Statistical Analysis of Repeated Measurements Data

Dimitris Rizopoulos

Department of Biostatistics, Erasmus University Medical Center

✉ d.rizopoulos@erasmusmc.nl

🐦 @drizopoulos

Contents

3 The Linear Mixed Effects Model 145

4   Marginal Models for Discrete Data   254

5 Mixed Models for Discrete Data   308

6 Statistical Analysis with Incomplete Grouped Data   370

7 Closing   417

Practicals   423

What is this Course About

Grouped data arise in a wide range of disciplines

- Typical examples of grouped data

  ▷ *repeated measurements:* measuring the same outcome multiple times on the same sample unit (e.g., biomarkers in patients)

  ▷ *multilevel data:* outcomes measured on sample units that are organized in different levels (e.g., patients in medical centers or students in schools)

What is this Course About (cont'd)

- Statistical analysis of clustered/grouped data

  ▷ Features of grouped data

  ▷ describe their distribution

  ▷ inference using suitable regression models

Lexical convention

- The following terms are used interchangeably to denote multivariate outcomes

  ▷ clustered data

  ▷ repeated measurements data

  ▷ multilevel data

  ▷ grouped data

Learning Objectives

- <u>Goals:</u> After this course participants will be able to

  ▷ identify settings in which a repeated measurements model is required,

  ▷ construct and fit an appropriate model to the data at hand, and

  ▷ correctly interpret the results

- Even though the course will be primarily explanatory

  ▷ sufficient mathematical detail will be provided in order participants to obtain a clear view on the different modeling approaches, and how they should be used in practice

- **Chapter 1:** Motivating Data Sets

  ▷ Data sets that we will use throughout the course

  ▷ Formulation of possible research questions

  ▷ Features of repeated measurements data

- **Chapter 2:** Marginal Models for Continuous Data

  ▷ Naive approaches

  ▷ Review linear regression

  ▷ Marginal models

Agenda (cont'd)

- **Chapter 3:** The Linear Mixed Effects Model

  ▷ Mixed effects models

  ▷ Mixed models with correlated errors

  ▷ Nested and cross random effects

  ▷ Time-varying covariates

- **Chapter 4:** Marginal Models for Discrete Data

  ▷ Review generalized linear models

  ▷ Generalized estimating equations

Agenda (cont'd)

- **Chapter 5:** Mixed Models for Discrete Data

  ▷ Generalized linear mixed effects models

  ▷ interpretation of parameters

  ▷ approximations of the integrand & integral

- **Chapter 6:** Statistical Analysis with Incomplete Grouped Data

  ▷ Problems with incomplete data

  ▷ Missing data mechanisms

  ▷ Valid inferential approaches

Structure of the Course & Material

- Lectures & software practicals using R

- Material:

  ▷ Course Notes

  ▷ R code in soft format

- Within the course notes there are several examples of R syntax – these are denoted by the symbol 'R> '

Software Requirements

- The up-to-date versions of R and Rstudio; downloadable from

  ▷ https://cran.r-project.org/

  ▷ https://www.rstudio.com/

- Additional required packages

  ▷ **nlme**, **lme4**, **GLMMadaptive**, **geepack**,

  ▷ **MASS**, **lattice**, **shiny**, **corrplot**

Software Requirements

- Up-to-date versions of these packages and their dependencies can be installed using the command

```
install.packages(c("shiny", "nlme", "lattice", "lme4",
                    "GLMMadaptive", "geepack", "MASS", "corrplot"),
                  dependencies = TRUE)
```

- Up-to-date version of a modern web browser, e.g.,

  ▷ Mozilla Firefox (https://www.mozilla.org/firefox/)

  ▷ Google Chrome (https://www.google.com/chrome/)

Software Requirements

- We will use a `shiny` web app that replicates all analyses in the course including also some additional illustrations

- The app is available on GitHub and can be invoked using the following two-step procedure (assuming internet connection is available and you have installed the aforementioned packages)

  1. Start R
  2. Run the command

     ```
     shiny::runGitHub("Repeated_Measurements", "drizopoulos")
     ```

  this will open a new web browser window (or tab) with the app

- <u>Note:</u> in order the app to be functional you should **not** close R

References

- Some texts in longitudinal data analysis

  ▷ Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: John Wiley & Sons.

  ▷ Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

  ▷ Galecki, A. and Burzykowski, T. (2013). *Linear Mixed-Effects Models Using R*. New York: Springer-Verlag.

  ▷ Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

  ▷ Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*, 2nd Ed. Hoboken: John Wiley & Sons.

  ▷ Hand, D. and Crowder, M. (1995). *Practical Longitudinal Data Analysis*. London: Chapman & Hall.

References (cont'd)

- Some texts in longitudinal data analysis

  ▷ Hedeker, D. and Gibbons, R. (2006). Longitudinal Data Analysis. New York: John Wiley & Sons.

  ▷ Lindsey, J. (1993). *Models for Repeated Measurements*. Oxford: Oxford University Press.

  ▷ Pinheiro, J. and Bates, D. (2000). *Mixed Effects Models in S and S-plus*. New York: Springer-Verlag.

  ▷ Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Exams

- Written exams

- Scope: Convincingly demonstrate that you have understood the basics of the statistical analysis of grouped data, and you can apply it on your own

- Format

  ▷ You will split in groups of 3-4 persons

  ▷ On the last day of the course you will receive a data set and specific questions

  ▷ On the exams deadline (see CANVAS) you will need to submit your report on CANVAS – **reports submitted after the deadline will NOT be accepted**

Disclaimer & Warning!

As you will see, the analysis of repeated measurements data is rather complex, and the statistical regression models we have available for these data **cannot** be introduced without the use of mathematical equations

I will try to explain all material simply and intuitively, nevertheless, a week of equations follows . . .

Use of Statistical Models

> *. . . the megalomaniacal strategy of fitting a grand unified model, supposedly capable of answering any conceivable question that might be posed, is, in our view, dangerous, unnecessary and counterproductive.*
>
> Drum and McCullach (1993, *Statistical Science* **8**, 300–301)

Chapter 1

Motivating Data Sets

1.1 Motivating Longitudinal Studies

- **AIDS:** 467 HIV infected patients who had failed or were intolerant to zidovudine therapy (AZT) (Abrams et al., NEJM, 1994)

- The aim of this study was to compare the efficacy and safety of two alternative antiretroviral drugs, didanosine (ddI) and zalcitabine (ddC)

- Outcomes of interest:

  ▷ CD4 cell count measurements at baseline, 2, 6, 12 and 18 months

  ▷ randomized treatment: 230 patients ddI and 237 ddC

  ▷ prevOI: previous opportunistic infections

- Research Questions:

    ▷ How CD4 cell count evolves over time for this cohort of patients?

    ▷ Does treatment improve average longitudinal evolutions?

- **PBC:** Primary Biliary Cirrhosis:

    ▷ a chronic, fatal but rare liver disease

    ▷ characterized by inflammatory destruction of the small bile ducts within the liver

- Data collected by Mayo Clinic from 1974 to 1984 (Murtaugh et al., Hepatology, 1994)

- Outcomes of interest:

    ▷ longitudinal serum bilirubin, serum cholesterol, prothrombin time

    ▷ randomized treatment: 158 patients received D-penicillamine and 154 placebo

- Research Questions:

  ▷ Do men have higher serum bilirubin during follow-up than women?

  ▷ Is there a difference in the average longitudinal evolutions of serum bilirubin between the two treatments when we correct for age and gender at baseline?

