

# Chapter 2

## Marginal Models for Continuous Data

## 2.1 Simple Methods

---

- The reason why classical statistical techniques fail in the context of longitudinal data is that observations within subjects are correlated
  - ▷ often the correlation between two repeated measurements decreases as the time span between those measurements increases
- The paired  $t$ -test accounts for this by considering subject-specific differences
$$\Delta_i = Y_{i1} - Y_{i2}$$
  - ▷ this reduces the number of measurements to just one per subject, which implies that classical techniques can be applied again

## 2.1 Simple Methods (cont'd)

---

- In the case of more than 2 measurements per subject, similar simple techniques are often applied to reduce the number of measurements for the  $i$ -th subject, from  $n_i$  to 1
  - ▷ Analysis at each time point separately
  - ▷ Analysis of Area Under the Curve (AUC)
  - ▷ Analysis of endpoints
  - ▷ Analysis of increments

## 2.1 Simple Methods (cont'd)

---

- **Analysis at each time point separately**

- ▷ **General idea:** The data are analyzed at each occasion separately

- ▷ **Advantages:**

- \* simple to interpret
    - \* uses all available data

- Disadvantages:**

- \* does not consider 'overall' differences
    - \* does not allow to study the evolution of differences
    - \* problem of multiple testing
    - \* possible problems with missing data

## 2.1 Simple Methods (cont'd)

---

- **Analysis of area under the curve (AUC)**

- ▷ **General idea:** For each subject, the area under her curve is calculated

$$\text{AUC}_i = (t_{i2} - t_{i1}) \times (y_{i2} + y_{i1})/2 + (t_{i3} - t_{i2}) \times (y_{i3} + y_{i2})/2 + \dots$$

Afterwards, these AUCs are analyzed

- ▷ **Advantages:**

- \* no problems of multiple testing
    - \* does not explicitly assume balanced data
    - \* compares 'overall' differences

## 2.1 Simple Methods (cont'd)

---

- Analysis of area under the curve (AUC)

- ▷ **Disadvantages:**

- \* subjects could have the same AUC but completely different profiles
    - \* possible problems with missing data

## 2.1 Simple Methods (cont'd)

---

- **Analysis of endpoints**

- ▷ **General idea:** Assess differences only on the last time point

- ▷ **Advantages:**

- \* no problems of multiple testing
    - \* does not explicitly assume balanced data

- Disadvantages:**

- \* applicable only in randomized trials
    - \* uses partial information
    - \* the last time point must be the same for all subjects
    - \* does not consider 'overall' differences
    - \* possible problems with missing data

## 2.1 Simple Methods (cont'd)

---

- **Analysis of increments**

- ▷ **General idea:** A simple method to compare evolutions between subjects, correcting for differences at baseline, is to analyze the subject-specific changes

$$y_{in_i} - y_{i1}$$

- ▷ **Advantages:**

- \* no problems of multiple testing
- \* does not explicitly assume balanced data

- Disadvantages:**

- \* uses partial information
- \* the last time point must be the same for all subjects
- \* possible problems with missing data



## 2.1 Simple Methods (cont'd)

---

- The AUC, endpoints and increments are examples of summary statistics
  - ▷ these statistics summarize the vector of repeated measurements for each subject separately
- This leads to the following general procedure:
  - ▷ **Step 1:** Summarize the data of each subject into one statistic
  - ▷ **Step 2:** Analyze the summary statistics, e.g. analysis of covariance to compare groups after correction for important covariates
- This way, the analysis of longitudinal data is reduced to the analysis of independent observations, for which classical statistical procedures are available

## 2.1 Simple Methods (cont'd)

---

- However, all these methods have the disadvantage that (lots of) information is lost

**This has led to the development of statistical techniques that overcome these disadvantages**

## 2.1 Simple Methods (cont'd)

---

- These techniques are based on extensions of simple regression models for univariate data
- Before introducing these extensions we start with a short review of the classical *linear regression model* for continuous outcomes. . .

## 2.2 Review of Linear Regression

---

- Suppose we have a continuous outcome  $Y$  measured *cross-sectionally*
  - ▷ Example: The serum bilirubin levels from the PBC dataset at baseline (i.e., time  $t = 0$ )
- We are interested in making statistical inferences for this outcome, e.g.,
  - ▷ is there any difference between placebo and D-penicillamine corrected for the age and sex of the patients?
  - ▷ which factors best predict serum bilirubin levels?



**Linear Regression Model**

## 2.2 Review of Linear Regression (cont'd)

---

- Definition of the linear regression model

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

where

- ▷  $y_i$  denotes the outcome for subject  $i$
- ▷  $x_{i1}, \dots, x_{ip}$  denote the  $p$  covariates for subject  $i$
- ▷  $\beta_0, \beta_1, \dots, \beta_p$  the regression coefficients
- ▷  $\varepsilon_i$  the error term for subject  $i$

## 2.2 Review of Linear Regression (cont'd)

---

- **Example:** For the PBC patients we postulate the linear regression model

$$\log(\text{serBilir}_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{D-penicil}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

where

- ▷  $\text{serBilir}_i$  denotes the serum bilirubin of patient  $i$  at baseline
- ▷  $\text{Age}_i$  and  $\text{D-penicil}_i$  denote the Age and whether patient  $i$  received D-penicil or placebo
- ▷  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  are the regression coefficients
- ▷  $\varepsilon_i$  are the error terms

## 2.2 Review of Linear Regression (cont'd)

---

- Behind this model there are several assumptions, some obvious, some hidden. In particular:
  - ▷ serum bilirubin is assumed to be only related to Age and treatment
  - ▷ the relation between serum bilirubin and Age is linear
  - ▷ the effect of Age is the same whatever the treatment the patient took, and vice versa
  - ▷ the error terms are normally distributed
  - ▷ the variance of the error terms does not depend on neither Age nor D-penicillamine
  - ▷ **measurements are independent of each other**

## 2.2 Review of Linear Regression (cont'd)

---

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.5395	0.2824	1.91	0.0570
age	0.0015	0.0056	0.28	0.7817
drugD-penicil	-0.0933	0.1174	-0.79	0.4274

- Interpretation

- ▷  $\beta_0 = 0.5$  average log(Ser. Bilir.) for Age = 0 and placebo patients
- ▷  $\beta_1 = 0.0015$  increase in average log(Ser. Bilir.) for every year increase for patients with the same treatment
- ▷  $\beta_2 = -0.1$  decrease in average log(Ser. Bilir.) when receiving D-penicil versus placebo for patients of the same age



## 2.2 Review of Linear Regression (cont'd)

---

- Linear regression model with *matrix notation*
  - ▷ the linear regression model for the  $n$  subjects

$$y_1 = \beta_0 + \beta_1 x_{11} + \dots + \beta_p x_{1p} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \dots + \beta_p x_{2p} + \varepsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \dots + \beta_p x_{np} + \varepsilon_n$$

## 2.2 Review of Linear Regression (cont'd)

- Linear regression model with *matrix notation*
  - ▷ the linear regression model for the  $n$  subjects

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ & \vdots & & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

## 2.2 Review of Linear Regression (cont'd)

---

- Linear regression model with *matrix notation*
  - ▷  $\mathbf{y}$ : response vector
  - ▷  $\mathbf{X}$ : design matrix
  - ▷  $\boldsymbol{\beta}$ : parameter vector
  - ▷  $\boldsymbol{\varepsilon}$ : measurement error vector

More on linear algebra?  $\Rightarrow$  Check the videos: <https://goo.gl/4zQfiu>

## 2.2 Review of Linear Regression (cont'd)

---

- Maximum likelihood estimators

$$\begin{cases} \hat{\beta} = (X^{\top} X)^{-1} X^{\top} y \\ \hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^{\top} (y - X\hat{\beta}) \end{cases}$$

where

- ▷  $X^{\top}$  denotes the *transpose* of matrix  $X$
- ▷  $X^{\top} X$  denotes the *matrix product* between matrices  $X^{\top}$  and  $X$
- ▷  $(X^{\top} X)^{-1}$  denotes the *matrix inverse* of matrix  $(X^{\top} X)$

## 2.3 Marginal Models

---

- Let's go back to the independence assumption
  - ▷ the first five rows of the data are:

id	serBilir	age	drug
1	14.50	58.77	D-penicil
2	1.10	56.45	D-penicil
3	1.40	70.07	D-penicil
4	1.80	54.74	D-penicil
5	3.40	38.11	placebo

Each row represents a different patient, and patients are **independent** of each other

## 2.3 Marginal Models (cont'd)

---

- When we have repeated measurements data, we have the form

id	serBilir	year	age	drug
1	14.50	0.00	58.77	D-penicil
1	21.30	0.53	58.77	D-penicil
2	1.10	0.00	56.45	D-penicil
2	0.80	0.50	56.45	D-penicil
2	1.00	1.00	56.45	D-penicil
2	1.90	2.10	56.45	D-penicil
2	2.60	4.90	56.45	D-penicil

## 2.3 Marginal Models (cont'd)

---

Multiple rows per subject, rows belonging to the same subject are **correlated**

- Note: Long vs Wide format
  - ▷ wide format can only be used when all subjects are measured at the same time points
  - ▷ long format can always be used
  - ▷ (almost) all software packages accept repeated measurements data in long format

## 2.3 Marginal Models (cont'd)

---

- How correlation affects modeling of the data?
- Say we are interested in the effect of time on serum bilirubin while also correcting for the age of the patients
  - ▷ the corresponding regression equation is

$$\log(\text{serBilir}_{ij}) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Age}_i + \varepsilon_{ij}$$

where

- \*  $\text{serBilir}_{ij}$  denotes the level of serum bilirubin of patient  $i$  at time point  $\text{Time}_{ij}$
- \*  $\varepsilon_{ij}$  is the corresponding error term



## 2.3 Marginal Models (cont'd)

- The fact that the responses of each patient are correlated translates to error terms that are correlated
  - ▷ based on the data of the first two patients (see pp.49) we have

$$\begin{bmatrix} 14.5 \\ 21.3 \\ 1.1 \\ 0.8 \\ 1.0 \\ 1.9 \\ 2.6 \end{bmatrix} = \begin{bmatrix} 1 & 0.0 & 58.8 \\ 1 & 0.5 & 58.8 \\ 1 & 0.0 & 56.5 \\ 1 & 0.5 & 56.5 \\ 1 & 1.0 & 56.5 \\ 1 & 2.1 & 56.5 \\ 1 & 4.9 & 56.5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix}$$

## 2.3 Marginal Models (cont'd)

---

- The direct approach to account for correlated data  $\Rightarrow$  *multivariate regression*

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, V_i),$$

where

- ▷  $y_i$  the vector of responses for the  $i$ -th subject
- ▷  $X_i$  design matrix describing the structural component
- ▷  $V_i$  covariance matrix describing the variance and correlation structures

