - Y10 == 0   2.9259     0.3924   7.457   <0.001 ***
Y14 - Y12 == 0   1.4444     0.3924   3.681   0.0015 **
```

This second model is most sensitive than the previous one: all differences are statistically significant.

Comparison of both models

The estimation method in **nlme** is REML. To compare the fixed part of the models we need a Likelihood Ratio Test based upon Maximum Likelihood (ML) method. This can be done by updating to ML the REML computed models prior to their comparison using the **anova** command.

```
> MODF1.ml<-update(MODF1, method="ML")
> MODF2.ml<-update(MODF2, method="ML")
> anova(MODF1.ml, MODF2.ml)
```

```
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
MODF1.ml     1 10 446.6329 473.4542 -213.3165
MODF2.ml     2  7 447.9443 466.7192 -216.9721 1 vs 2 7.311335  0.0626
```

The preferred model is the one giving the lower AIC and BIC values. In this case the values of AIC and BIC for both models contradict each other. The *p*-value of the L.Ratio only approaches statistical significance. We can not conclude the superiority of any of the two models. We could use the simplest one.

Bayesian information criterion

The Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) was developed by G. E. Schwarz. It is a criterion for **model selection** among a finite set of models. It is based, in part, on the likelihood function and it is closely related to the Akaike Information Criterion that we have seen in a previous lesson.

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model. The **penalty term is larger in BIC than in AIC**.

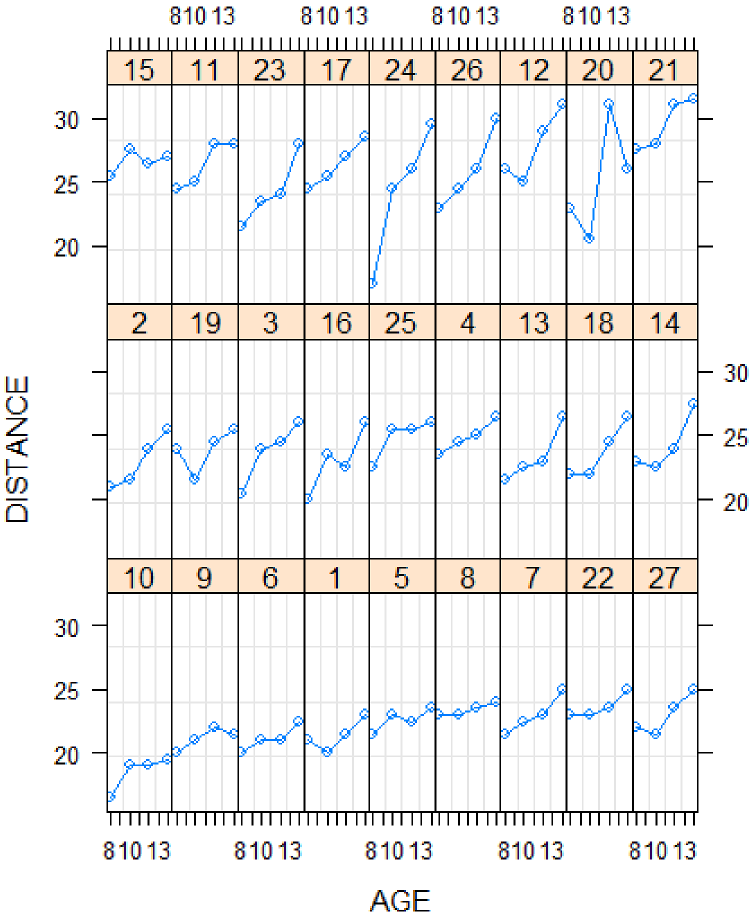
$$BIC = -2 \ln L + k \ln(n)$$

L being the estimated likelihood, *k* the number of parameters and *n* the number of observations. **Models with lower BIC are preferred**.

(Adapted from Wikipedia)

Evolution of growth through time