- **Prothro:** Prednisone versus placebo in liver cirrhosis patients

  ▷ slowly progressing disease in which healthy liver tissue is replaced with scar tissue, eventually preventing the liver from functioning properly

- Randomized trial in Denmark (Andersen et al., Springer, 1993)

- Outcomes of interest:

  ▷ randomized treatment: 237 patients received prednisone and 251 placebo

  ▷ longitudinal prothrombin times

- **Glaucoma:** A group of eye conditions resulting in optic nerve damage, which may cause loss of vision

- Ongoing prospective cohort study on 139 patients (80% men) conducted by the Rotterdam Eye Hospital in the Netherlands http://rod-rep.com

- Outcome of interest:

  ▷ Visual field (VF) sensitivity collected at approximately 6-months intervals
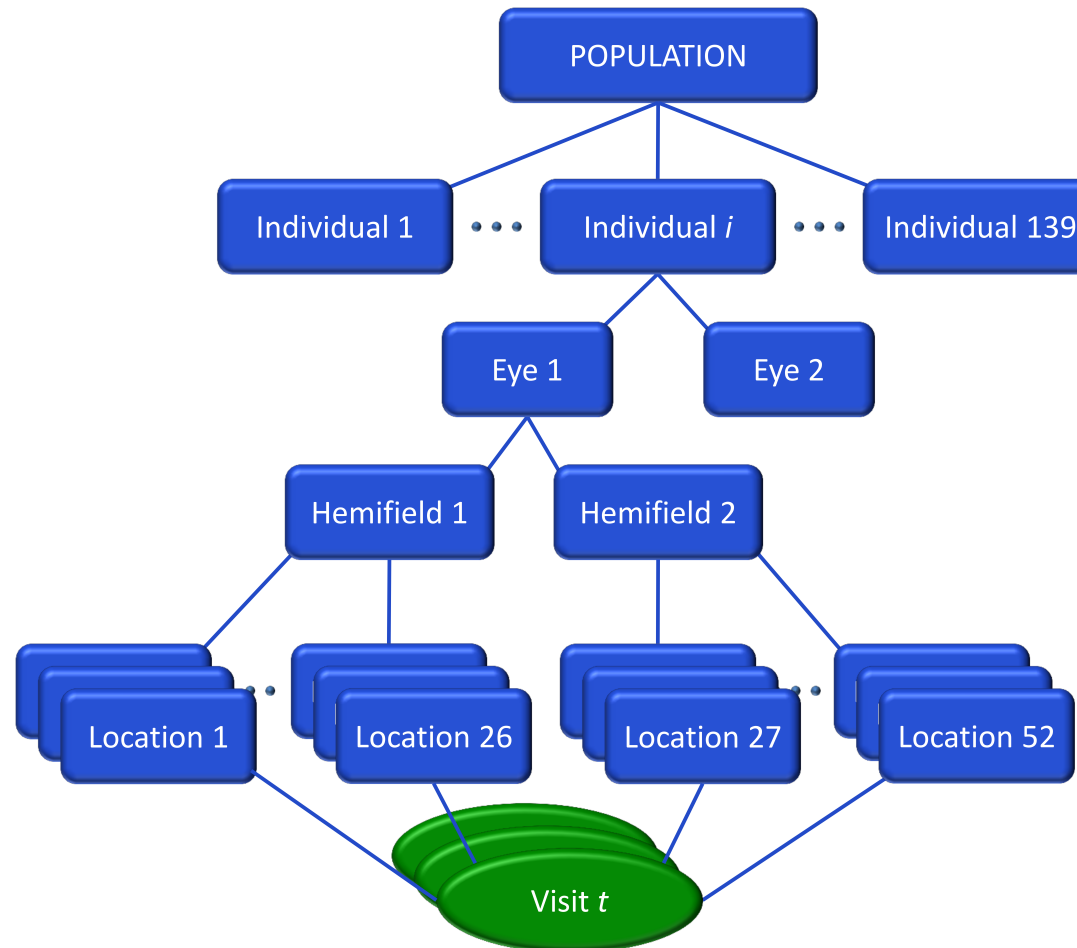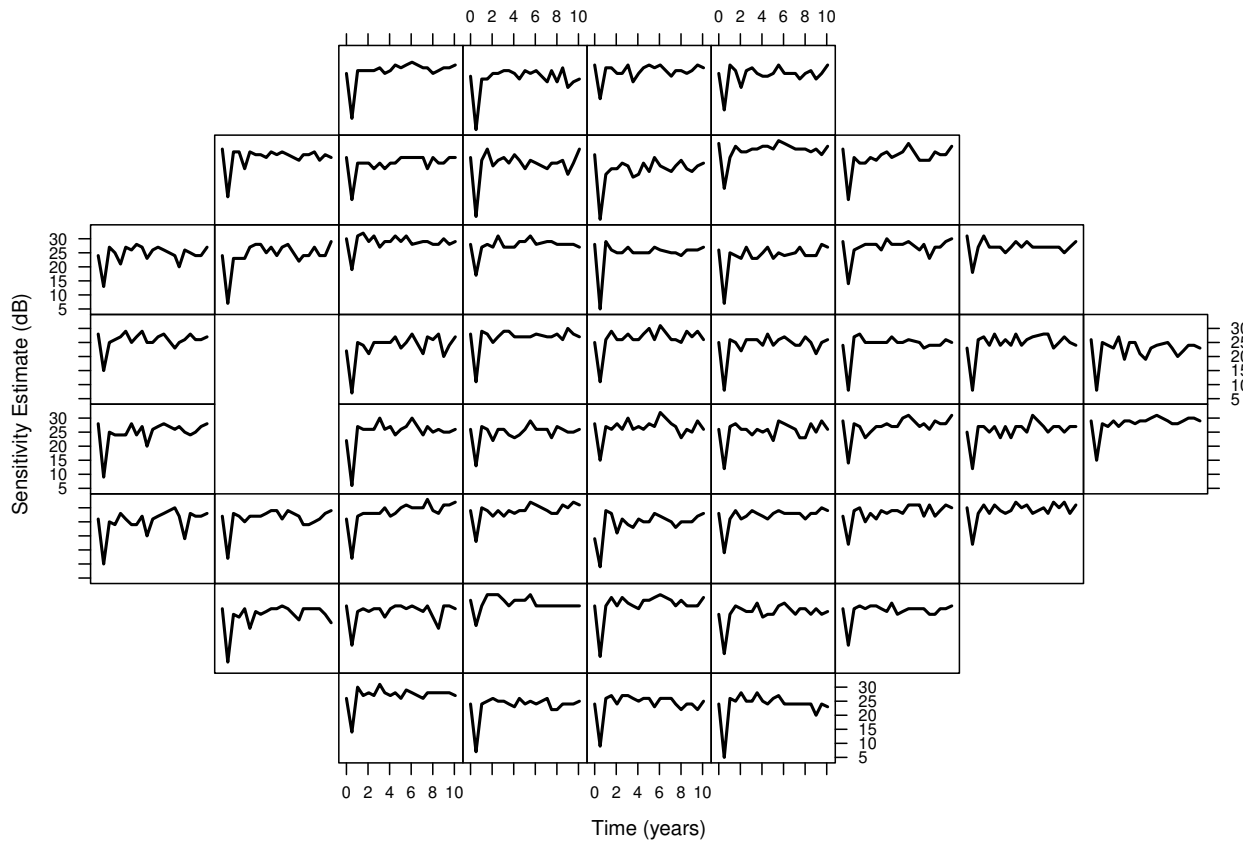
- Research Questions:

  ▷ Study disease progression using VF sensitivity

  ▷ Predict rate of progression for future patients

- Repeated evaluations of the same outcome in each subject over time

  ▷ CD4 cell count in HIV-infected patients

  ▷ serum bilirubin in PBC patients

- Visiting process

  ▷ some times fixed by design (e.g., in randomized trials) but often not everybody adheres to them