**The covariance matrix  $V_i$  explicitly accounts for the correlations**

## 2.4 Interpretation

---

- Interpretation of  $\beta$ 
  - ▷  $\beta_j$  denotes the change in the average  $y_i$  when  $x_j$  is increased by one unit and all other covariates are fixed
- Example: In the AIDS dataset we are interested in the effect of treatment on the average longitudinal evolutions – we fit a marginal model with
  - ▷ different average longitudinal evolutions per treatment group ( $X\beta$  part)
  - ▷ compound symmetry covariance matrix ( $V_i$  part)

$$\left\{ \begin{array}{l} \sqrt{\text{CD4}}_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \{\text{ddI}_i \times \text{Time}_{ij}\} + \varepsilon_{ij}, \\ \varepsilon_i \sim \mathcal{N}(0, V_i) \end{array} \right.$$

## 2.4 Interpretation (cont'd)

	Value	Std.Err.	t-value	p-value
$\beta_0$	7.189	0.221	32.593	< 0.001
$\beta_1$	-0.156	0.017	-9.247	< 0.001
$\beta_2$	0.016	0.024	0.662	0.508

- ▷ Coefficient  $\beta_1$ : For patients in the ddC group, every month the average  $\sqrt{\text{CD4}}$  changes by  $-0.156$
- ▷ Coefficient  $\beta_2$ :
  - \* Is the difference of the time effect between ddl and ddC
  - \* For patients in the ddl group, every month the average  $\sqrt{\text{CD4}}$  changes by  $(-0.156 + 0.016)$

## 2.4 Interpretation (cont'd)

---

- The estimated covariance matrix  $V_i$  is

	$t = 0$	$t = 2$	$t = 6$	$t = 12$	$t = 18$
$t = 0$	24.15	20.30	20.30	20.30	20.30
$t = 2$	20.30	24.15	20.30	20.30	20.30
$t = 6$	20.30	20.30	24.15	20.30	20.30
$t = 12$	20.30	20.30	20.30	24.15	20.30
$t = 18$	20.30	20.30	20.30	20.30	24.15

$$\triangleright \text{corr}(CD4_{t=0}, CD4_{t=2}) = \frac{\text{cov}(CD4_{t=0}, CD4_{t=2})}{\sqrt{\text{var}(CD4_{t=0})} \sqrt{\text{var}(CD4_{t=2})}} = \frac{20.3}{24.15} = 0.84$$

## 2.4 Interpretation (cont'd)

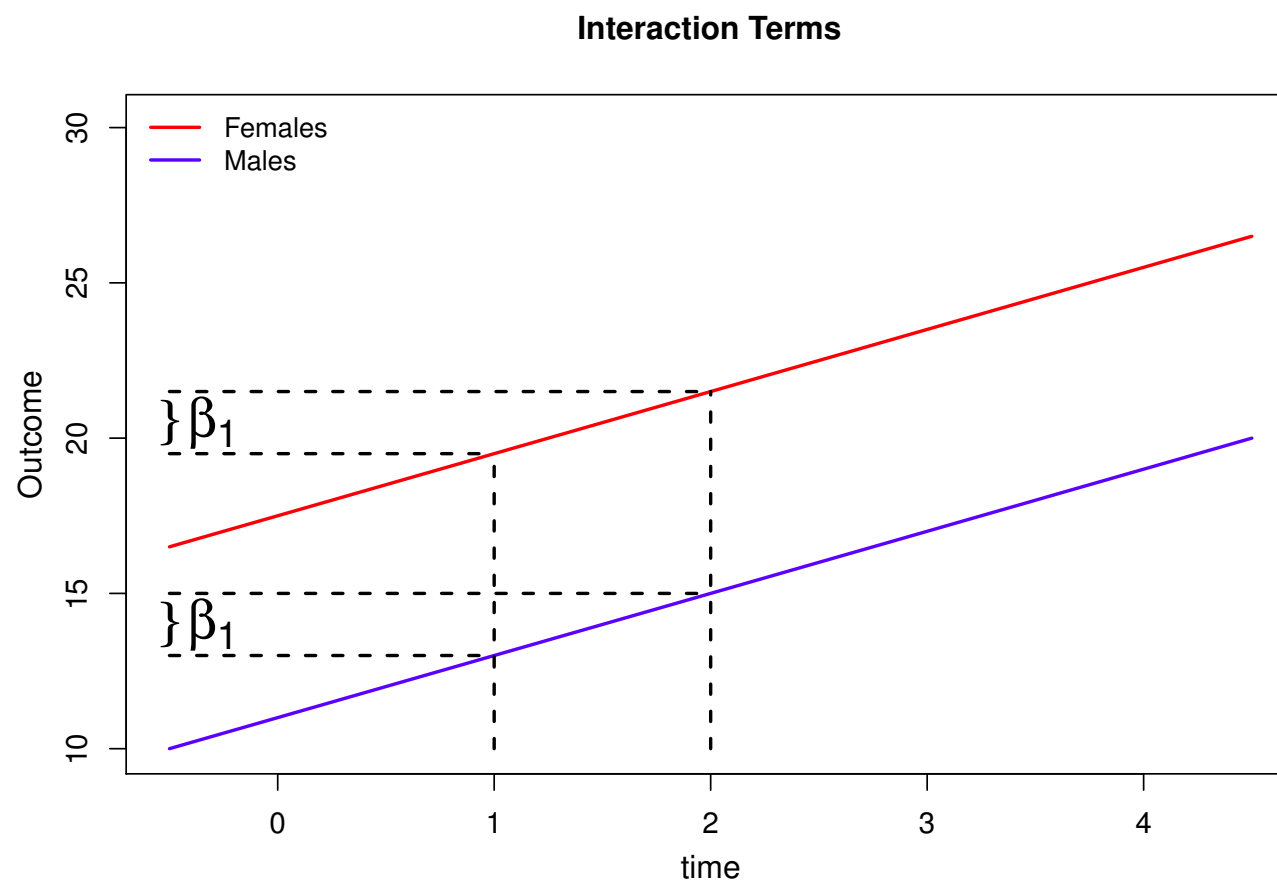
---

- Note: Interaction terms for longitudinal data
  - ▷ Consider the model

$$y_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Sex}_i + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, V_i)$$

- \* we include the time effect and we also control for sex
- \* the model assumes that the effect of time is the same for the two sexes  
(*parallel lines*)

## 2.4 Interpretation (cont'd)



## 2.4 Interpretation (cont'd)

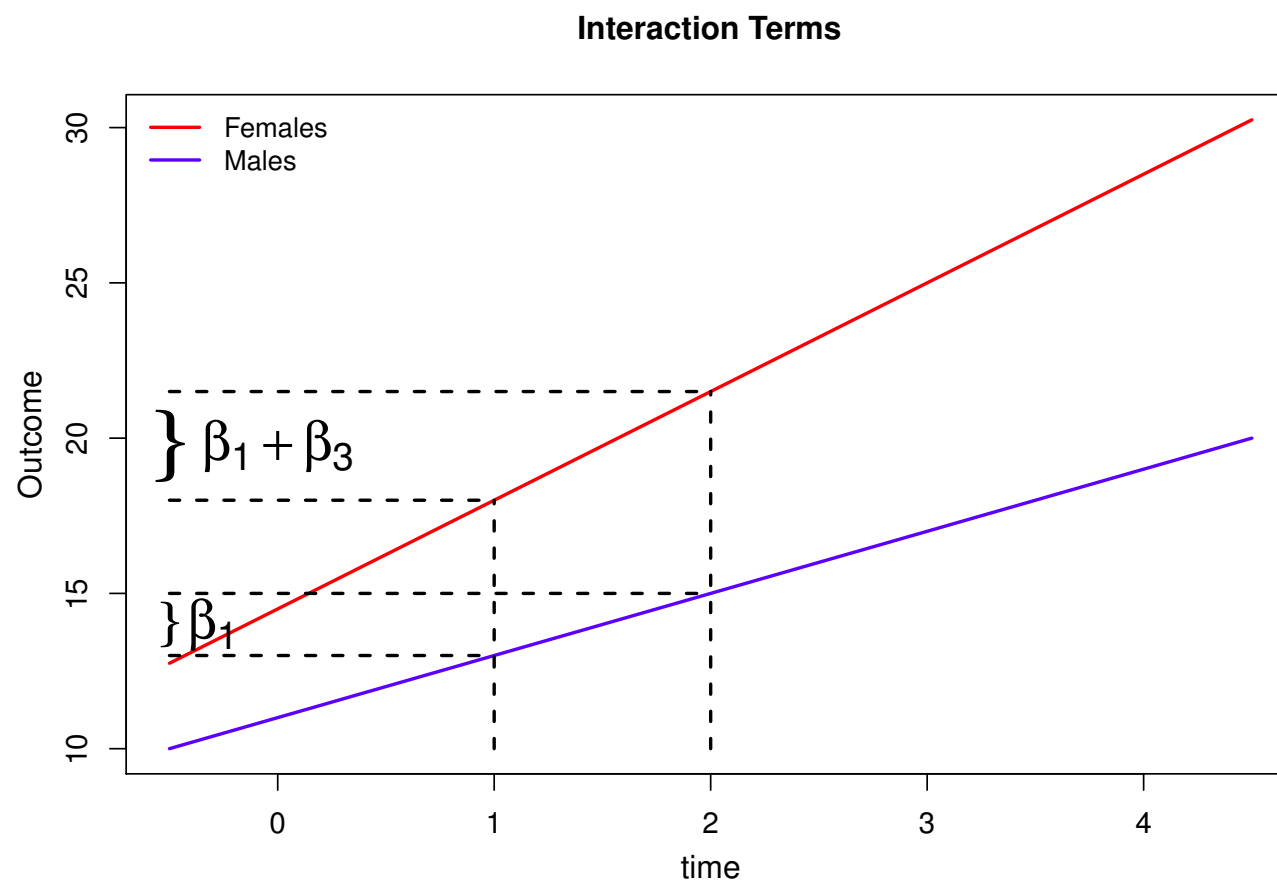
---

- Note: Interaction terms for longitudinal data
  - ▷ if we would like different longitudinal evolutions for the two sexes we need to include the *interaction term*

$$y_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Sex}_i + \beta_3 \{\text{Sex}_i \times \text{Time}_{ij}\} + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, V_i)$$



