

# 01NAEX - Lecture 12

## Introduction to Longitudinal Data Analysis

Jiri Franc

Czech Technical University  
Faculty of Nuclear Sciences and Physical Engineering  
Department of Mathematics

# Introduction to Longitudinal Data Analysis

## **Longitudinal Data Analysis:**

- ▶ Experimental designs in which each subject is measured at several points in time.
- ▶ Often called "Repeated measures analysis".

Analysis of these data types can be done by

- ▶ Separate analysis for each time–point.
- ▶ Analysis of summary statistic.
- ▶ Random effects approach with different covariance models.

## Basic example

### **Dataset: Dental study from Pothoff and Roy (1964):**

- ▶ Sample of 27 children (16 boys, 11 girls)
- ▶ Response variable: Distance in mm between the center of the pituitary and the pterygomaxillary fissure.
- ▶ Measurement repeated at 8, 10, 12, and 14 years of age.

### **Questions of interest:**

- ▶ Does distance change over time?
- ▶ What is the pattern of change?
- ▶ Is the pattern different for boys and girls and how?

## Basic example - data preview

```
> library(nlme) > library(lattice) > data(Orthodont)
> head(Orthodont)
```

Grouped Data: distance ~ age | Subject

	distance	age	Subject	Sex
1	26.0	8	M01	Male
2	25.0	10	M01	Male
3	29.0	12	M01	Male
4	31.0	14	M01	Male
5	21.5	8	M02	Male
6	22.5	10	M02	Male

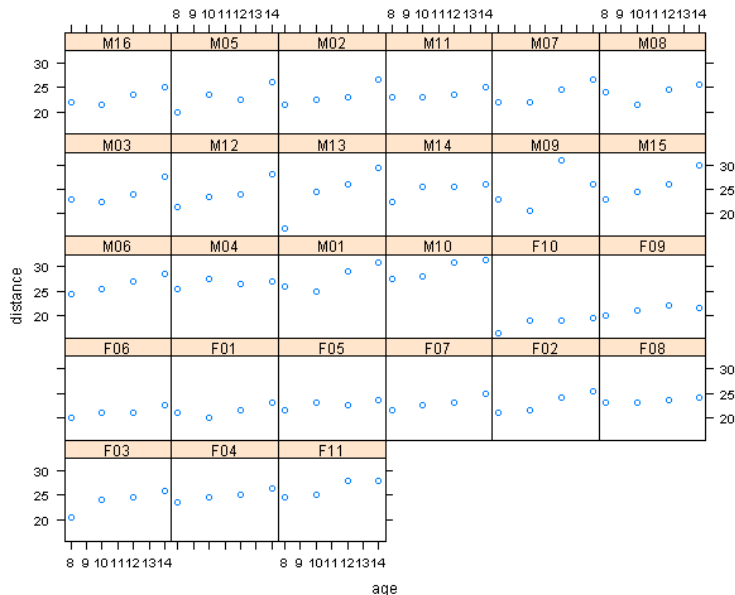
```
> summary(Orthodont)
```

distance	age	Subject	Sex
Min. :16.50	Min. : 8.0	M16 : 4	Male :64
1st Qu.:22.00	1st Qu.: 9.5	M05 : 4	Female:44
Median :23.75	Median :11.0	M02 : 4	
Mean :24.02	Mean :11.0	M11 : 4	
3rd Qu.:26.00	3rd Qu.:12.5	M07 : 4	
Max. :31.50	Max. :14.0	M08 : 4	
(Other):84			