  ▷ completely determined by the physicians and/or the patients

> **Measurements on the same subject are expected to be (positively) correlated**

- This implies that standard statistical tools, such as the t-test and simple linear regression that assume independent observations, are not optimal for longitudinal data analysis

1.2 Features of Longitudinal Data (cont'd)

- Let's see why: The simplest case of longitudinal data are paired data

- Example: We consider the baseline and 6-month longitudinal measurements of square root CD4 cell count from the AIDS dataset

	n	mean	sd
$month = 0$	294	7.73	4.69
$month = 6$	294	6.71	4.96

- There is an average decrease of about 1 unit

- The classical analysis of paired data is based on comparisons within subjects:

$$\Delta_i \;=\; Y_i(t = 0) - Y_i(t = 6), \qquad i = 1, \ldots, n$$

- A positive Δ_i corresponds to a decrease of the square root CD4 cell count, while a negative Δ_i is equivalent to an increase

- Testing for a time effect is now equivalent to testing whether the average difference μ_Δ equals zero

- The paired t-test yields

```
        Paired t-test

data:   CD4 by obstime t = 6.472, df = 293, p-value = 4.057e-10
alternative hypothesis: true difference in means is not equal to 0 95 percent
confidence interval:
 0.7105585 1.3315439
sample estimates: mean of the differences
                1.021051
```

- What if we had ignored the paired nature of the data?

- We then could have used a two-sample (unpaired) t-test to compare the average CD cell count at the two time points

```
        Welch Two Sample t-test

data:  CD4 by obstime t = 2.565, df = 584.229, p-value = 0.01056
alternative hypothesis: true difference in means is not equal to 0 95 percent
confidence interval:
 0.2392406 1.8028617
sample estimates: mean in group 0 mean in group 6
        7.730128          6.709077
```

1.2 Features of Longitudinal Data (cont'd)

- We would still have found a significant difference (p = 0.0106), but the p-value would have been several orders of the magnitude larger than the one obtained from the paired t-test

- The two-sample t-test does not take into account that the measurements are not independent

  ▷ p-values **wrongly** too small for *between subjects* effects

  ▷ p-values **wrongly** too large for *within subjects* effects

- The different effects

  ▷ *between subjects:* examine differences between subjects (e.g., males vs females)

  ▷ *within subjects:* examine how much subjects tend to change over time

- This illustrates that classical statistical models which assume independent observations will not be optimal for the analysis of clustered data

- Longitudinal studies allow to investigate

    1. how treatment means differ at specific time points, e.g., at the end of the study (*cross-sectional effect*)

    2. how treatment means or differences between means of treatments change over time (*longitudinal effect*)

- An example: Suppose it is of interest to study the relation between some response Y and age

    ▷ a cross-sectional study yields the following data:
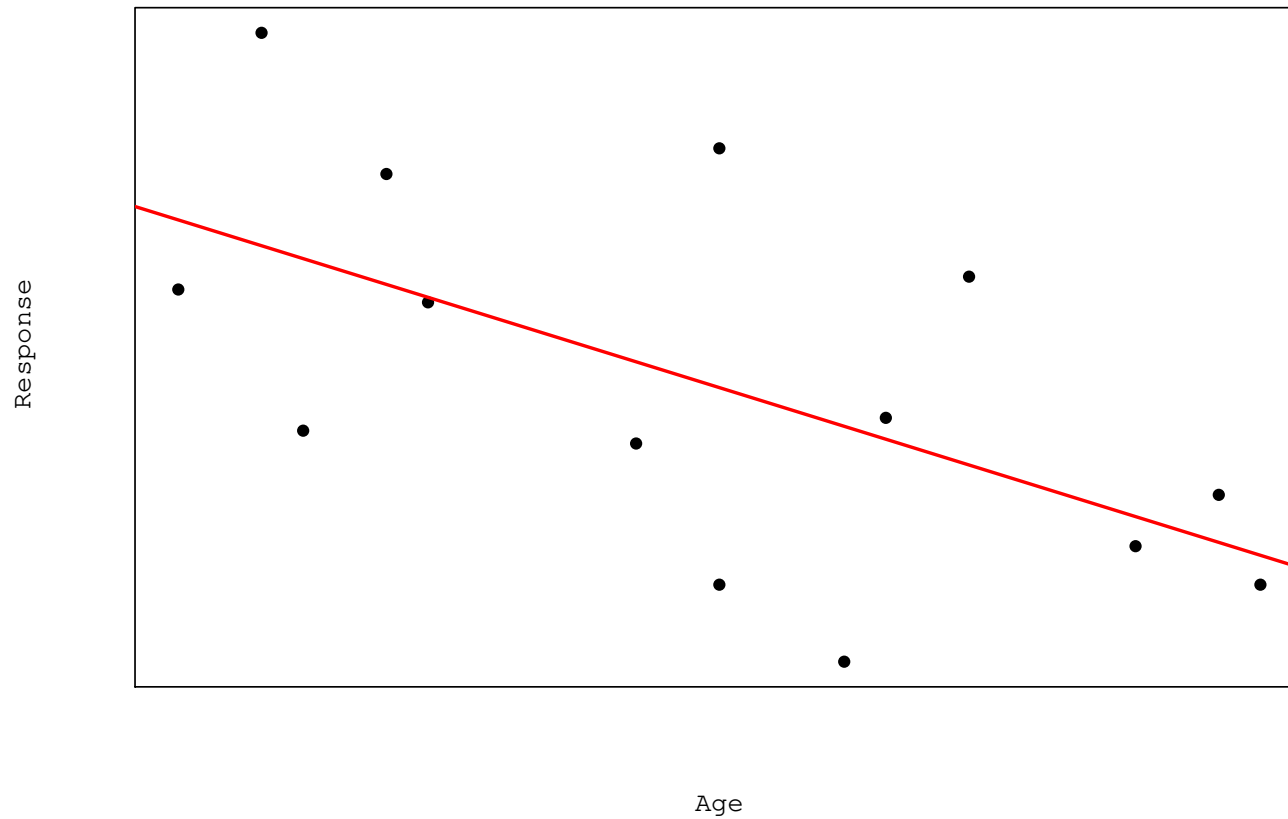
- The graph clearly suggests a negative relation between Y and age