## 2.4 Interpretation (cont'd)



## 2.4 Interpretation (cont'd)

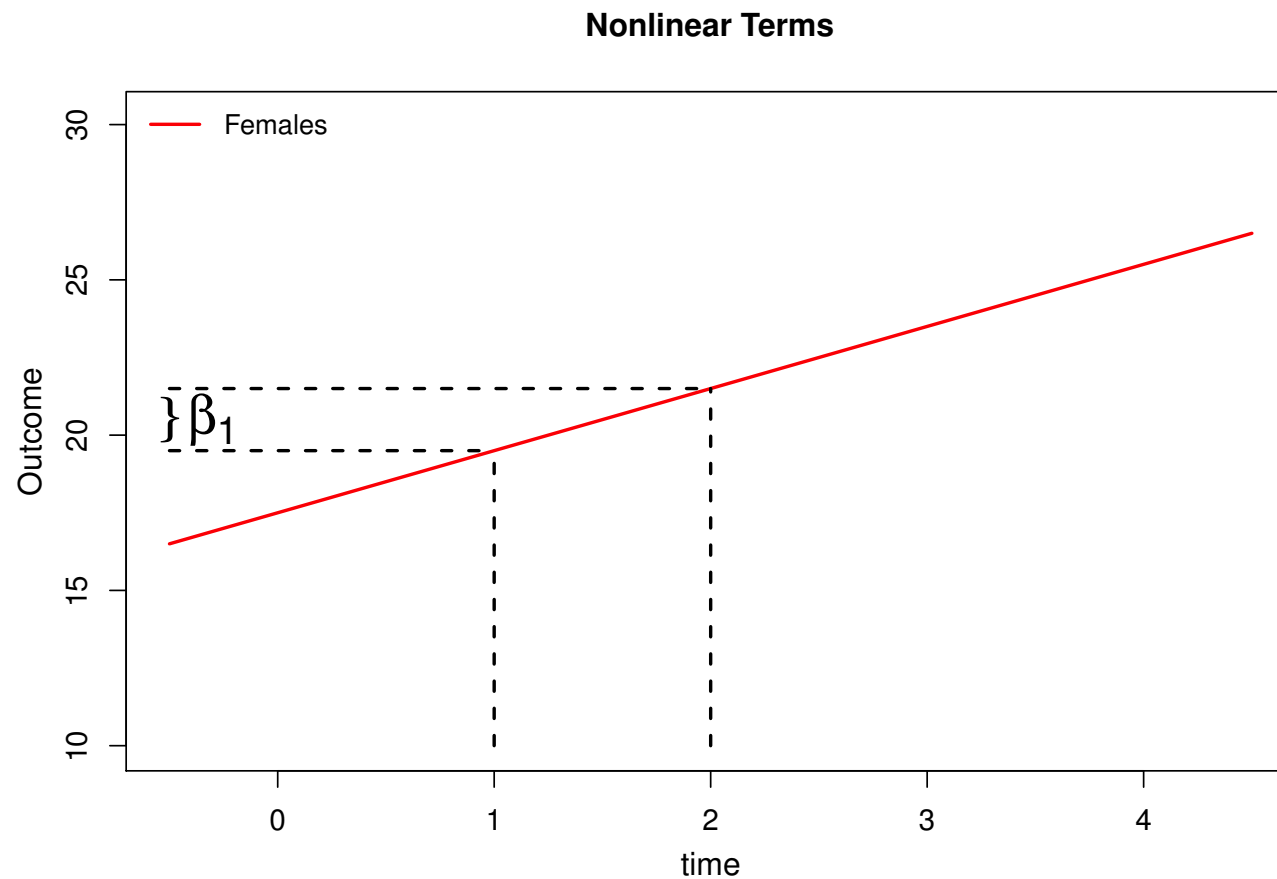
---

- Note: Nonlinear terms for longitudinal data
  - ▷ Consider the model

$$y_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Sex}_i + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, V_i)$$

- \* we include the time effect and we also control for sex
- \* the model assumes that the effect of time is linear

## 2.4 Interpretation (cont'd)



## 2.4 Interpretation (cont'd)

---

- Note: Nonlinear terms for longitudinal data
  - ▷ to relax this assumption, we need to include **nonlinear terms** of time
  - ▷ two popular choices are
    - \* **polynomials**

$$y_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Time}_{ij}^2 + \beta_3 \text{Time}_{ij}^3 + \beta_4 \text{Sex}_i + \varepsilon_{ij}$$

- \* and **splines**

$$y_{ij} = \beta_0 + \beta_1 N(\text{Time}_{ij})_1 + \beta_2 N(\text{Time}_{ij})_2 + \beta_3 N(\text{Time}_{ij})_3 + \beta_4 \text{Sex}_i + \varepsilon_{ij}$$

## 2.4 Interpretation (cont'd)

---

- Brief background on splines:
  - ▷ splines are *local* polynomials
  - ▷ *local* means that we split the follow-up period in a number of intervals
  - ▷ the limits of these intervals are defined from the *knots* of the spline
    - \* we have two boundary notes, and
    - \* a number of internal knots
  - ▷ in each interval we assume a polynomial (typically cubic)
  - ▷ restrictions are put such that the polynomials in each interval connect with each other

## 2.4 Interpretation (cont'd)

---

- In both polynomials and splines, increasing
  - ▷ the degree in the former, and
  - ▷ the number of internal knots in the latterallows the time effect to be modeled more flexibly
- **However**, we should not overdo it because of the risk of over-fitting
  - ▷ in the majority of the cases, a 2nd or 3rd degree polynomial or 2 or 3 internal knots are sufficient to capture nonlinearities

**From the two approaches, splines are preferable**

## 2.4 Interpretation (cont'd)

---

- Note: How to place the knots in splines
  - ▷ *Boundary knots*:
    - \* By default (i.e., what function `ns()` in R does), these are placed in the minimum and maximum follow-up times
    - \* **However**, this default choice may lead to problems when very few subjects have long profiles, and the majority has much shorter ones
    - \* In these cases, place the boundary knots at the 5% and 95% percentiles of the follow-up times

## 2.4 Interpretation (cont'd)

---

- Note: How to place the knots in splines
  - ▷ *internal knots*:
    - \* By default (i.e., what function `ns()` in R does), these are placed in percentiles follow-up times
    - \* This is a sensible choice
    - \* **However**, some times the placing of these knots may be driven by subject-matter knowledge



## 2.4 Interpretation (cont'd)

---

- **Communicating a model with complex terms:** Due to the elaborate structure of repeated measurements data it is often required to include complex terms in a model
  - ▷ interaction terms (e.g., between baseline and time-varying predictors)
  - ▷ nonlinear terms (e.g., nonlinear evolutions over time modeled with polynomials or splines)
- In such cases the regression coefficients  $\beta$  we obtain in the output do not often have a straightforward interpretation

## 2.4 Interpretation (cont'd)

- To overcome this issue we can use **effect plots**
  - ▷ this is a figure that depicts the average outcome along with 95% confidence intervals for specific combinations of the predictors' levels
- Example: We have fitted the following model to the PBC dataset:

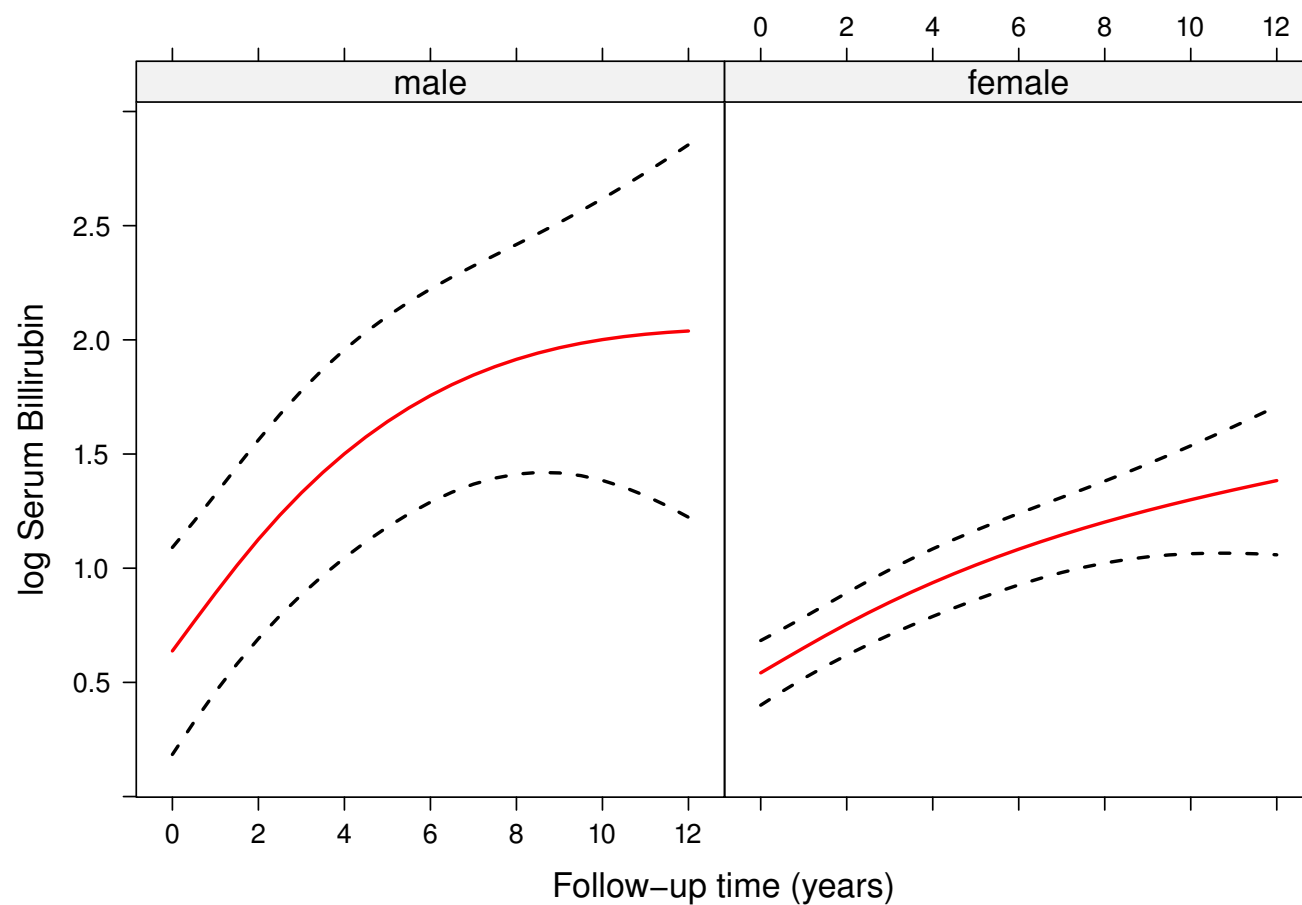
$$\left\{ \begin{array}{l} \log(\text{serBilir}_{ij}) = \beta_0 + \beta_1 N(\text{Time}_{ij})_1 + \beta_2 N(\text{Time}_{ij})_2 + \beta_3 \text{Female}_i + \beta_4 \text{Age}_i + \\ \quad \beta_5 \{ \text{Female}_i \times N(\text{Time}_{ij})_1 \} + \beta_6 \{ \text{Female}_i \times N(\text{Time}_{ij})_2 \} + \\ \quad \beta_7 \{ \text{Female}_i \times \text{Age}_i \} + \varepsilon_{ij} \\ \\ \varepsilon_i \sim \mathcal{N}(0, V_i) \quad V_i \text{ has a continuous AR1 structure} \end{array} \right.$$

## 2.4 Interpretation (cont'd)

---

- The terms  $N(\text{Time}_{ij})_1$  and  $N(\text{Time}_{ij})_2$  denote the basis for a natural cubic spline with two degrees of freedom to model possible nonlinearities in the time effect
- In this model not all coefficients have a direct interpretation in isolation
- Hence to understand the model we depict
  - ▷ how the average longitudinal profiles evolve over time,
  - ▷ separately for males and females, and
  - ▷ for the average age of 49 years old (in the app different ages can be selected)
  - ▷ including also the corresponding 95% pointwise confidence intervals