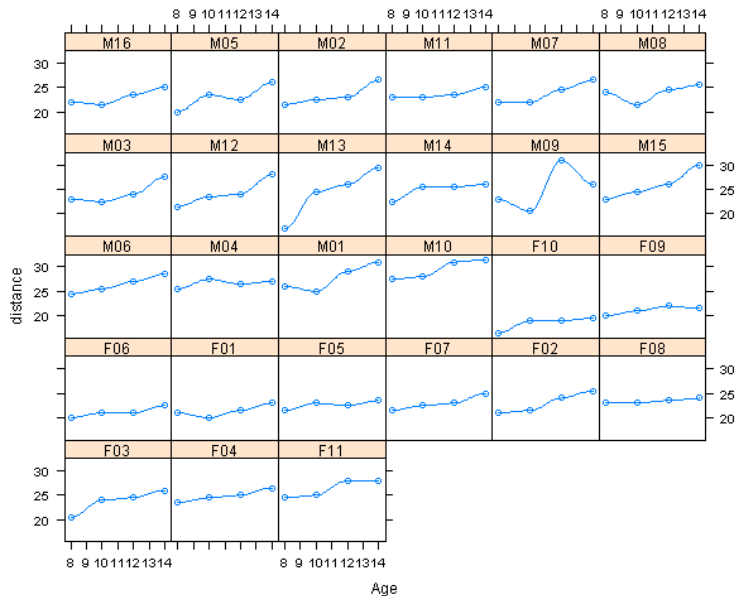
```
> attach(Orthodont)
```

## Basic example - visualization

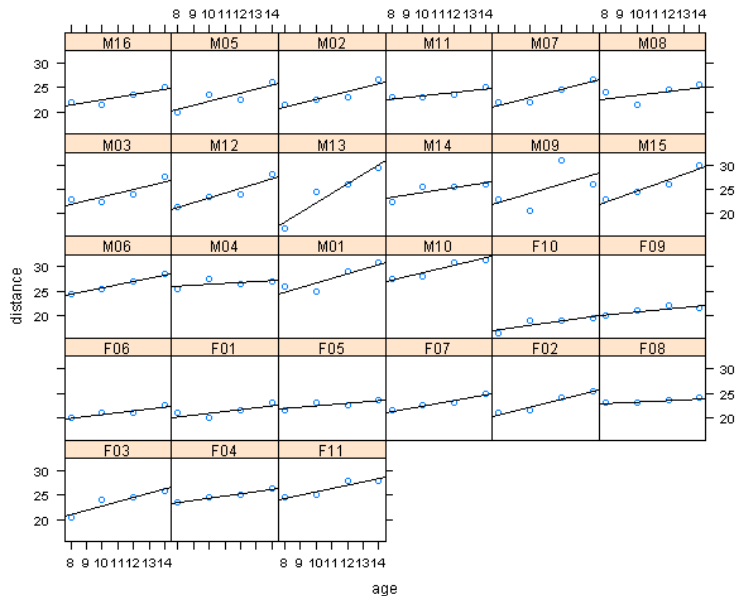
```
> xyplot(distance ~ age | Subject, data=Orthodont, as.table=T)
```



## Basic example - vizualization



## Basic example - vizualization







## Pooling dataset - wrong way analysis

### What can we conclude from figures?

- ▶ The male trajectories appear steeper than the female ones.
- ▶ Slopes are relatively consistent within each sex.
- ▶ Male seems to be higher than female overall.
- ▶ Subject who starts high (or low) tend to stay high (or low).
- ▶ The individual pattern for most subjects follows a rough straight line increase.

### Pooling dataset - Balanced data analysis:

- ▶ All subjects are measured at the same set of ages.
- ▶ Cross-sectional analysis comparing means (male vs. female) at each age 8, 10, 12, 14 by two-sample t-tests.
- ▶ Residuals within clusters are not independent; they tend to be highly correlated with each other

## Basic example - results 1

```
> summary(lm ( distance ~ age * Sex , Orthodont))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	16.3406	1.4162	11.538	< 2e-16 ***
age	0.7844	0.1262	6.217	1.07e-08 ***
SexFemale	1.0321	2.2188	0.465	0.643
age:SexFemale	-0.3048	0.1977	-1.542	0.126

Residual standard error: 2.257 on 104 degrees of freedom

Multiple R-squared: 0.4227, Adjusted R-squared: 0.4061

F-statistic: 25.39 on 3 and 104 DF, p-value: 2.108e-12

```
> summary(aov( distance ~ age * Sex , Orthodont))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	235.4	235.36	46.204	6.88e-10 ***
Sex	1	140.5	140.46	27.576	8.05e-07 ***
age:Sex	1	12.1	12.11	2.378	0.126
Residuals	104	529.8	5.09		