- **Nevertheless**, exactly the same observations also could have been obtained in a longitudinal study, with 2 measurements per subject
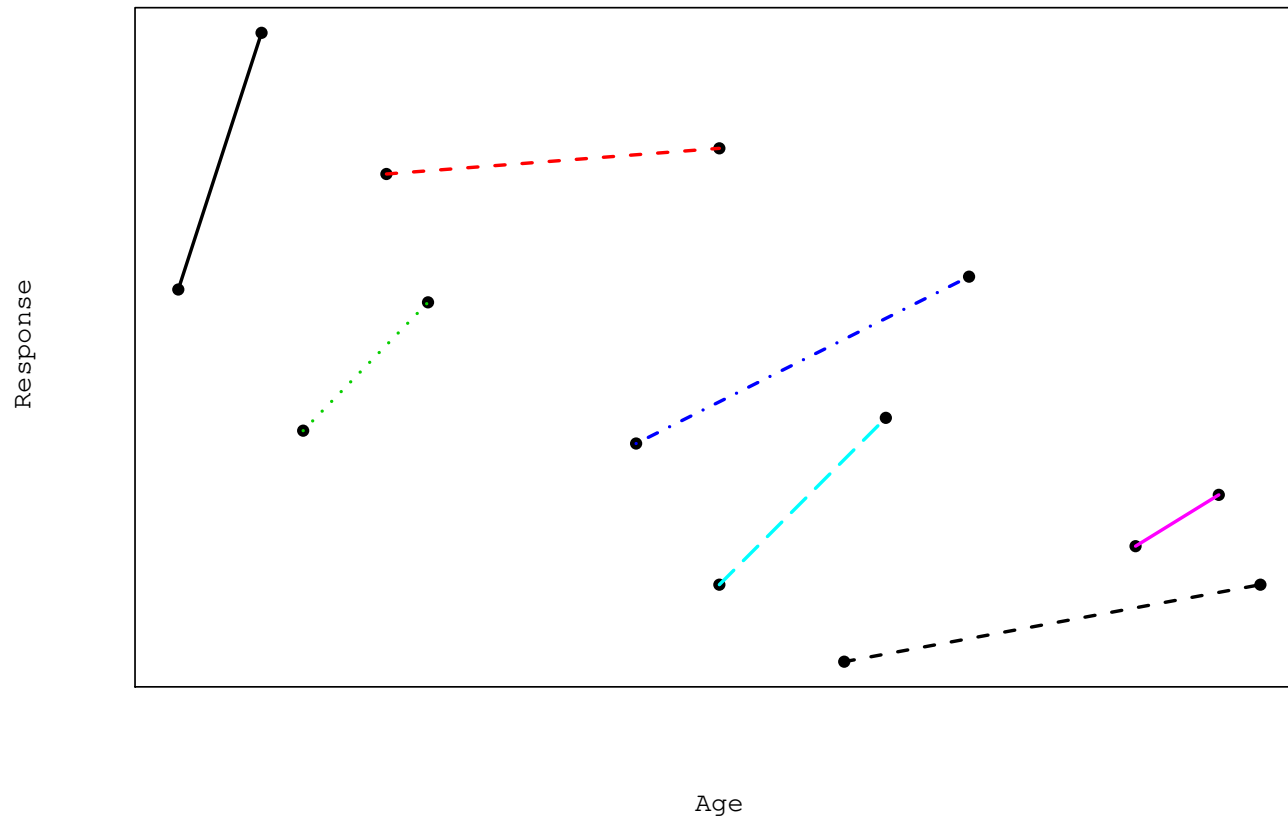
Are we now still inclined to conclude that there is a negative relation between Y and age?

- <u>Conclusion:</u> Longitudinal data allow to distinguish differences between subjects from changes within subjects

1.3 Review of Key Points

- Grouped & longitudinal data: Features

  ▷ measurements on the same subject are correlated

  ▷ allow to distinguish within and between subjects effects

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

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

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

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

  Afterwards, these AUCs are analyzed

  ▷ **Advantages:**
    * no problems of multiple testing
    * does not explicitly assume balanced data
    * compares 'overall' differences

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

  ▷ **Disadvantages:**

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

- **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

- **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

• <u>However</u>, all these methods have the disadvantage that (lots of) information is lost

<div style="border:1px solid black; padding:1em; text-align:center; color:red; font-weight:bold;">

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

</div>

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?

$$\Downarrow$$

Linear Regression Model

- Definition of the linear regression model

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

where

▷ y_i denotes the outcome for subject i

▷ $x_{i1}, \ldots, x_{ip}$ denote the p covariates for subject i

▷ $\beta_0, \beta_1, \ldots, \beta_p$ the regression coefficients

▷ ε_i the error term for subject i

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

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

  where

  ▷ $\texttt{serBilir}_i$ denotes the serum bilirubin of patient i at baseline

  ▷ $\texttt{Age}_i$ and $\texttt{D-penicil}_i$ denote the Age and whether patient i received D-penicil or placebo

  ▷ β_0, β_1, and β_2 are the regression coefficients

  ▷ ε_i are the error terms

- 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 log 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**

	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

- Linear regression model with *matrix notation*

  ▷ the linear regression model for the n subjects

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

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

$$\vdots$$

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

- Linear regression model with *matrix notation*

  ▷ the linear regression model for the n subjects

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

$$
\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}
$$

- 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`

• Maximum likelihood estimators

$$
\begin{cases}
\widehat{\beta} = (X^\top X)^{-1} X^\top y \\[2em]
\widehat{\sigma}^2 = \dfrac{1}{n}(y - X\widehat{\beta})^\top (y - X\widehat{\beta})
\end{cases}
$$

where

$\triangleright$ $X^\top$ denotes the *transpose* of matrix X

$\triangleright$ $X^\top X$ denotes the *matrix product* between matrices $X^\top$ and X

$\triangleright$ $(X^\top X)^{-1}$ denotes the *matrix inverse* of matrix $(X^\top X)$

- 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**

- <u>Note:</u> 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

- 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

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

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

- 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

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

- 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

$\triangleright$ y_i the vector of responses for the i-th subject

$\triangleright$ X_i design matrix describing the structural component

$\triangleright$ V_i covariance matrix describing the variance and correlation structures

The covariance matrix V_i explicitly accounts for the correlations

- Interpretation of β

  ▷ β_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)

$$\begin{cases} \sqrt{\texttt{CD4}_{ij}} = \beta_0 + \beta_1 \texttt{Time}_{ij} + \beta_2 \{\texttt{ddI}_i \times \texttt{Time}_{ij}\} + \varepsilon_{ij}, \\[2em] \varepsilon_i \sim \mathcal{N}(0, V_i) \end{cases}$$

	Value	Std.Err.	t-value	p-value
β_0	7.189	0.221	32.593	< 0.001
β_1	-0.156	0.017	-9.247	< 0.001
β_2	0.016	0.024	0.662	0.508

▷ Coefficient β_1: For patients in the ddC group, every month the average $\sqrt{\text{CD4}}$ changes by -0.156

▷ Coefficient β_2:

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

- 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

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

- Note: Interaction terms for longitudinal data

  ▷ Consider the model

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

  * 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*)
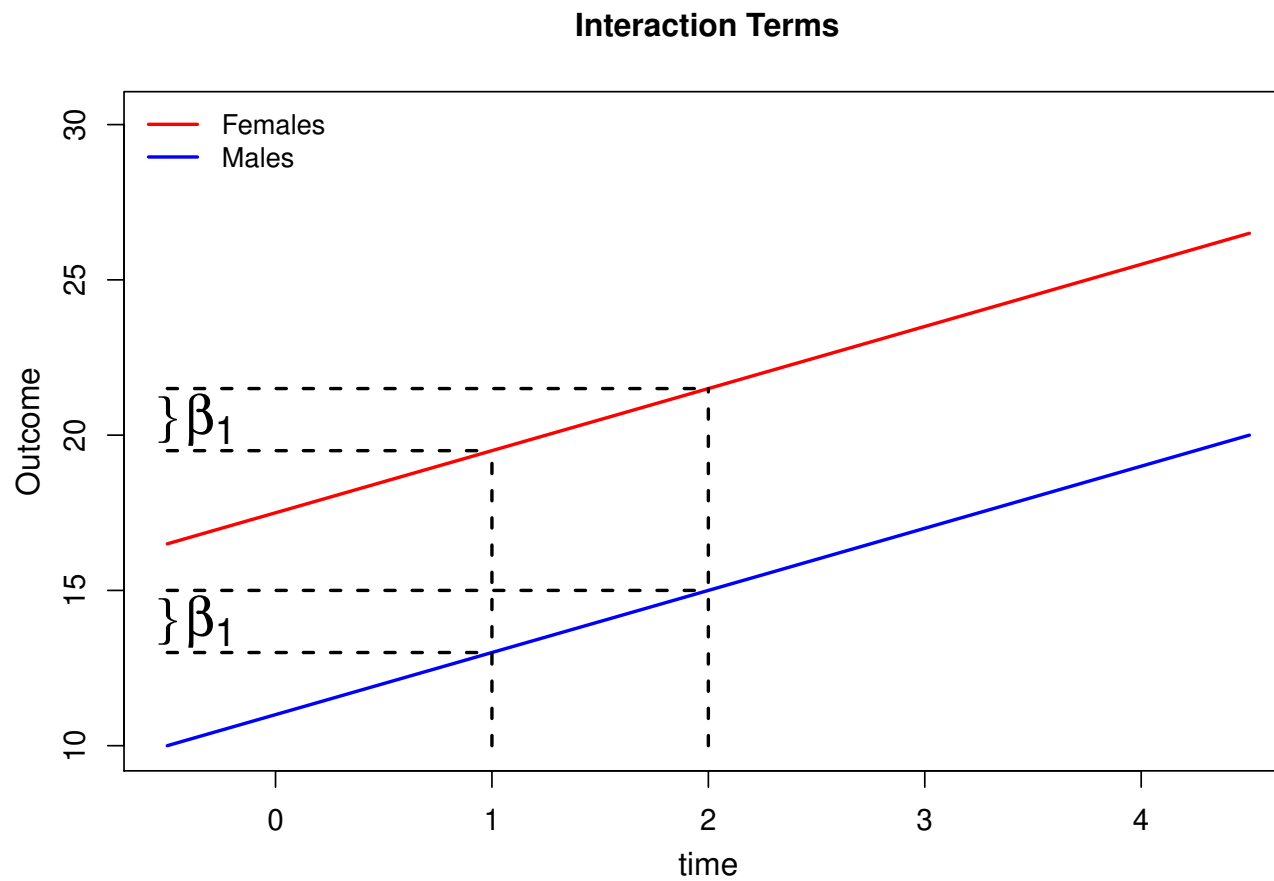
Interaction Terms