## 2.4 Interpretation (cont'd)



## 2.5 Estimation

---

- Estimation of model parameters
  - ▷ For known covariance matrix  $V_i$ , the regression coefficients are estimated using generalized least squares

$$\hat{\beta} = \left( \sum_{i=1}^n X_i^T V_i^{-1} X_i \right)^{-1} \sum_{i=1}^n X_i^T V_i^{-1} y_i$$

- ▷ Variance Components – matrix  $V_i$ :
  - \* Maximum Likelihood (ML)
  - \* restricted maximum likelihood (REML)

## 2.5 Estimation (cont'd)

---

- What's the difference between ML and REML?
  - ▷ ML estimates of variances are known to be biased in small samples
  - ▷ the simplest case: Sample variance

$$\text{var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- ▷ to obtain an unbiased estimate we need to divide by  $n-1$  because we estimate the mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

## 2.5 Estimation (cont'd)

---

**The REML estimation is a generalization of this idea**

- It provides unbiased estimates of the parameters in the covariance matrix  $V_i$  in small samples
- **Example:** To illustrate the difference between REML and ML we consider fitting the same model for the AIDS dataset we have seen before but using only the first 50 rows

## 2.5 Estimation (cont'd)

### ▷ REML Estimation

	$t = 0$	$t = 2$	$t = 6$	$t = 12$	$t = 18$
$t = 0$	16.03	13.48	13.48	13.48	13.48
$t = 2$	13.48	16.03	13.48	13.48	13.48
$t = 6$	13.48	13.48	16.03	13.48	13.48
$t = 12$	13.48	13.48	13.48	16.03	13.48
$t = 18$	13.48	13.48	13.48	13.48	16.03



## 2.5 Estimation (cont'd)

### ▷ ML Estimation

	$t = 0$	$t = 2$	$t = 6$	$t = 12$	$t = 18$
$t = 0$	14.97	12.56	12.56	12.56	12.56
$t = 2$	12.56	14.97	12.56	12.56	12.56
$t = 6$	12.56	12.56	14.97	12.56	12.56
$t = 12$	12.56	12.56	12.56	14.97	12.56
$t = 18$	12.56	12.56	12.56	12.56	14.97

- \* We observe some visible differences because of small  $n$
- \* In the full dataset the differences are negligible

## 2.5 Estimation (cont'd)

---

- Features of REML estimation:
  - ▷ Available in all software that fit marginal and mixed effects models
  - ▷ The way it works is by applying a transformation in the longitudinal outcome  $y$  based on the chosen structure of the design matrix  $X$  (i.e., which predictors you have included in the model)
  - ▷ **Hence, we cannot compare the likelihoods of models fitted with REML and have different  $X\beta$  part**

## 2.6 Fitting Marginal Models in R

---

R> Marginal models can be fitted using function `gls()` from the **nlme** package

R> It has four basic arguments

- ▷ `model`: a formula specifying the response vector and the covariates to include in the model
- ▷ `data`: a data frame containing all the variables
- ▷ `correlation`: a function describing the assumed correlation structure
- ▷ `weights`: a function describing the assumed within-group heteroscedasticity structure

## 2.6 Fitting Marginal Models in R (cont'd)

**R>** The data frame that contains all variables should be in the *long format*

Subject	y	time	gender	age
1	5.1	0.0	male	45
1	6.3	1.1	male	45
2	5.9	0.1	female	38
2	6.9	0.9	female	38
2	7.1	1.2	female	38
2	7.3	1.5	female	38
⋮	⋮	⋮	⋮	⋮

## 2.6 Fitting Marginal Models in R (cont'd)

---

R> Using formulas in R

▷ CD4 = Time + Gender

⇒ `cd4 ~ time + gender`

▷ CD4 = Time + Gender + Time\*Gender

⇒ `cd4 ~ time + gender + time:gender`

⇒ `cd4 ~ time*gender` (the same)

▷ CD4 = Time + Time<sup>2</sup>

⇒ `cd4 ~ time + I(time^2)`

R> Note: the intercept term is included by default

## 2.6 Fitting Marginal Models in R (cont'd)

---

**R>** The following code fits a marginal model for the square root CD4 cell count with a compound symmetry correlation structure

```
glsFit <- gls(CD4 ~ obstime + obstime:drug, data = aids,  
             correlation = corCompSymm(form = ~ obstime | patient))
```

```
summary(glsFit)
```

(Note: In the aids database CD4 is the square root transformed CD4 cell count)

## 2.7 Covariance Matrix

---

- Reminder: What is a variance-covariance matrix?
  - ▷ we have the dataset:

Subject	$Y_1$	$Y_2$	$Y_3$	$Y_4$
1	2.1	3.2	2.9	3.3
2	1.8	3.1	4.2	5.1
3	3.1	3.2	3.5	3.3
⋮	⋮	⋮	⋮	⋮

## 2.7 Covariance Matrix (cont'd)

- The variance-covariance matrix is the matrix whose element in the  $i, j$ -th position is the covariance between  $Y_i$  and  $Y_j$ , e.g.,

$$\begin{bmatrix} \text{var}(Y_1) & \text{cov}(Y_1, Y_2) & \text{cov}(Y_1, Y_3) & \text{cov}(Y_1, Y_4) \\ \text{cov}(Y_2, Y_1) & \text{var}(Y_2) & \text{cov}(Y_2, Y_3) & \text{cov}(Y_2, Y_4) \\ \text{cov}(Y_3, Y_1) & \text{cov}(Y_3, Y_2) & \text{var}(Y_3) & \text{cov}(Y_3, Y_4) \\ \text{cov}(Y_4, Y_1) & \text{cov}(Y_4, Y_2) & \text{cov}(Y_4, Y_3) & \text{var}(Y_4) \end{bmatrix}$$

- Properties
  - ▷ on the diagonal the **variances**, off diagonal **covariances**
  - ▷ symmetric  $\Rightarrow \text{cov}(Y_1, Y_2) = \text{cov}(Y_2, Y_1)$



## 2.7 Covariance Matrix (cont'd)

---

- Variances, covariances and correlations
  - ▷ **variance** measures how far a set of numbers is spread out (always positive)
  - ▷ **covariance** is a measure of how much two random variables change together (positive or negative)
  - ▷ **correlation** a measure of the linear correlation (dependence) between two variables (between  $-1$  and  $1$ ;  $0$  no correlation)

$$\text{corr}(Y_1, Y_2) = \frac{\text{cov}(Y_1, Y_2)}{\sqrt{\text{var}(Y_1)} \sqrt{\text{var}(Y_2)}}$$

## 2.7 Covariance Matrix (cont'd)

---

- Due to the fact that the magnitude of the covariance between  $Y_1$  and  $Y_2$  depends on their variability, we translate the covariance matrix into a correlation matrix

$$\begin{bmatrix} 1 & \text{corr}(Y_1, Y_2) & \text{corr}(Y_1, Y_3) & \text{corr}(Y_1, Y_4) \\ & 1 & \text{corr}(Y_2, Y_3) & \text{corr}(Y_2, Y_4) \\ & & 1 & \text{corr}(Y_3, Y_4) \\ & & & 1 \end{bmatrix}$$

## 2.7 Covariance Matrix (cont'd)

---

- Coming back to our model

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, V_i)$$

- We need an appropriate choice for  $V_i$  in order to appropriately describe the correlations between the repeated measurements
  - ▷ compound symmetry
  - ▷ autoregressive process
  - ▷ exponential spatial correlation
  - ▷ Gaussian spatial correlation
  - ▷ Toeplitz
  - ▷ ...

## 2.7 Covariance Matrix (cont'd)

- Let's see some of those
  - ▷ General/Unstructured

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

- ▷ Diagonal

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

## 2.7 Covariance Matrix (cont'd)

---

▷ First-order autoregressive

$$\begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

▷ Toeplitz

$$\begin{bmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \rho_2\sigma_1\sigma_3 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 & \rho_1\sigma_2\sigma_3 \\ \rho_2\sigma_1\sigma_3 & \rho_1\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma^2 & \sigma_{12} \\ \sigma_{13} & \sigma_{12} & \sigma^2 \end{bmatrix}$$

## 2.7 Covariance Matrix (cont'd)

---

- The aforementioned structures for the covariance matrix are applicable in cases we have discrete and equally spaced time points
- For continuous time and unbalanced data, alternative options are:
  - ▷ continuous AR1
  - ▷ exponential serial correlation
  - ▷ linear correlation
  - ▷ Gaussian serial correlation

## 2.7 Covariance Matrix (cont'd)

---

- These serial correlation structures are defined using the semi-variogram
  - ▷ which we are not going to cover here because it is a bit technical (more info in any standard text for mixed models / longitudinal data analysis)
- The basic assumption is that correlations decay with the time lag  $|t_i - t_j| \Rightarrow$  measurements at closer time points are more strongly correlated than measurements at more distant time points
  - ▷ the aforementioned structures for unbalanced data have one parameter that controls how the correlations decay over time

## 2.7 Covariance Matrix (cont'd)

---

- Notes: On building covariance matrices
  - ▷ *variance function*: in some cases, and especially for longitudinal data, it may **not** be reasonable to assume that the variance of the outcome remains constant over time
    - \* we have seen versions of heteroscedastic covariance matrices, but these are only applicable when we have balanced data and few time points
    - \* for unbalanced designs we can specify other variance functions, e.g., that variances increase linearly or exponentially over time
  - ▷ *correlation at the same point*: is it **always** reasonable that the correlation of the outcome at the same point is set to 1?



## 2.7 Covariance Matrix (cont'd)

---

- Let's try the app...

## 2.8 Model Building

---

- We have seen that marginal models consist of two parts:
  - ▷ Mean part –  $X\beta$ : that describes how covariates we have put in the model explain the average of the repeated measurements
  - ▷ Covariance part –  $V_i$ : assumed covariance structure between the repeated measurements
- In the majority of the cases scientific interest focuses on the mean part

**However, to obtain valid and efficient inferences for the mean part, the covariance part needs to be adequately specified**

## 2.8 Model Building (cont'd)

---

- Hence, the general strategy for building models for repeated measurements data proceeds as follows:
  1. Put all the covariates of interest in the mean part, considering possible nonlinear and interaction terms – **do NOT** remove the ones that are not significant
  2. Then select an appropriate covariance matrix  $V_i$  that adequately describes the correlations in the repeated measurements
    - \* in this step you should be a bit anti-conservative, i.e., do not favor a simpler covariance matrix if the  $p$ -value is just non-significant
  3. Finally, return to the mean part and exclude non significant covariates
    - \* first start by testing the interaction terms, and
    - \* then the nonlinear terms

## 2.8 Model Building (cont'd)

---

- How many coefficients can we reliably estimate in the mean part?
- It depends on how strong the correlations between the repeated measurements are
  - ▷ weak correlations  $\Rightarrow N/10$  ( $N$  total number of measurements)
  - ▷ strong correlations  $\Rightarrow n/10$  ( $n$  number of subjects)

## 2.9 Hypothesis Testing

---

- Having fitted a marginal model using maximum likelihood we can use standard inferential tools for performing hypothesis testing
  - ▷ Wald tests / t-tests / F-tests
  - ▷ Score tests
  - ▷ Likelihood ratio tests
- Following the model building strategy described above, we will
  - ▷ first, describe how we can choose the appropriate covariance matrix, and
  - ▷ then focus on hypothesis testing for the mean part of the model