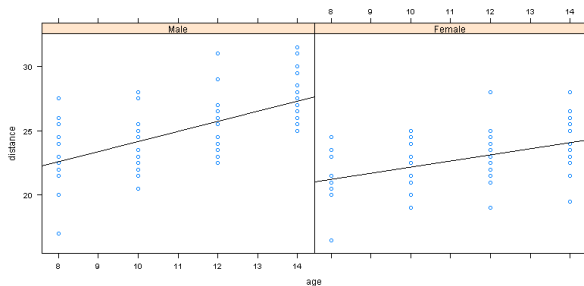
## Basic example - results 2

```
> summary(lm ( distance ~ age + Sex , Orthodont))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.70671      1.11221  15.920  < 2e-16 ***
age          0.66019      0.09776   6.753 8.25e-10 ***
SexFemale    -2.32102      0.44489  -5.217 9.20e-07 ***
---
Residual standard error: 2.272 on 105 degrees of freedom
Multiple R-squared:  0.4095, Adjusted R-squared:  0.3983
F-statistic: 36.41 on 2 and 105 DF,  p-value: 9.726e-13

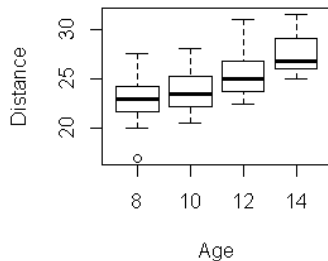
> summary(aov( distance ~ age + Sex , Orthodont))
Df Sum Sq Mean Sq F value    Pr(>F)
age          1   235.4   235.36    45.61 8.25e-10 ***
Sex          1   140.5   140.46    27.22 9.20e-07 ***
Residuals   105   541.9     5.16
```

Same slope for all subjects and estimated variance within each subject:  
 $\sigma^2 = 2.272^2$ .

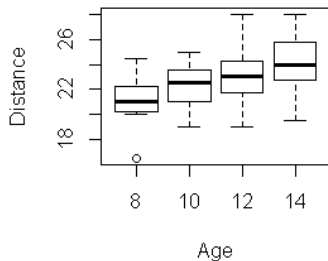
## Pooling dataset - wrong way analysis



**Boxplot for Male**



**Boxplot for Female**



## Basic example - results 3

```
> summary(lm( distance ~ age *Subject , Orthodont))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    16.76111      0.63281   26.487 < 2e-16 ***
age             0.66019      0.05638   11.711 < 2e-16 ***
Subject.L       5.07509      3.28817    1.543  0.12857
Subject.Q       0.59068      3.28817    0.180  0.85811
...
age:Subject.L  -0.49787      0.29293   -1.700  0.09496 .
age:Subject.Q  -0.19737      0.29293   -0.674  0.50334
age:Subject.C   0.69724      0.29293    2.380  0.02086 *
```

Residual standard error: 1.31 on 54 degrees of freedom  
Multiple R-squared: 0.899, Adjusted R-squared: 0.7999  
F-statistic: 9.07 on 53 and 54 DF, p-value: 6.568e-14

Different slope for each subjects and estimated variance for each subject:  
 $\sigma^2 = 1.31^2$ .

## Remarks to bad approach

### Why is the approach "single slope for single subject" wrong?

- ▶ No estimate of sex effect .
- ▶ Can't be generalized to population. We can't estimate new observation.
- ▶ No autocorrelation in time

### F-tests for no Sex effect

```
> names(fValues) <- levels(as.factor(age))
> for (i in 1:4){
+   fValues[as.integer(i)] <-
      anova(lm(distance[age == unique(age)[i]]
                ~ Sex[age == unique(age)[i]]))[1, 4]
+ }
> fValues
      8          10          12          14
3.450811  3.914354  6.972702 14.917559
> qf(0.95,1,25)
[1] 4.241699
```

A few significant values are found, so we can conclude that Sex evidence of group difference have been seen.

## New approach

### What can we do?

- ▶ Make a correct analysis time-by-time, but it is weak and often confusing, because it does not combine all information into one test.
- ▶ Randomize some variables (use lme, lmer)
- ▶ Add correlation structure (use gls for fixed and lme for mixed models)

**Remark:** For mixed model we have

$$\mathbf{y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZGZ}' + \mathbf{R}),$$

- ▶  $\mathbf{X}$  is the design matrix for the fixed effects part of the model;
- ▶  $\boldsymbol{\beta}$  is the fixed effects parameters;
- ▶  $\mathbf{Z}$  is the design matrix for the random effects;
- ▶  $\mathbf{G}$  the covariance between the random effects in the model;
- ▶  $\mathbf{R}$  covariance between the residual measurement errors.