```
> GROWTH$AGE<-as.numeric(gsub("Y", "", GROWTH$AGE))
> GROWTH<-groupedData(DISTANCE~AGE|PERSON,GROWTH)
> plot(GROWTH)
```



To transform **AGE** in a number, we need to eliminate **Y**, using **gsub** and specify the new variable as a number by writing **as.numeric**.

groupedData allows us to specify in the data frame which is the variable (**DISTANCE**), the primary covariate (**AGE**) and the grouping factor (**PERSON**).

We can observe a linear trend for the variable DISTANCE according to AGE

Age as a regressor

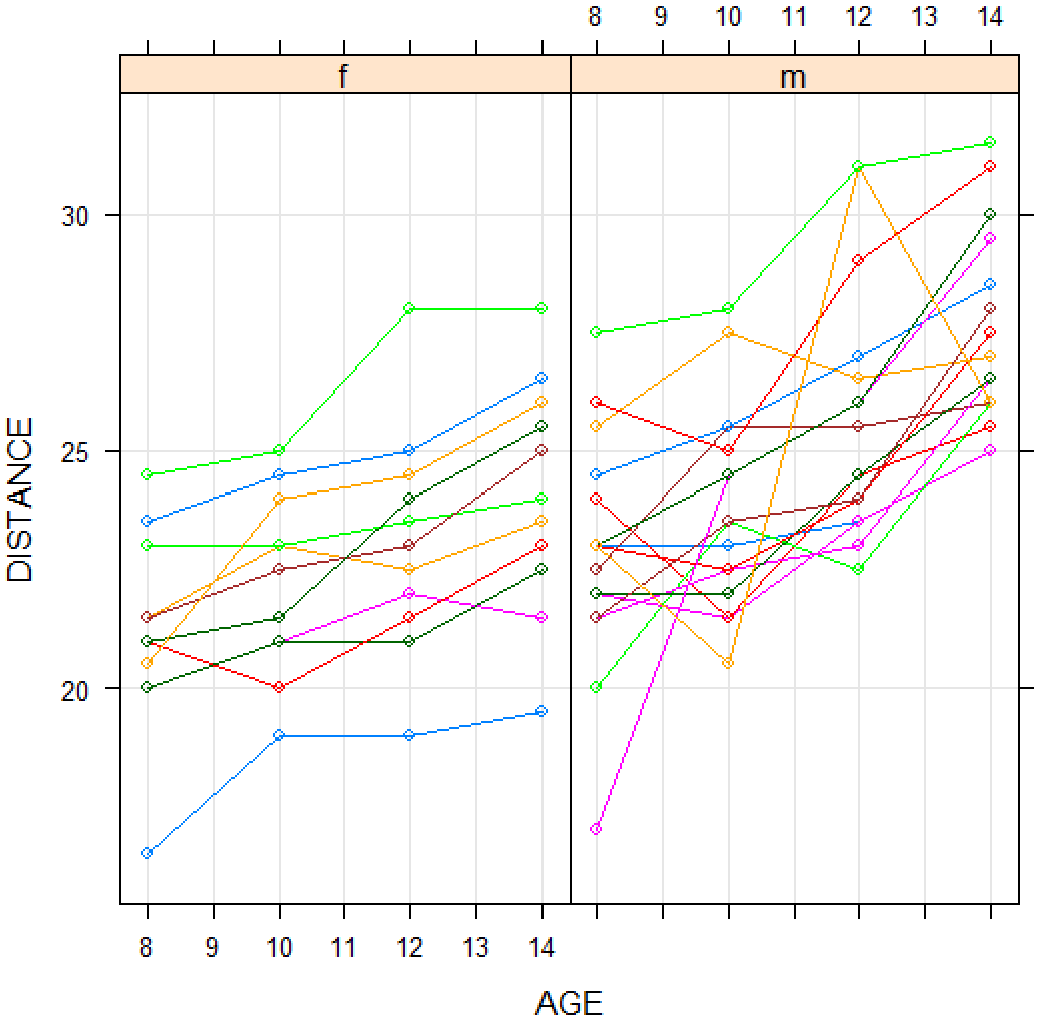
```
> summary(MODR1<-lme(DISTANCE~AGE,GROWTH,random=~1|PERSON))

Linear mixed-effects model fit by REML
 Data: GROWTH
       AIC      BIC    logLik
  455.0025 465.6563 -223.5013

Random effects:
 Formula: ~1 | PERSON
        (Intercept) Residual
StdDev:    2.114724 1.431592

Fixed effects: DISTANCE ~ AGE
                Value Std.Error DF  t-value p-value
(Intercept) 16.761111 0.8023952 80 20.88885       0
AGE          0.660185 0.0616059 80 10.71626       0
```

We assume a common intercept and slope for age in the two sexes. Both estimates are statistically significant. The regression equation is:

$$\hat{y} = 17.7611 + 0.6602 \times AGE$$

Evolution of growth through time for each sex

```
> library(lattice)
> xyplot(DISTANCE~AGE|SEX,GROWTH,type=c("o","g"),groups=PERSON)
```



`lattice` displays multivariate relationships: ***Trellis graphs*** that display a variable or the relationship between variables, conditioned on one or more other variables.

“o” : overplots points and lines for each person.

“g” adds a reference grid.

Age as a regressor according to sex

```
> summary(MODR2<-lme(DISTANCE~AGE*SEX,GROWTH,random=~1|PERSON))

Linear mixed-effects model fit by REML
 Data: GROWTH
       AIC      BIC    logLik
  445.7572 461.6236 -216.8786

Random effects:
 Formula: ~1 | PERSON
        (Intercept) Residual
StdDev:    1.816214 1.386382

Fixed effects: DISTANCE ~ AGE * SEX
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 1.1835071 79 14.679023  0.0000
AGE          0.479545 0.0934698 79  5.130483  0.0000
SEXm        -1.032102 1.5374208 25 -0.671321  0.5082
AGE:SEXm     0.304830 0.1214209 79  2.510520  0.0141
```

Estimates of Intercept and AGE correspond to females. For males, SEXm and AGE:SEXm must be added. The regression equations are:

$$\hat{y}_f = 17.3727 + 0.4796 \times AGE; \quad \hat{y}_m = 16.3406 + 0.7844 \times AGE$$

Some final considerations

This has been a brief introduction to the use of mixed models to analyze data with repeated measures in an individual.

Many more possibilities can be contemplated: random regression models, models with two or more random effects, models with other covariance structures, non normal data, etc.

Some other commands (gls) or library (lme4) can be used to develop them.

The book of Badiella and Sánchez (2011) presents an excellent development of these topics.

References

Badiella L., Sánchez J.A. 2011. *Modelos mixtos con R*. Servei d'Estadística Aplicada, UAB. Bellaterra (Cerdanyola del Vallés), Spain.

Potthoff R.F., Roy S.N. 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313-326.