## 2.9 Hypothesis Testing (cont'd)

---

- **Hypothesis testing for  $V_i$ :** Assuming the same mean structure we can fit a series of models and choose the one that best describes the covariances
- In general, we distinguish between two cases
  - ▷ comparing two models with *nested* covariance matrices
  - ▷ comparing two models with *non-nested* covariance matrices
- **Note:** Model A is nested in Model B, when Model A is a special case of Model B
  - ▷ i.e., by setting some of the parameters of Model B at some specific value we obtain Model A

## 2.9 Hypothesis Testing (cont'd)

---

- For **nested** models the preferable test for selecting  $V_i$  is the likelihood ratio test (LRT):

$$\text{LRT} = -2 \times \{\ell(\hat{\theta}_0) - \ell(\hat{\theta}_a)\} \sim \chi_p^2$$

where

- ▷  $\ell(\hat{\theta}_0)$  the value of the log-likelihood function under the null hypothesis, i.e., the special case model
  - ▷  $\ell(\hat{\theta}_1)$  the value of the log-likelihood function under the alternative hypothesis, i.e., the general model
  - ▷  $p$  denotes the number of parameters being tested
- 
- **Note:** Provided that the mean structure in the two models is the same, we can either compare the REML or ML likelihoods of the models (preferable is REML)

## 2.9 Hypothesis Testing (cont'd)

---

- **Example:** In the model we fitted for the AIDS dataset (see pp.54) we had assumed a compound symmetry covariance matrix – we would like to see if this option sufficiently describes the correlations and variances in the data
  - ▷ we will compare the compound symmetry model:

$$H_0 : V_i = \begin{bmatrix} t=0 & t=2 & t=6 & t=12 & t=18 \\ \sigma^2 & \tilde{\sigma} & \tilde{\sigma} & \tilde{\sigma} & \tilde{\sigma} \\ & \sigma^2 & \tilde{\sigma} & \tilde{\sigma} & \tilde{\sigma} \\ & & \sigma^2 & \tilde{\sigma} & \tilde{\sigma} \\ & & & \sigma^2 & \tilde{\sigma} \\ & & & & \sigma^2 \end{bmatrix}$$



## 2.9 Hypothesis Testing (cont'd)

---

▷ versus the unstructured model

$$H_a : V_i = \begin{bmatrix} t = 0 & t = 2 & t = 6 & t = 12 & t = 18 \\ \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ & & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ & & & \sigma_4^2 & \sigma_{45} \\ & & & & \sigma_5^2 \end{bmatrix}$$

## 2.9 Hypothesis Testing (cont'd)

---

- We can rewrite the two hypothesis as

$$H_0 : \begin{cases} \sigma_1^2 = \sigma_2^2 = \dots = \sigma_5^2 = \sigma^2 \\ \sigma_{12} = \sigma_{13} = \dots = \sigma_{45} = \tilde{\sigma} \end{cases}$$

$H_a$  : at least one variance or a covariance is not equal to the others

- The likelihood ratio test gives:

	df	logLik	LRT	p-value
Comp Symm	5.00	−3586.91		
General	18.00	−3547.72	78.39	<0.0001

## 2.9 Hypothesis Testing (cont'd)

---

- When we have **non-nested** models we **cannot** use standard tests anymore
- As an alternative for this case we use information criteria – the two standard ones are:

$$\text{AIC} = -2\ell(\hat{\theta}) + 2n_{par}$$

$$\text{BIC} = -2\ell(\hat{\theta}) + n_{par} \log(n)$$

where

- ▷  $\ell(\hat{\theta})$  is the value of the log-likelihood function
- ▷  $n_{par}$  the number of parameters in the model
- ▷  $n$  the number of subjects (independent units)

## 2.9 Hypothesis Testing (cont'd)

When we compare two **non-nested** models we choose the model that has the **lowest** AIC/BIC value

- **Example:** For the Prothrombin data we compare the exponential and Gaussian serial correlation structures – the models are:

$$\left\{ \begin{array}{l} M_1 : \text{pro}_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \{\text{predn}_i \times \text{Time}_{ij}\} + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, V_i^{\text{Exp}}) \\ M_2 : \text{pro}_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \{\text{predn}_i \times \text{Time}_{ij}\} + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, V_i^{\text{Gauss}}) \end{array} \right.$$

## 2.9 Hypothesis Testing (cont'd)

---

- The AIC and BIC values for the two models are:

	df	logLik	AIC	BIC
Exp	5.00	-13468.84	26947.67	26977.65
Gauss	5.00	-13750.88	27511.76	27541.73

- ▷ Both AIC and BIC suggest that the model with the exponential correlation structure is better

## 2.9 Hypothesis Testing (cont'd)

---

- The models we have assumed for the Prothrombin data assumed constant variance over time – as we have mentioned earlier (see pp. 91), this assumption is not often justified for longitudinal data
- We extend models  $M_1$  and  $M_2$  by assuming that the variances are an exponential function of time, i.e.,

$$\text{var}(\varepsilon_{ij}) = \sigma^2 \exp(\delta \text{Time}_{ij})$$

where

- ▷  $\delta$  is a parameter that controls how fast the variance changes over time
  - \* if  $\delta < 0$ , the variance decreases over time
  - \* if  $\delta = 0$ , the variance remains constant
  - \* if  $\delta > 0$ , the variance increases over time

## 2.9 Hypothesis Testing (cont'd)

- This means that models  $M_1$  and  $M_2$  are nested within their heteroscedastic cousins, i.e.,

$H_0 : \delta = 0$  homoscedastic model

$H_a : \delta \neq 0$  heteroscedastic model

- This implies that we can perform a likelihood ratio test

	df	logLik	AIC	BIC	LRT	p-value
Exp - homoscedastic	5.00	-13468.84	26947.67	26977.65		
Exp - heteroscedastic	6.00	-13459.99	26931.97	26967.94	17.70	<0.0001
Gauss - homoscedastic	5.00	-13750.88	27511.76	27541.73		
Gauss - heteroscedastic	6.00	-13748.10	27508.21	27544.18	17.70	0.0185

## 2.9 Hypothesis Testing (cont'd)

---

- Notes: Hypothesis testing for the covariance matrix  $V_i$ 
  - ▷ The unstructured covariance matrix is the most general matrix we can assume:
    - \* all other covariance matrices are a special case of the unstructured matrix
    - \* **but** realistically it can only be fitted when we have balanced data and relatively few time points
  - ▷ The AIC and BIC do not always select the same model – when they disagree
    - \* AIC typically selects the more elaborate model, whereas
    - \* BIC the more parsimonious model



## 2.9 Hypothesis Testing (cont'd)

---

- **Hypothesis testing for the regression coefficients  $\beta$** : We assume that first a suitable choice for the covariance matrix has been made
- In the majority of the cases we compare nested models, and hence standard tests can be used
- We distinguish between two cases
  - ▷ tests for individual coefficients
  - ▷ tests for groups of coefficients

## 2.9 Hypothesis Testing (cont'd)

---

- Tests for individual coefficients are based on the Wald-type statistic but assume the  $t$  distribution for calculating  $p$ -values
  - ▷ the set of hypotheses is:

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

- ▷ and we use the  $t$  test statistic

$$\frac{\hat{\beta}}{s.e.(\hat{\beta})} \sim t_{df}$$

where  $\hat{\beta}$  is the MLE,  $s.e.(\hat{\beta})$  is the standard error of the MLE, and  $df$  are specified according to the number of subjects and number of repeated measurements per subject

## 2.9 Hypothesis Testing (cont'd)

---

- Tests for groups of coefficients are based on the F-test
  - ▷ the set of hypotheses is:

$$H_0 : L\beta = 0$$

$$H_a : L\beta \neq 0$$

where  $L$  is the contrasts matrix

- ▷ the  $F$  test statistic is

$$\frac{\hat{\beta}^\top L^\top \left\{ L \left( \sum_{i=1}^n X_i^\top V_i^{-1} X_i \right)^{-1} L^\top \right\}^{-1} L \hat{\beta}}{\text{rank}(L)} \sim F_{df_1, df_2}$$

## 2.9 Hypothesis Testing (cont'd)

---

- Tests for groups of coefficients are based on the F-test
  - ▷ The numerator degrees of freedom are always equal to the rank of the contrast matrix  $L$
  - ▷ Denominator degrees of freedom need to be estimated from the data:
    - \* Containment method
    - \* Satterthwaite approximation
    - \* Kenward and Roger approximation

**There is no single method that provides satisfactory results in all settings – it matters more what you do in small samples**

## 2.9 Hypothesis Testing (cont'd)

---

- **Example:** We have fitted the following model to the PBC dataset:

$$\left\{ \begin{array}{l} \log(\text{serBilir}_{ij}) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \\ \quad \beta_4 \{ \text{D-penicil}_i \times \text{Time}_{ij} \} + \beta_5 \{ \text{Female}_i \times \text{Time}_{ij} \} + \varepsilon_{ij} \\ \varepsilon_i \sim \mathcal{N}(0, V_i) \quad \text{where } V_i \text{ has a continuous AR1 structure} \end{array} \right.$$

- We are interested in
  - ▷ the effect of Age, and
  - ▷ the overall effect of Sex

## 2.9 Hypothesis Testing (cont'd)

---

- For the effect of Age we set the hypotheses:

$$H_0 : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

- The output of the model gives: ...

## 2.9 Hypothesis Testing (cont'd)

---

	Value	Std.Err.	<i>t</i> -value	<i>p</i> -value
$\beta_0$	0.940	0.395	2.382	0.017
$\beta_1$	0.154	0.034	4.546	< 0.001
$\beta_2$	-0.281	0.218	-1.291	0.197
$\beta_3$	-0.002	0.006	-0.361	0.718
$\beta_4$	-0.014	0.020	-0.670	0.503
$\beta_5$	-0.064	0.034	-1.862	0.063

- Hence, a non-significant Age effect
  - ▷ the *t*-value in the output is the estimated coefficient divided by its standard error

## 2.9 Hypothesis Testing (cont'd)

---

- For the overall effect of Sex we set the hypotheses:

$$H_0 : \beta_2 = \beta_5 = 0$$

$$H_a : \text{either } \beta_2 \text{ or } \beta_5 \text{ are not equal to 0}$$

- We **cannot** obtain the  $p$ -value for this test directly from the output
- We have six parameters, the contrast matrix  $L$  is

$$L = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$



## 2.9 Hypothesis Testing (cont'd)

---

- We obtain

$F$ -value	$df_1$	$df_2$	$p$ -value
4.458	2	1939	0.0117

- Hence, a significant overall sex effect
- We could also test the same hypotheses using a likelihood ratio test
  - ▷ in this case we compare the models under the null and alternative hypothesis

## 2.9 Hypothesis Testing (cont'd)

---

- The two models are:

$$H_0 : \log(\text{serBilir}_{ij}) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_3 \text{Age}_i + \beta_4 \{ \text{D-penicil}_i \times \text{Time}_{ij} \} + \varepsilon_{ij}$$

$$H_a : \log(\text{serBilir}_{ij}) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \beta_4 \{ \text{D-penicil}_i \times \text{Time}_{ij} \} + \beta_5 \{ \text{Female}_i \times \text{Time}_{ij} \} + \varepsilon_{ij}$$