## Theory of Mixed Models

Regression model of **Mixed linear model** approach:

$$y_{ij} = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Age}_{ij} + \beta_3 \text{Sex}_i \text{Age}_{ij} + u_{i0} + u_{i1} \text{Age}_{ij} + e_{ij}$$

$$\underbrace{\begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & \text{Sex}_i & \text{Age}_{i1} & \text{Sex}_i \text{Age}_{i1} \\ 1 & \text{Sex}_i & \text{Age}_{i2} & \text{Sex}_i \text{Age}_{i2} \\ 1 & \text{Sex}_i & \text{Age}_{i3} & \text{Sex}_i \text{Age}_{i3} \\ 1 & \text{Sex}_i & \text{Age}_{i4} & \text{Sex}_i \text{Age}_{i4} \end{pmatrix}}_{\mathbf{X}} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} 1 & \text{Age}_{i1} \\ 1 & \text{Age}_{i2} \\ 1 & \text{Age}_{i3} \\ 1 & \text{Age}_{i4} \end{pmatrix}}_{\mathbf{Z}} \cdot \underbrace{\begin{pmatrix} u_{i0} \\ u_{i1} \end{pmatrix}}_{\mathbf{u}} + \underbrace{\begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \\ e_{i4} \end{pmatrix}}_{\mathbf{e}},$$

equivalently,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where,

$$u_i \sim N(0, \mathbf{G}), \quad e_i \sim N(0, \mathbf{R}_i) \quad i \in \{1, 2, \dots, N\}.$$

and

$$\delta_i = \mathbf{Z}_i \mathbf{u}_i + \mathbf{e}_i \quad \delta_i \sim N(0, \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i) \quad y_i \sim N(\mu, \mathbf{V}).$$

Estimation of  $\boldsymbol{\beta}$  by Generalized Least Squares:

$$\hat{\boldsymbol{\beta}}^{GLS} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}.$$



### Compound symmetry correlation structure

$$y_i \sim N(\mu, \mathbf{V})$$

$$\text{cov}(y_{i_1}, y_{i_2}) = V_{i_1, i_2} = \begin{cases} 0 & \text{if } \text{Subject}_{i_1} \neq \text{Subject}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_{\text{Subj}}^2 & \text{if } \text{Subject}_{i_1} = \text{Subject}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_{\text{Subj}}^2 + \sigma^2 & \text{if } i_1 = i_2 \end{cases}$$

Two measurements from the same individual are correlated, but equally correlated no matter how far apart the measurements were taken.

This is counterintuitive if some measurements are close (in time or space) and some are far apart. To fix this the correlation structure need to be set.

## Correlation structures

### AR(1) correlation structure

$y_i \sim N(\mu, \mathbf{V})$  and  $e_i \sim N(0, \mathbf{R}_i)$ ,  $\rho \in (0, 1)$ , where

$$\mathbf{R}_i = \sigma^2 \begin{pmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{pmatrix}$$

Two observations "very close" together have covariance  $\sigma_{Subj}^2 + \tau^2$  and two observations "very far" apart have covariance  $\sigma_{Subj}^2$

$$\text{cov}(y_{i_1}, y_{i_2}) = V_{i_1, i_2} = \begin{cases} 0 & \text{if } \text{Subject}_{i_1} \neq \text{Subject}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_{Subj}^2 + \tau^2 \rho^{|i_1 - i_2|} & \text{if } \text{Subject}_{i_1} = \text{Subject}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_{Subj}^2 + \tau^2 + \sigma^2 & \text{if } i_1 = i_2 \end{cases}$$

## Correlation structures

```
m_gls<-glS(distance ~ age + Sex,  
  correlation=corCompSymm(form=~1|Subject),data=Orthodont)  
  
> summary(m_gls)  
Generalized least squares fit by REML  
Model: distance ~ age + Sex  
Data: Orthodont  
AIC      BIC      logLik  
447.5125 460.7823 -218.7563  
Correlation Structure: Compound symmetry  
Formula: ~1 | Subject  
Parameter estimate(s):  
Rho  
0.6144914  
Coefficients:  
                Value      Std.Error    t-value p-value  
(Intercept) 17.706713 0.8339225 21.233044 0.0000  
age           0.660185 0.0616059 10.716263 0.0000  
SexFemale    -2.321023 0.7614169 -3.048294 0.0029  
Correlation:  
(Intr) age  
age          -0.813  
SexFemale    -0.372 0.000  
Residual standard error: 2.305697  
Degrees of freedom: 108 total; 105 residual
```

## Correlation structures

```
> m_lme<-lme(distance ~ age * Sex,  
+           random=~1+age|Subject,  
+           correlation=corAR1(form=~1|Subject),  
+           data=Orthodont)
```

```
> summary(m_lme)  
Linear mixed-effects model fit by REML  
Data: Orthodont  
AIC      BIC      logLik  
446.8076 470.6072 -214.4038
```

Random effects:

Formula: ~1 + age | Subject

Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	3.3730482	(Intr)
age	0.2907673	-0.831
Residual	1.0919754	