- <u>Note:</u> Interaction terms for longitudinal data

  ▷ if we would like different longitudinal evolutions for the two sexes we need to include the *interaction term*

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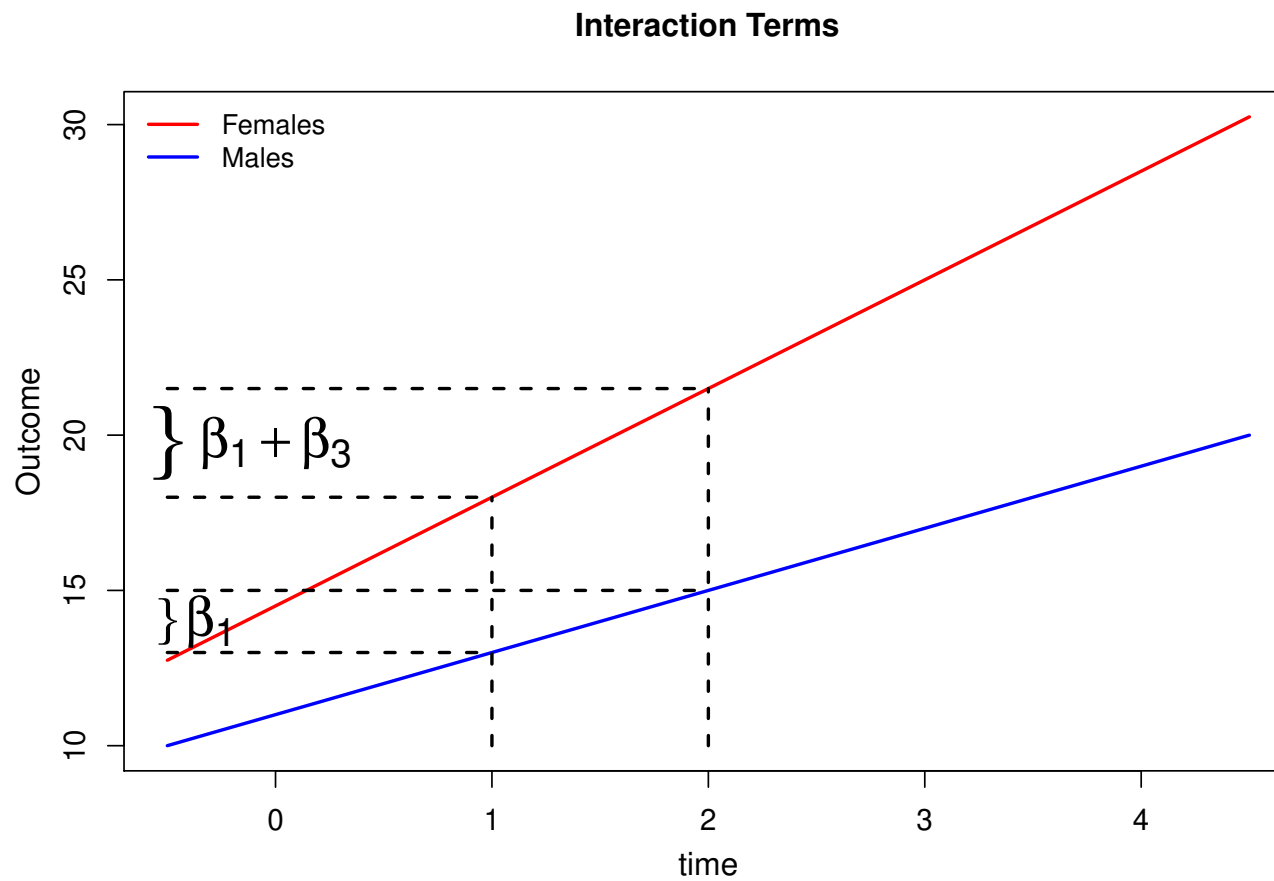
Interaction Terms

- <u>Note:</u> Nonlinear terms for longitudinal data

  ▷ Consider the model

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

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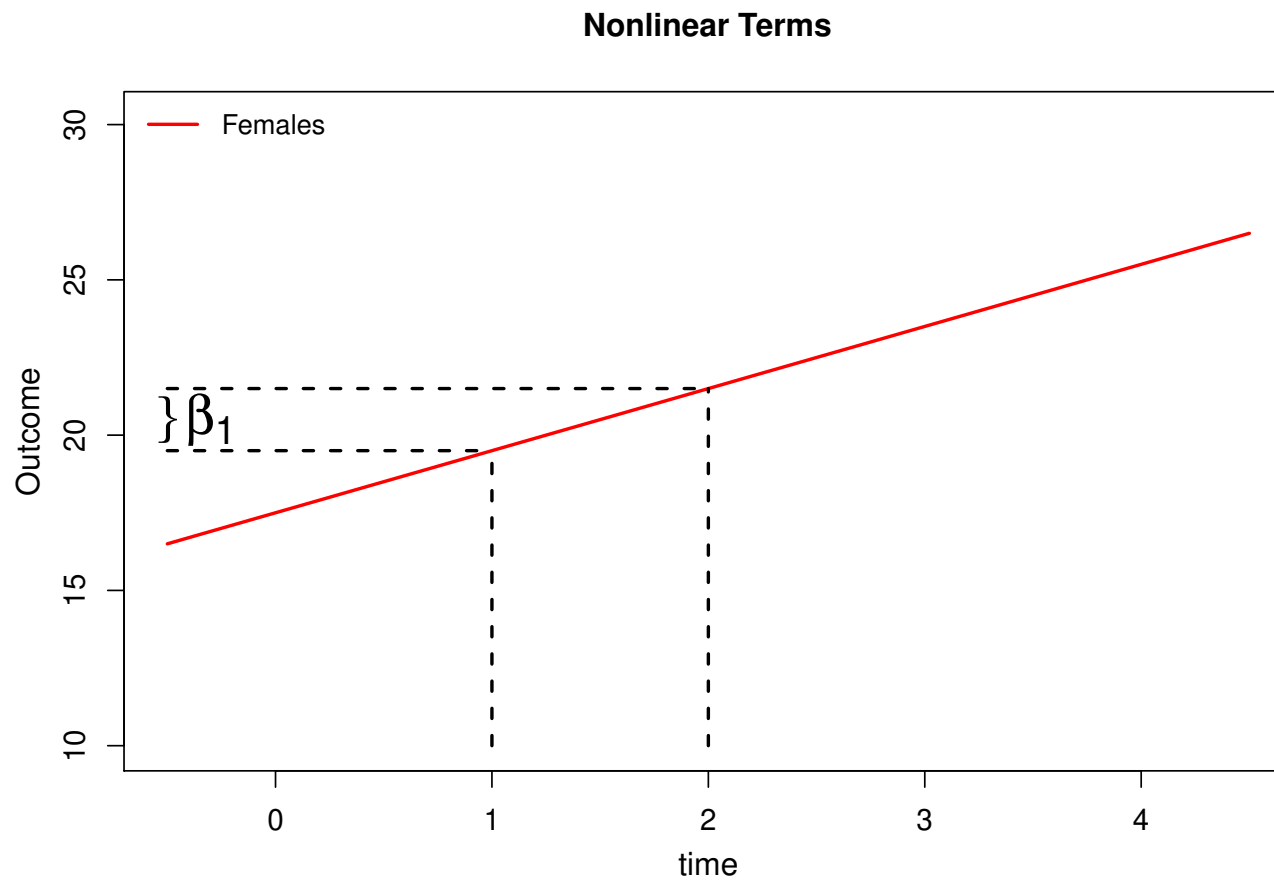
Nonlinear Terms