▷ for both models  $V_i$  has a continuous AR1 structure

- If we compare the two models we again end up in the same hypotheses:

$$H_0 : \beta_2 = \beta_5 = 0$$

$$H_a : \text{either } \beta_2 \text{ or } \beta_5 \text{ are not equal to 0}$$

## 2.9 Hypothesis Testing (cont'd)

---

- The likelihood ratio test gives

	df	logLik	AIC	BIC	LRT	p-value
without Sex	6.00	−1618.23	3248.46	3281.90		
with Sex	8.00	−1613.76	3243.52	3288.10	8.94	0.0114

- Hence, again the same conclusion, i.e., a significant overall sex effect

## 2.9 Hypothesis Testing (cont'd)

---

- Notes: Hypothesis testing for the regression coefficients  $\beta$ 
  - ▷ The likelihood ratio test, and the classical univariate and multivariate Wald tests (i.e., using the  $\chi^2$  distribution instead of the  $t$  or  $F$  distributions) are 'liberal'
    - \* they give smaller  $p$ -values than the ones they should give, especially in small samples
  - ▷ **Important:** The likelihood ratio test for comparing models with different  $X\beta$  parts is only valid when the models have been fitted using maximum likelihood and not REML (see also pp. 73–77)

## 2.10 Confidence Intervals

---

- Confidence intervals for model parameters are obtained from the approximate distribution of the maximum likelihood estimates (MLEs)

$$\hat{\beta} \sim \mathcal{N}(\beta^*, \text{var}(\hat{\beta}))$$

where

- ▷  $\hat{\beta}$  are the MLEs
- ▷  $\beta^*$  the true parameter values
- ▷  $\text{var}(\hat{\beta}) = \left( \sum_{i=1}^n X_i^\top V_i^{-1} X_i \right)^{-1}$  is the covariance matrix of the MLEs

## 2.10 Confidence Intervals (cont'd)

---

- For example, for the  $k$ -th regression coefficient  $\beta_k$ , the 95% Wald-based CI is

$$\hat{\beta}_k \pm 1.96 \times \text{s.e.}(\hat{\beta}_k)$$

- To obtain confidence intervals for the whole mean evolution we need to multiply with a corresponding design matrix  $X$  (see pp. 45–46), i.e.,

$$X\hat{\beta} \pm 1.96 \times \sqrt{\text{diag}\{X\text{var}(\hat{\beta})X^\top\}}$$

- ▷ this type of confidence intervals have been used in the effect plots we have seen earlier (see pp. 68–71)

## 2.11 Design Considerations - Sample Size

---

- Two interrelated questions relevant to hypothesis testing are how to perform **power** & **sample size** calculations
  - ▷ **power:** is the probability that we will find a statistically significant difference between the two groups, given that this difference truly exists
  - ▷ **sample size:** in the design phase of a study, and for a given a priori postulated setting, we often want to find how many subjects we need to enrol to detect the difference of interest, with a prespecified level of power (and a prespecified significance level)

## 2.11 Design Considerations - Sample Size (cont'd)

---

- In the literature several formulas for sample size calculations have been developed for marginal and linear mixed models (see Chapter 3)
- **However**, in the majority of the cases these formulas are only applicable in simple settings, and **cannot** account for common features of longitudinal data, e.g.,
  - ▷ complex correlation structures
  - ▷ unbalanced data
  - ▷ missing data (see Chapter 6)



## 2.11 Design Considerations - Sample Size (cont'd)

---

- The only viable and trustworthy approach is to use simulation  
This entails the following generic steps

S1: Simulate longitudinal responses under the postulated model, and a specific sample size  $n$

\* in this step the covariates could be set fixed or also simulated

S2: Fit the postulated model in the simulated data

S3: Perform the hypothesis test of interest and retain the  $p$ -value

## 2.11 Design Considerations - Sample Size (cont'd)

---

- Repeat Steps 1–3  $M$  times (e.g.,  $M = 500$  or  $M = 1000$ ), and calculate how many times the  $p$ -value was significant at significance level  $\alpha$  (e.g.,  $\alpha = 0.05$ )
  - ▷ the percentage of times the test was significant is the estimated power for the specific setting under consideration

## 2.11 Design Considerations - Sample Size (cont'd)

---

- Notes: On power calculation for repeated measurement models
  - ▷ To perform a sample size calculation we just repeat the above simulation procedure with increasing  $n$  until the power reaches the prespecified level
  - ▷ The simulation approach allows very easily to investigate how power is affected by specific changes in the design, e.g.,
    - \* increasing the number of repeated measurements per subject  $n_i$  versus increasing the number of subjects  $n$
    - \* different percentages of missing data
    - \* ...
  - ▷ The downside is that each time a new syntax needs to be written to do these calculations

## 2.12 Residuals

---

**All statistical models are based on assumptions**

- Hence, to extract meaningful conclusions we need to check whether these assumptions are (crudely) violated

## 2.12 Residuals (cont'd)

---

- The marginal model for multivariate continuous data makes analogous assumptions to the linear regression model

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, V_i)$$

namely

- ▷ the error terms  $\varepsilon_i$  follow the normal distribution  $\mathcal{N}(0, V_i)$
- ▷ the error terms are independent from the covariates  $X$
- ▷ the covariates act linearly on the average outcome

## 2.12 Residuals (cont'd)

---

- To validate these assumptions we need an estimate of the error terms  $\varepsilon_{ij}$
- Based on the fitted model we obtain the estimate

$$r_{ij} = y_{ij} - x_{ij}^{\top} \hat{\beta}$$

- ▷  $\hat{\beta}$  are the (restricted) maximum likelihood estimates
- ▷ the  $r_{ij}$  are called *residuals*

**When the model is correctly specified**, we expect these residuals to have a  $\mathcal{N}(0, V_i)$  distribution

## 2.12 Residuals (cont'd)

---

- Hence, we expect these residuals to be correlated and possibly also heteroscedastic
  - ▷ 'heteroscedastic' means that they exhibit non-constant variance
- This feature complicates matters because it is not easy to assess if the residuals exhibit the assumed properties
- To overcome this problem we need to transform  $r_{ij}$  to a scale that has easier to check properties
  - ▷ for example, in general, it is easier to assess whether a particular variable has a standard normal distribution

## 2.12 Residuals (cont'd)

---

- To achieve this we multiply the residual with the inverse Choleski factor

$$r_i^{norm} = \hat{H}_i^{-1} r_i = \hat{H}_i^{-1} (y_i - X_i \hat{\beta})$$

where

- ▷  $\hat{H}_i$  is an upper-triangular matrix with the property  $\hat{H}_i^\top \hat{H}_i = \hat{V}_i$ , with  $\hat{V}_i$  denoting the estimated covariance matrix
- ▷  $r_{ij}^{norm}$  are called *normalized residuals* and when the covariance matrix is correctly specified, they should be approximately distributed as  $\mathcal{N}(0, 1)$  random variables



## 2.12 Residuals (cont'd)

---

- When we have assumed a homoscedastic covariance matrix (i.e., variance remains constant), another transformation that it is often used is

$$r_i^{Pears} = \hat{\sigma}^{-1} r_i = \sigma^{-1}(y_i - X_i \hat{\beta})$$

where

- ▷  $\hat{\sigma}$  denotes the estimated standard deviation of the error term, i.e.,  $V_i$  has the structure  $\sigma^2 R_i$ , with  $R_i$  denoting a correlation matrix
- ▷  $r_{ij}^{Pears}$  are called *Pearson residuals* and when the covariance matrix is correctly specified, they should be approximately distributed as  $\mathcal{N}(0, R_i)$  random variables

## 2.12 Residuals (cont'd)

---

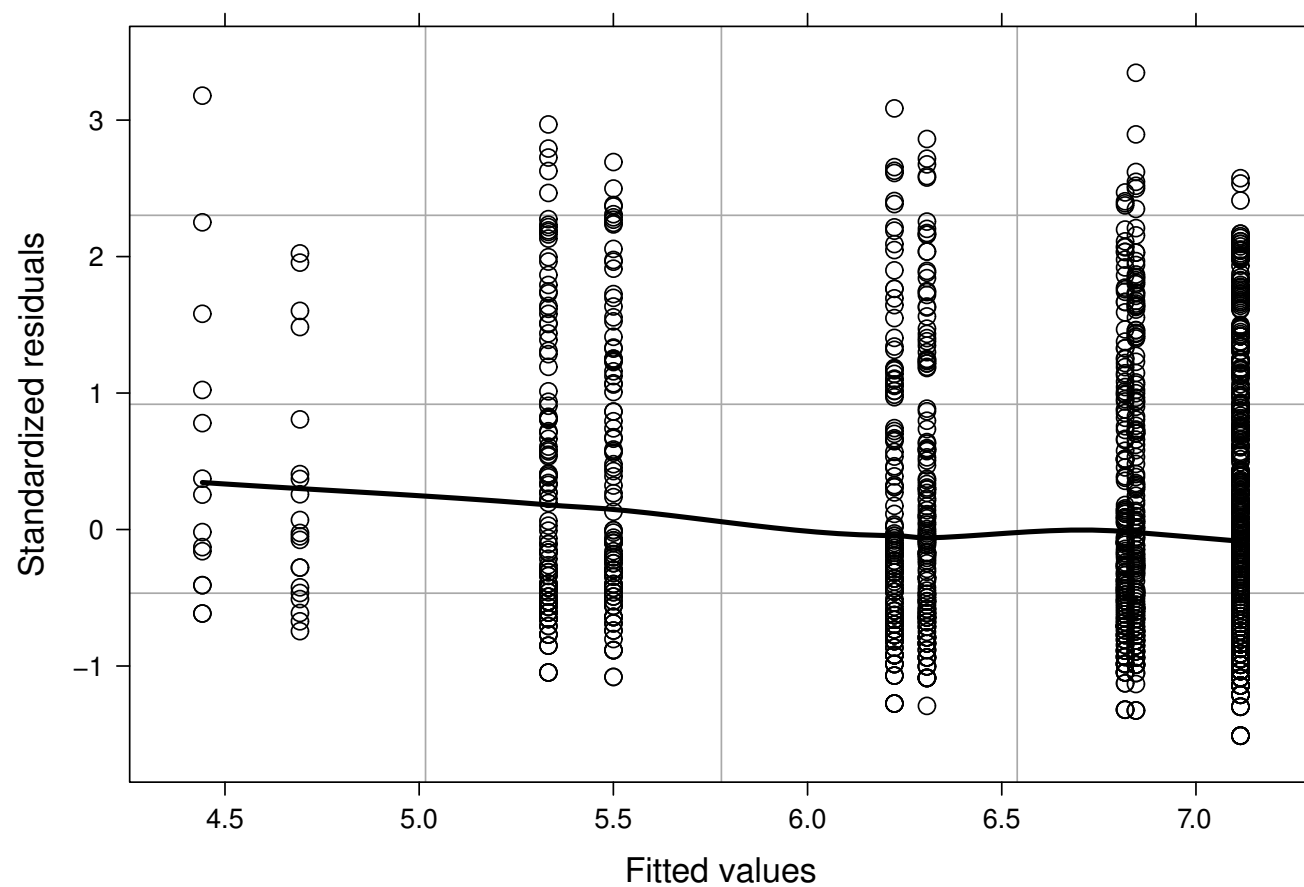
- **Example:** We evaluate the assumptions behind the following model fitted to the AIDS dataset:

$$\begin{cases} \sqrt{\text{CD4}_{ij}} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \{\text{ddI}_i \times \text{Time}_{ij}\} + \varepsilon_{ij}, \\ \varepsilon_i \sim \mathcal{N}(0, V_i), \quad V_i \text{ is unstructured} \end{cases}$$