Correlation Structure: AR(1)

Formula: ~1 | Subject

Parameter estimate(s):

Phi  
-0.47328

## Correlation structures

```
> m_lme<-lme(distance ~ age * Sex,  
+           random=~1+age|Subject,  
+           correlation=corAR1(form=~1|Subject),  
+           data=Orthodont)
```

```
> summary(m_lme) (..continue)
```

Fixed effects: distance ~ age \* Sex

	Value	Std.Error	DF	t-value	p-value
(Intercept)	16.152435	0.9984616	79	16.177323	0.0000
age	0.797950	0.0870677	79	9.164702	0.0000
SexFemale	1.264698	1.5642886	25	0.808481	0.4264
age:SexFemale	-0.322243	0.1364089	79	-2.362334	0.0206

Correlation:

(Intr) age	SexFml
age	-0.877
SexFemale	-0.638 0.559
age:SexFemale	0.559 -0.638 -0.877

Standardized Within-Group Residuals:

Min	Q1	Med	Q3	Max
-3.288886631	-0.419431536	-0.001271185	0.456257976	4.203271248

Number of Observations: 108

Number of Groups: 27

## Correlation structures

```
>intervals(m_lme)
Approximate 95% confidence intervals
Fixed effects:
```

	lower	est.	upper
(Intercept)	14.1650475	16.1524355	18.13982351
age	0.6246456	0.7979496	0.97125348
SexFemale	-1.9570145	1.2646982	4.48641100
age:SexFemale	-0.5937584	-0.3222434	-0.05072829

```
attr(,"label")
[1] "Fixed effects:"

Random Effects:
Level: Subject
```

	lower	est.	upper
sd((Intercept))	2.2057787	3.3730482	5.1580217
sd(age)	0.1848096	0.2907673	0.4574740
cor((Intercept),age)	-0.9377525	-0.8309622	-0.5806162

```
Correlation structure:
```

	lower	est.	upper
Phi	-0.7560625	-0.47328	-0.04159463

## Correlation Matrix G

```
> VarCorr( m_lme )
Subject = pdLogChol(1 + age)
Variance   StdDev    Corr
(Intercept) 11.3774543 3.3730482 (Intr)
age          0.0845456 0.2907673 -0.831
Residual     1.1924103 1.0919754

> getVarCov( m_lme )
Random effects variance covariance matrix
(Intercept)      age
(Intercept)    11.37700 -0.814980
age             -0.81498  0.084546

Standard Deviations: 3.373 0.29077
```

## Check and Study

- ▶ Variogram nlme: Calculate Semi-variogram
- ▶ corClasses nlme: Correlation Structure Classes

## Summary of the example:

- ▶ The estimated autocorrelation is negative and  $\rho = -0.47328$ .
- ▶ Strong positive autocorrelation can be a symptom of lack of fit.
- ▶ Occasional large measurement errors will contribute negatively to the estimate of autocorrelation.

## Random effects approach in Longitudinal Data Analysis:

- ▶ Good method for short series.
- ▶ Uses all observations
- ▶ Usually not good for long series.



### **Final Project Assignment.**

Discussion about the final project - individual consultation.

### **Solve pig problem: .**

Investigate the effect of injection of Porcine Growth Hormone (PGH) on pH. Experiment was carried out with two pigs from each of 6 litters. There were two treatments: 1) control and 2) pgh (daily injection with 0.08 mg PGH)

The pH in the meat was measured 20 times from 30 minutes after until 24 hours after slaughter. There were 10 litters in the experiment but pH was measured for only 6 of these. The order of the data is: treatment, litter, pig number, followed by pH measurements at 30, 45, 60, 75, 90, 105, 120, 150, 180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480, 1440 minutes after slaughter.

In this analysis the focus should be on the effect of the treatment over time.

### Solve pig problem.

1. Make one or more plots of the data. Comment on the plot(s).
2. Setup a suitable model for this data set, including both fixed and random effects, but no correlation structure. (Notice that besides the “pig” variable we also have information about litter, which could be included as an additional random effect.)
3. Reduce the initial model (if possible), both the random effects and fixed effects parts.
4. Extend the model by adding a correlation structure.
5. Use information criteria and/or semi-variograms to select an appropriate correlation structure.
6. Explain in words the correlation structure that was chosen.
7. Repeat the model reduction process.
8. What is the conclusion about the treatment?

Hint: Use time transformation:  $\frac{1}{\log()}$ .