- <u>Note:</u> 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

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

- 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 latter

  allows 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

<div style="border:1px solid black; text-align:center; color:red; font-weight:bold;">

From the two approaches, splines are preferable

</div>

- <u>Note:</u> 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

- 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

- **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 β we obtain in the output do not often have a straightforward interpretation

- 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:

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

- The terms $N(\mathtt{Time}_{ij})_1$ and $N(\mathtt{Time}_{ij})_2$ 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
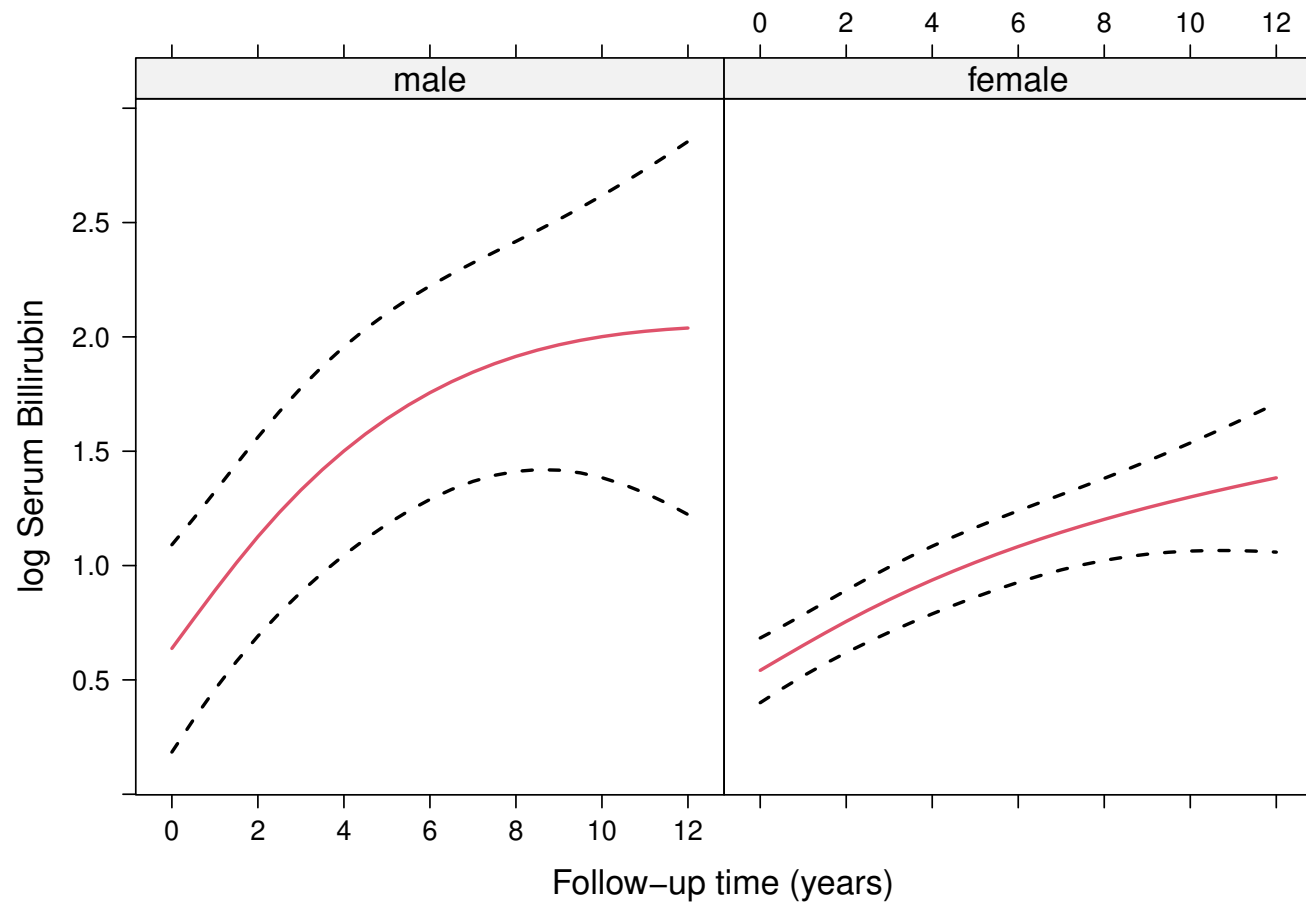
- Estimation of model parameters

  ▷ For known covariance matrix V_i, the regression coefficients are estimated using generalized least squares

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

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

- 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

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

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

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

> **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

▷ **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 <u>cannot</u> 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

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
⋮	⋮	⋮	⋮	⋮

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^2
⇒ `cd4 ~ time + I(time^2)`

R> <u>Note:</u> 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

- <u>Reminder:</u> 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
⋮	⋮	⋮	⋮	⋮

- 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.,

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

- Properties

  ▷ on the diagonal the variances, off diagonal covariances

  ▷ symmetric $\Rightarrow \mathrm{cov}(Y_1, Y_2) = \mathrm{cov}(Y_2, Y_1)$

- 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)

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

- 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   ▷ Gaussian spatial correlation

  ▷ autoregressive process   ▷ Toeplitz

  ▷ exponential spatial correlation ▷ . . .

- 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

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

▷ 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

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

- 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

- 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?

- Let's try the app. . .