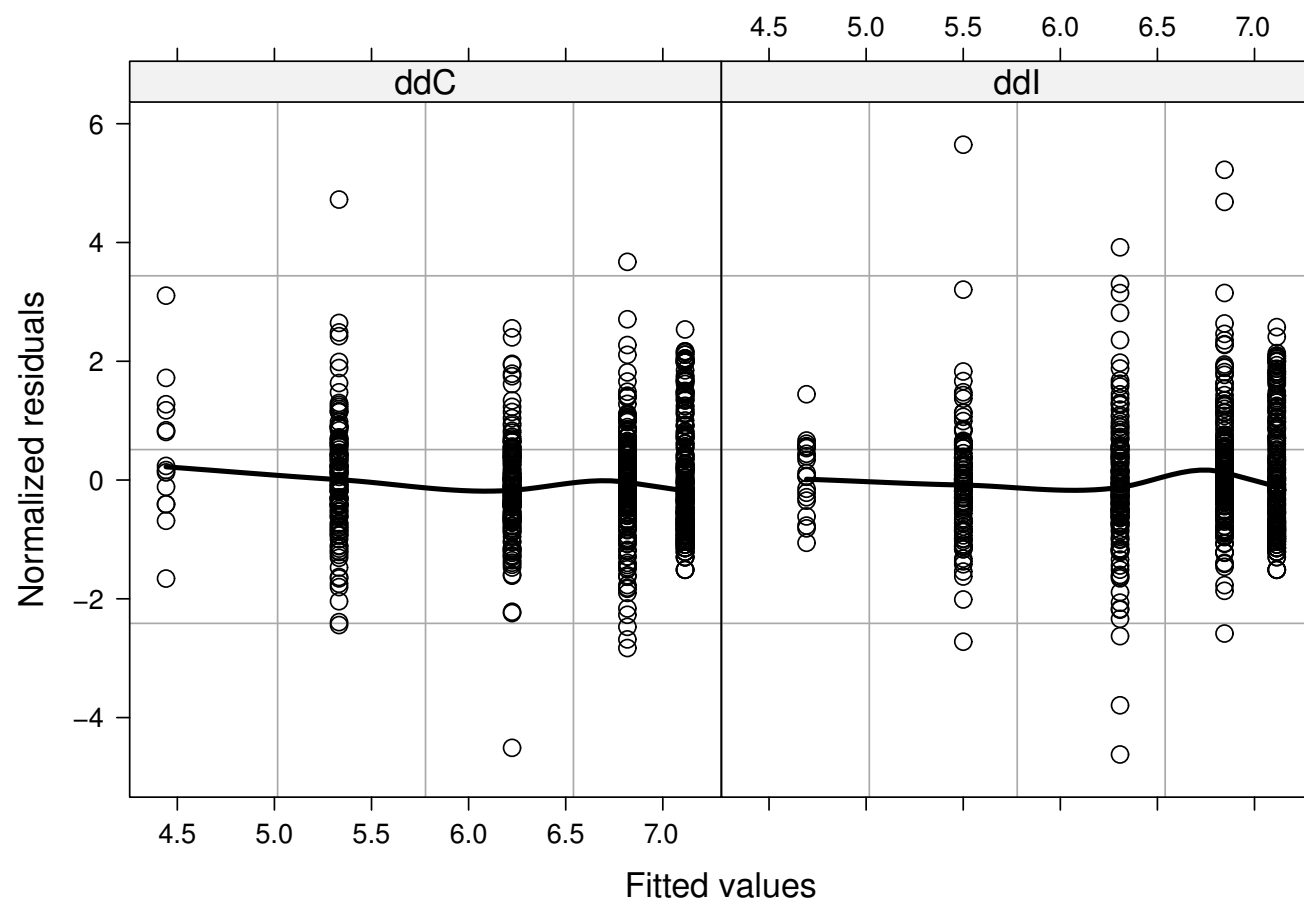
by plotting

- ▷ the standardized residuals versus fitted values
- ▷ the normalized residuals versus fitted values per treatment group
- ▷ QQ-plot of the standardized residuals

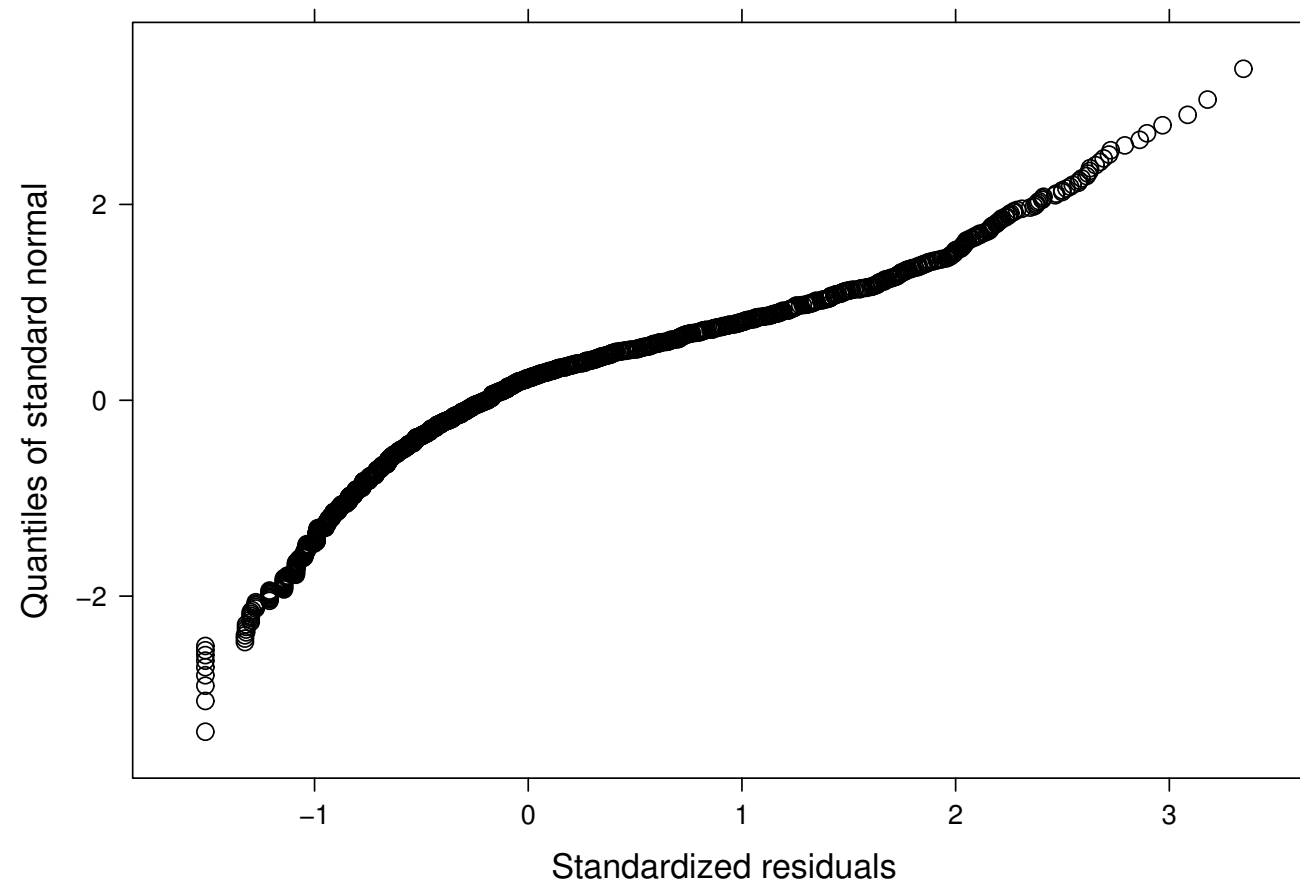
## 2.12 Residuals (cont'd)



## 2.12 Residuals (cont'd)



## 2.12 Residuals (cont'd)



## 2.12 Residuals (cont'd)

---

- Observations

- ▷ the plots of the residuals versus the fitted values do show a slightly systematic behavior with more positive residuals in the range of low fitted values
- ▷ the QQ-plot is not perfect, but does not show a big discrepancy from normality

## 2.12 Residuals (cont'd)

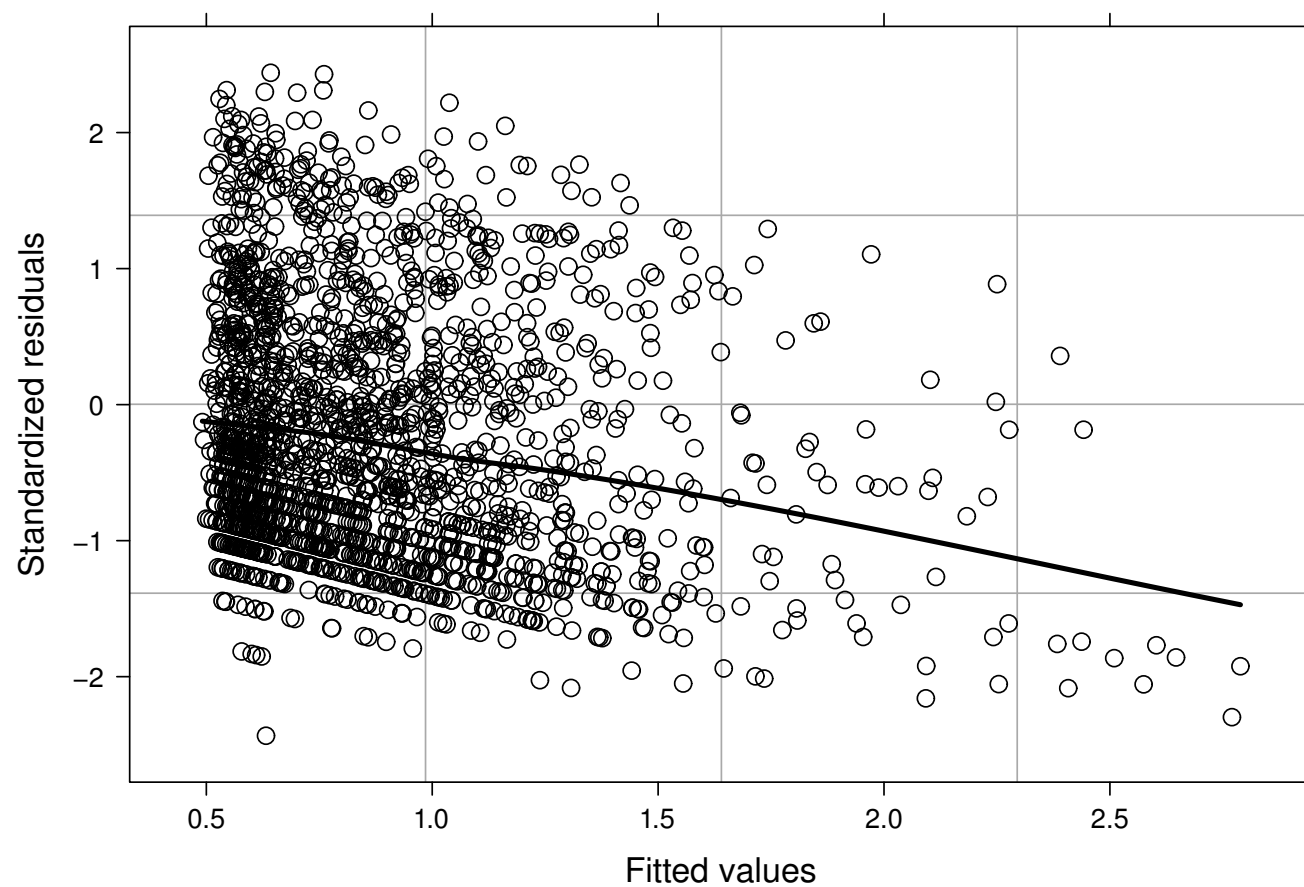
- **Example:** We continue by evaluating the assumptions of the model we have fitted to the PBC dataset:

$$\left\{ \begin{array}{l} \log(\text{serBilir}_{ij}) = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \\ \quad \beta_4 \{ \text{D-penicil}_i \times \text{Time}_{ij} \} + \beta_5 \{ \text{Female}_i \times \text{Time}_{ij} \} + \varepsilon_{ij} \\ \varepsilon_i \sim \mathcal{N}(0, V_i) \quad V_i \text{ has a continuous AR1 structure} \end{array} \right.$$

by plotting again

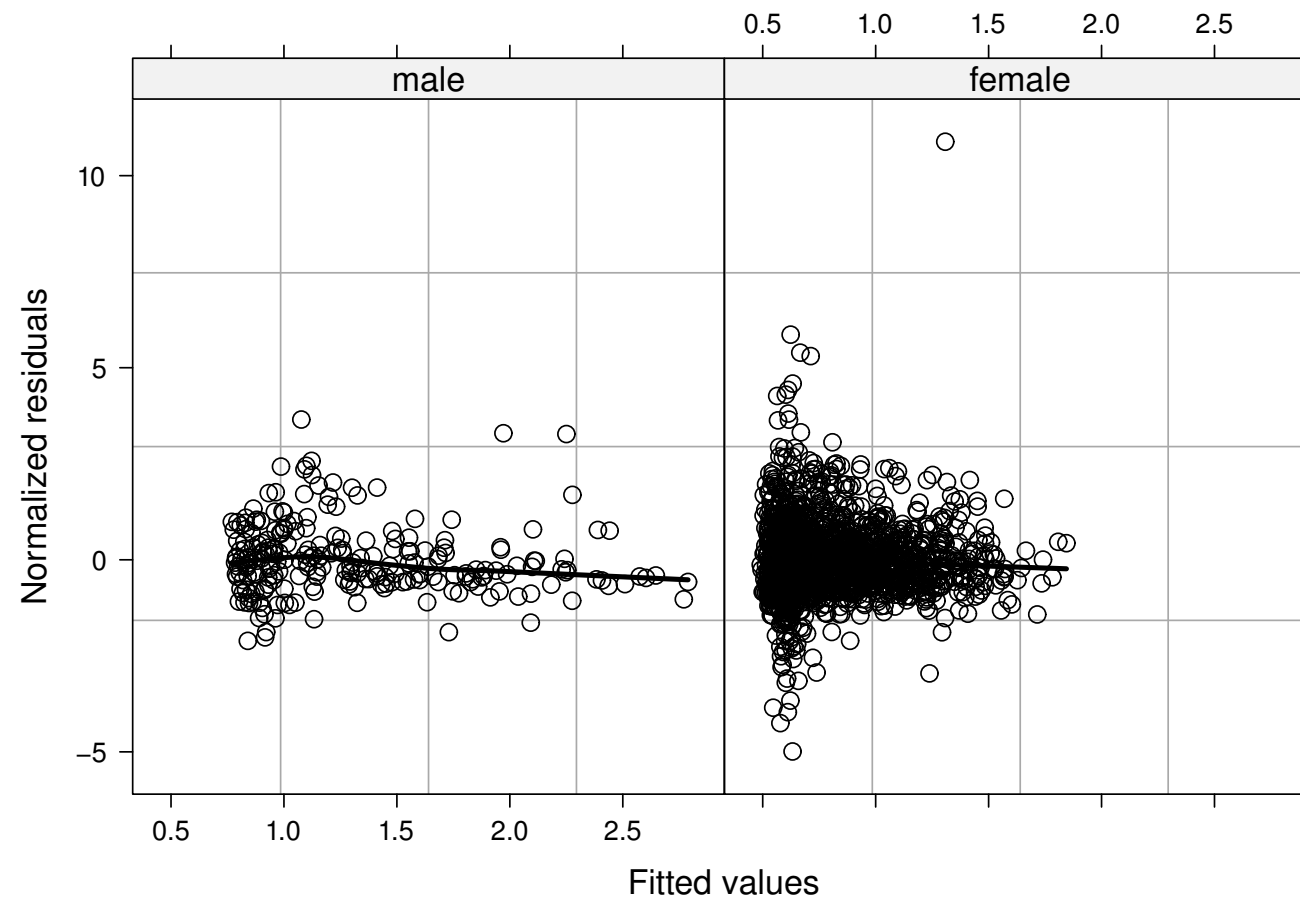
- ▷ the standardized residuals versus fitted values
- ▷ the normalized residuals versus fitted values per gender
- ▷ QQ-plot of the standardized residuals

## 2.12 Residuals (cont'd)

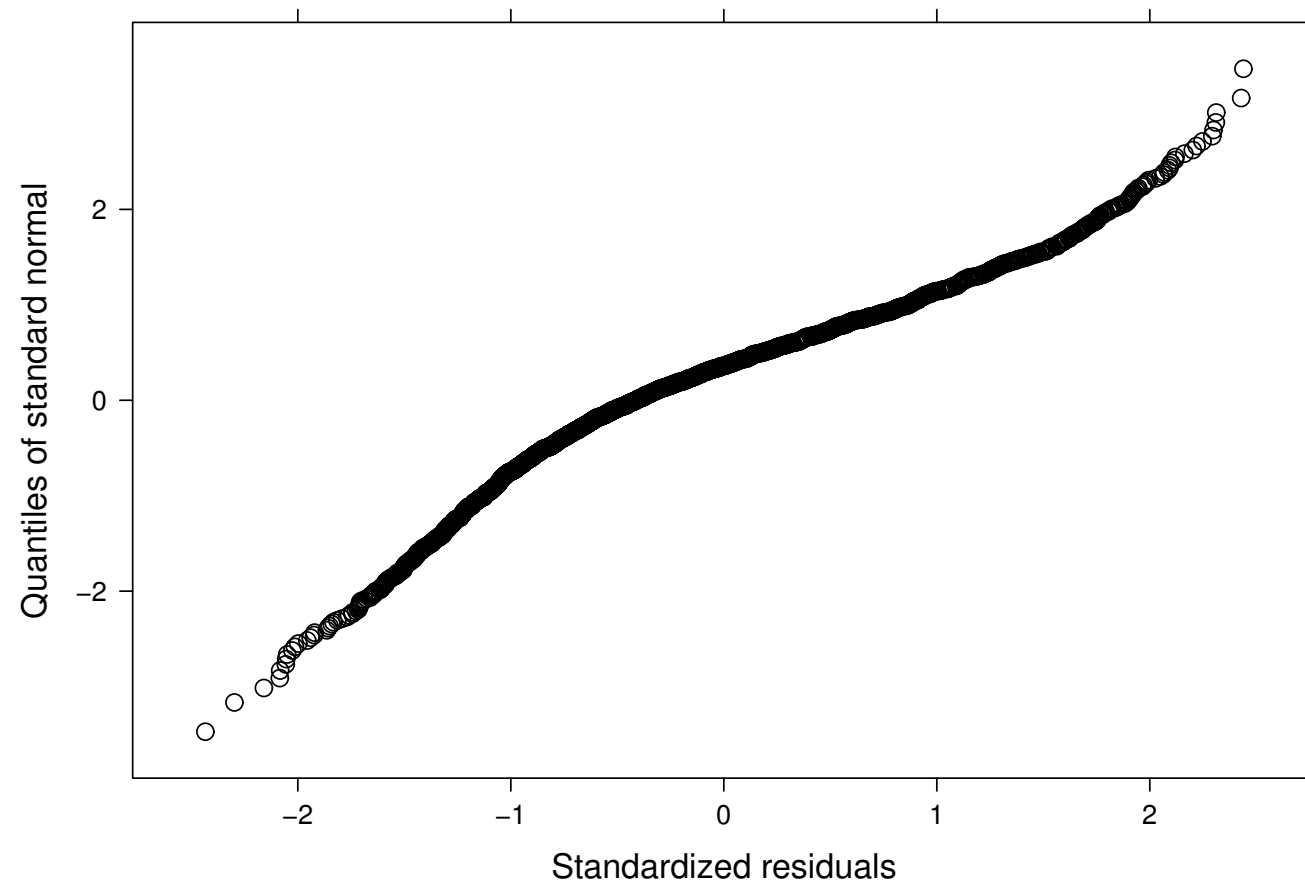




## 2.12 Residuals (cont'd)



## 2.12 Residuals (cont'd)



## 2.12 Residuals (cont'd)

---

- Observations

- ▷ the plot of the standardized residuals versus fitted values shows a clear systematic trend with more negative residuals in the range of high fitted values
- ▷ the plot of normalized residuals versus fitted values shows an outlying observation for female and some slight heteroscedasticity (higher spread of residuals for low fitted values than for high)
- ▷ the QQ-plot suggests a good fit of the normal distribution

## 2.13 Review of Key Points

---

- Methods for analyzing grouped/correlated data
  - ▷ naive approaches working on parts or summaries of the data  $\Rightarrow$  loss of information
  - ▷ marginal models  $\Rightarrow$  extension of simple linear regression to the context of correlated data
- Marginal models: Features
  - ▷ error terms are assumed correlated  $\Rightarrow$  we need to make an appropriate assumption
  - ▷ mean structure is build as in standard regression models – however, need to account for potential nonlinear effects of time and/or interaction terms
  - ▷ model building: we start from a ‘fully’ specified mean structure, we select an appropriate covariance structure, and then the return to make inference for the mean

## 2.13 Review of Key Points (cont'd)

---

- Hypothesis testing
  - ▷ for the covariance structure and for nested models likelihood ratio tests are most often used, for non-nested models AIC/BIC
  - ▷ for the mean structure  $t$  and  $F$  tests with appropriate degrees of freedom
  
- Residuals
  - ▷ standard residuals plots are used to check the model assumptions
  - ▷ standardized and normalized residuals