- 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

- 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

- 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

- **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

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

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

  where

  - $\triangleright$ $\ell(\hat{\theta}_0)$ the value of the log-likelihood function under the null hypothesis, i.e., the special case model

  - $\triangleright$ $\ell(\hat{\theta}_a)$ the value of the log-likelihood function under the alternative hypothesis, i.e., the general model

  - $\triangleright$ 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 of ML likelihoods of the models (preferable is REML)

* 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:

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

$\triangleright$ versus the unstructured model

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

- We can rewrite the two hypothesis as

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

$$H_a: \quad \text{at least one variance or a covariance is not equal to the others}$$

- The likelihood ratio test gives:

	df	logLik	LRT	p-value
Comp Symm	5.00	-3586.91		
General	18.00	-3544.46	84.90	<0.0001

- 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

$\triangleright$ $\ell(\hat{\theta})$ is the value of the log-likelihood function

$\triangleright$ n_{par} the number of parameters in the model

$\triangleright$ n the number of subjects (independent units)

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

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

$$
\begin{cases}
M_1 : \mathtt{pro}_{ij} = \beta_0 + \beta_1 \mathtt{Time}_{ij} + \beta_2\{\mathtt{predn}_i \times \mathtt{Time}_{ij}\} + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, V_i^{Exp}) \\[2em]
M_2 : \mathtt{pro}_{ij} = \beta_0 + \beta_1 \mathtt{Time}_{ij} + \beta_2\{\mathtt{predn}_i \times \mathtt{Time}_{ij}\} + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, V_i^{Gauss})
\end{cases}
$$

- 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

- The models we have assumed for the Prothrobin 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.,

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

where

  ▷ δ 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

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

$$H_0 : \delta = 0 \text{ homoscedastic model}$$
$$H_a : \delta \neq 0 \text{ heteroscedastic model}$$

- 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	5.55	0.0185

- <u>Notes:</u> 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

- **Hypothesis testing for the regression coefficients** β: 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

- 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

- Tests for <u>groups of coefficients</u> are based on the F-test

  ▷ the set of hypotheses is:

$$
\begin{aligned}
H_0 : & \quad L\beta = 0 \\
H_a : & \quad L\beta \neq 0
\end{aligned}
$$

  where L is the contrasts matrix

  ▷ the F test statistic is

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

- 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

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

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

- We are interested in

  - ▷ the effect of Age, and

  - ▷ the overall effect of Sex

- For the effect of Age we set the hypotheses:

$$H_0: \quad \beta_3 = 0$$
$$H_a: \quad \beta_3 \neq 0$$

- The output of the model gives: . . .

	Value	Std.Err.	t-value	p-value
β_0	0.940	0.395	2.382	0.017
β_1	0.154	0.034	4.546	< 0.001
β_2	-0.281	0.218	-1.291	0.197
β_3	-0.002	0.006	-0.361	0.718
β_4	-0.014	0.020	-0.670	0.503
β_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

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

$$H_0: \quad \beta_2 = \beta_5 = 0$$
$$H_a: \quad \text{either } \beta_2 \text{ or } \beta_5 \text{ are not equal to } 0$$

- We **cannot** obtain the p-value for this test directly from the output

- We have six parameters, the contrast matrix L is

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

- 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

- The two models are:

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

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

  ▷ for both models V_i has a continuous AR1 structure

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

$$\begin{aligned} H_0 : \quad & \beta_2 = \beta_5 = 0 \\ H_a : \quad & \text{either } \beta_2 \text{ or } \beta_5 \text{ are not equal to } 0 \end{aligned}$$

- 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

- <u>Notes:</u> Hypothesis testing for the regression coefficients β

  ▷ The likelihood ratio test, and the classical univariate and multivariate Wald tests (i.e., using the χ^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

  ▷ β^* 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

- For example, for the k-th regression coefficient β_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.,

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

  $\triangleright$ 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)

- 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

- 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 α (e.g., $\alpha = 0.05$)

  - ▷ the percentage of times the test was significant is the estimated power for the specific setting under consideration

- <u>Notes:</u> 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

All statistical models are based on assumptions

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

- 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 ε_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

- To validate these assumptions we need an estimate of the error terms ε_{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

---

- 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

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

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

where

▷ $\widehat{H}_i$ is an upper-triangular matrix with the property $\widehat{H}_i^\top \widehat{H}_i = \widehat{V}_i$, with $\widehat{V}_i$ 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
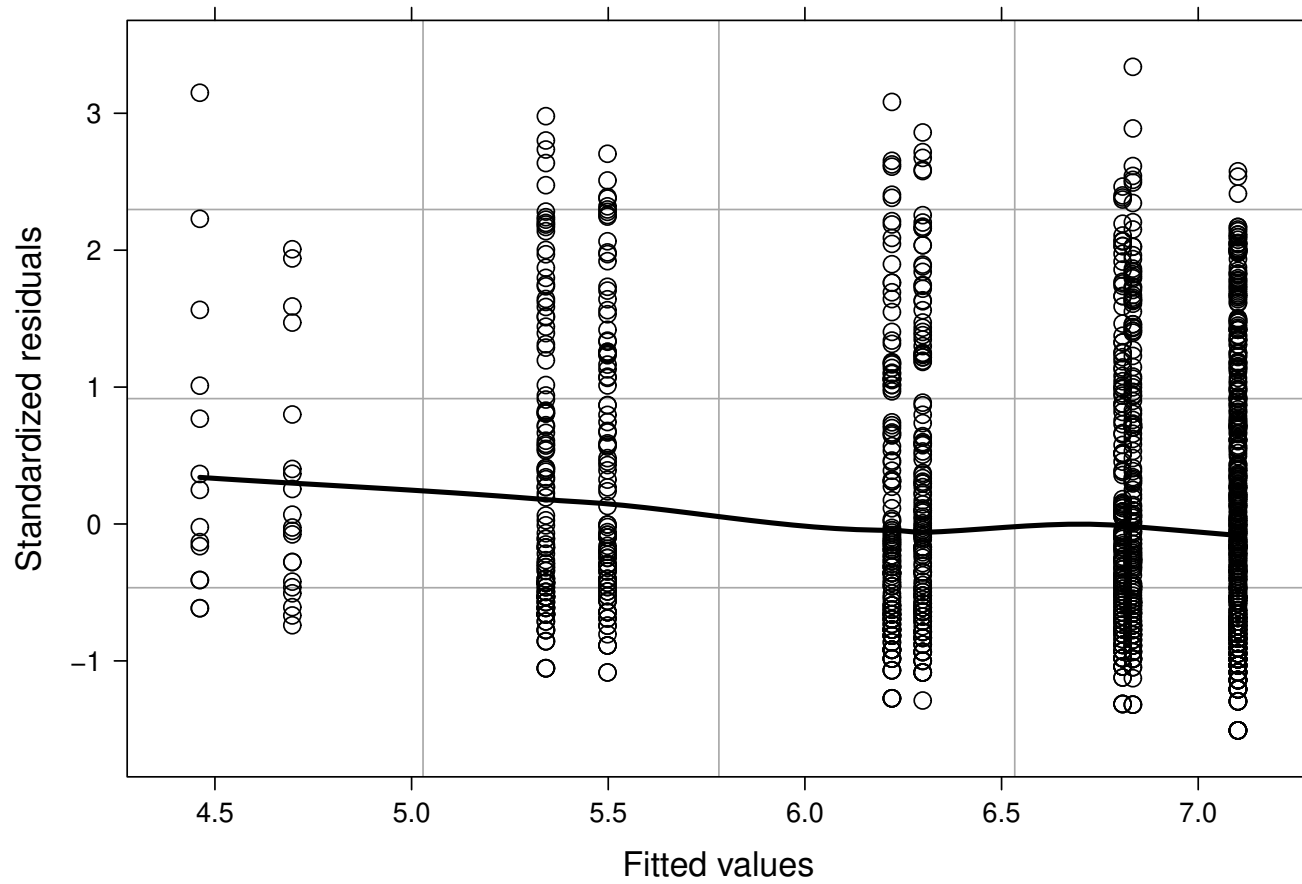
2.12 Residuals (cont'd)

- 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

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

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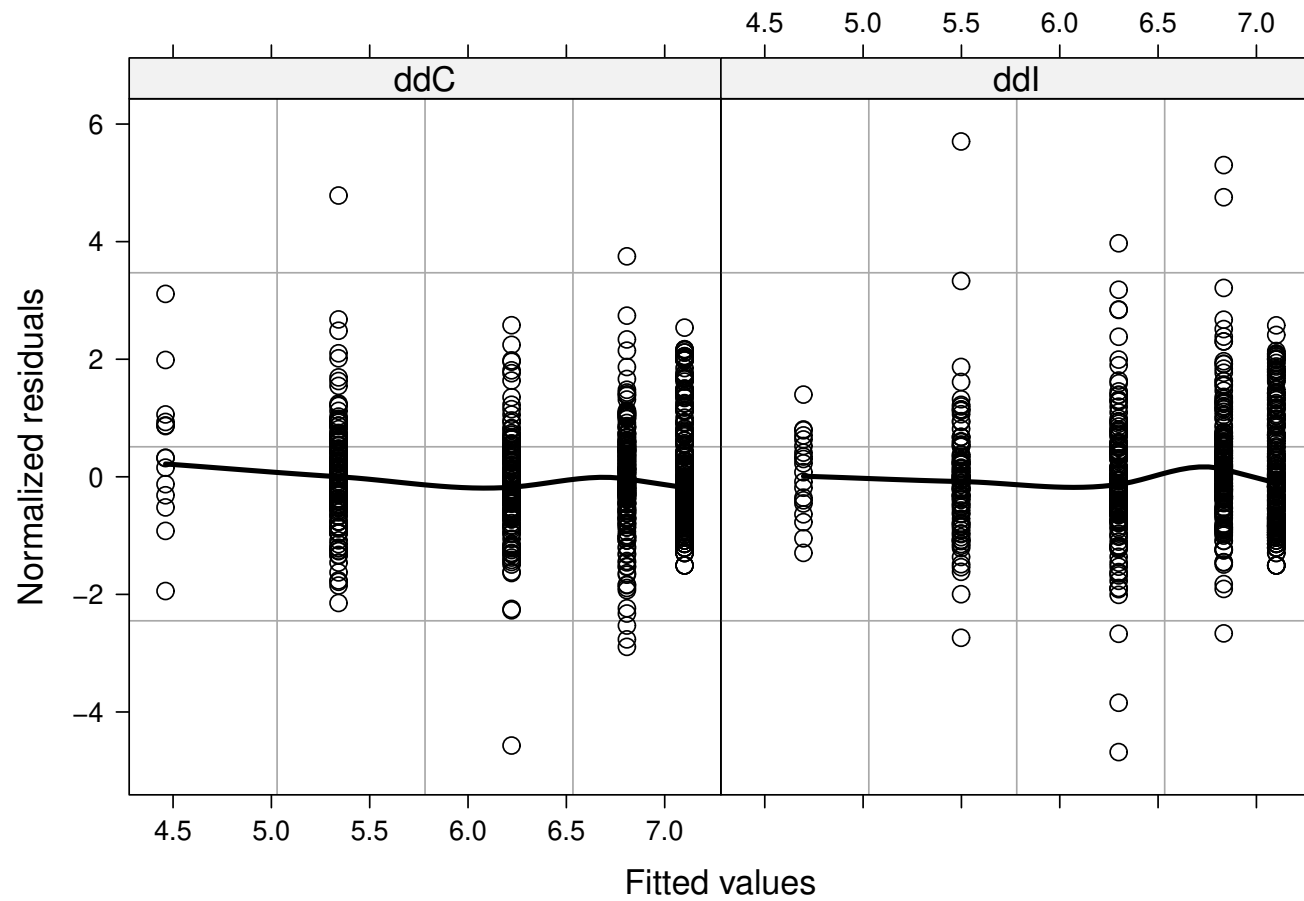\varepsilon_i \sim \mathcal{N}(0, V_i), \quad V_i \text{ is unstructured}
\end{cases}
$$

by plotting

▷ the standardized residuals versus fitted values

▷ the normalized residuals versus fitted values per treatment group

▷ QQ-plot of the standardized residuals

- 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

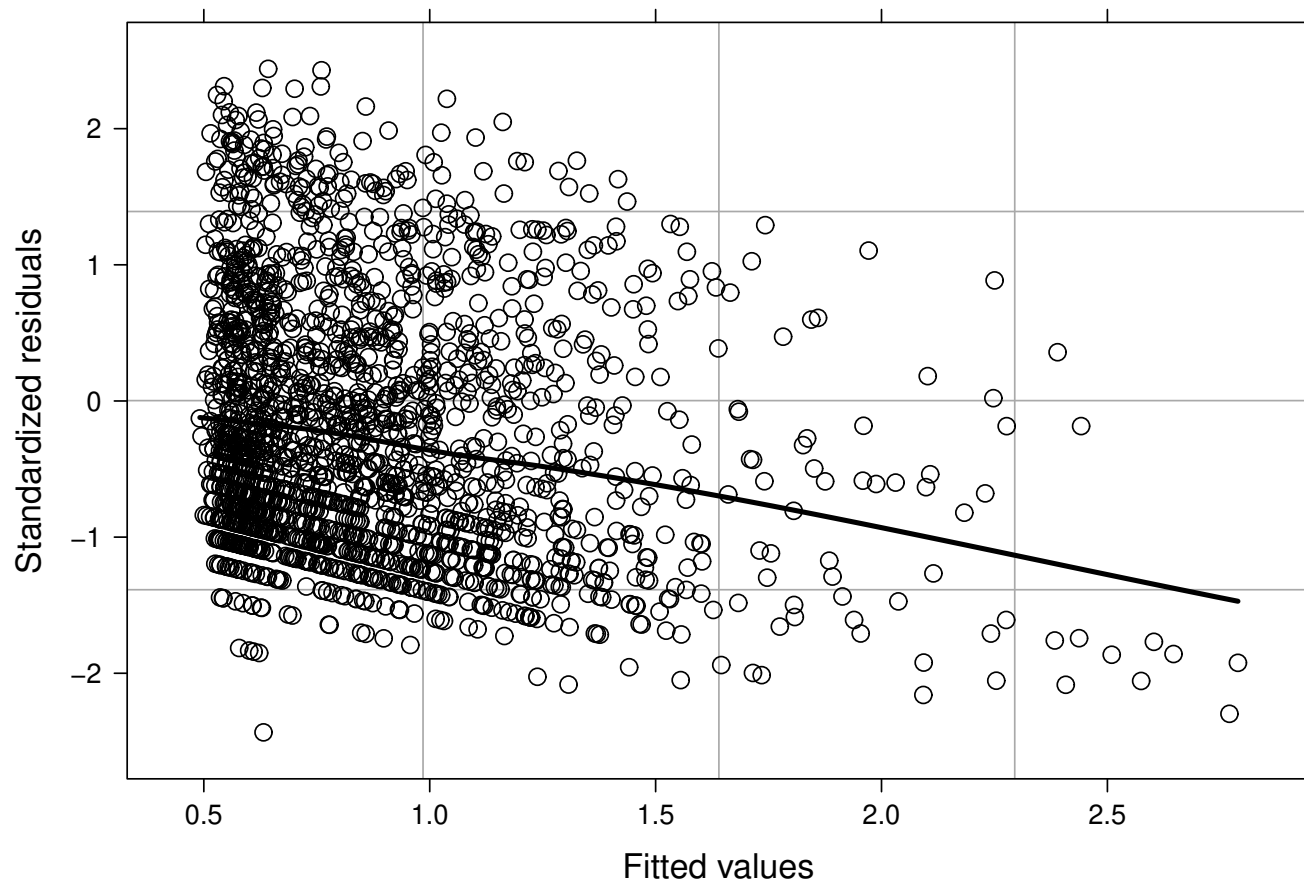- $\boxed{\text{Example:}}$ We continue by evaluating the assumptions of the model we have fitted to the PBC dataset:

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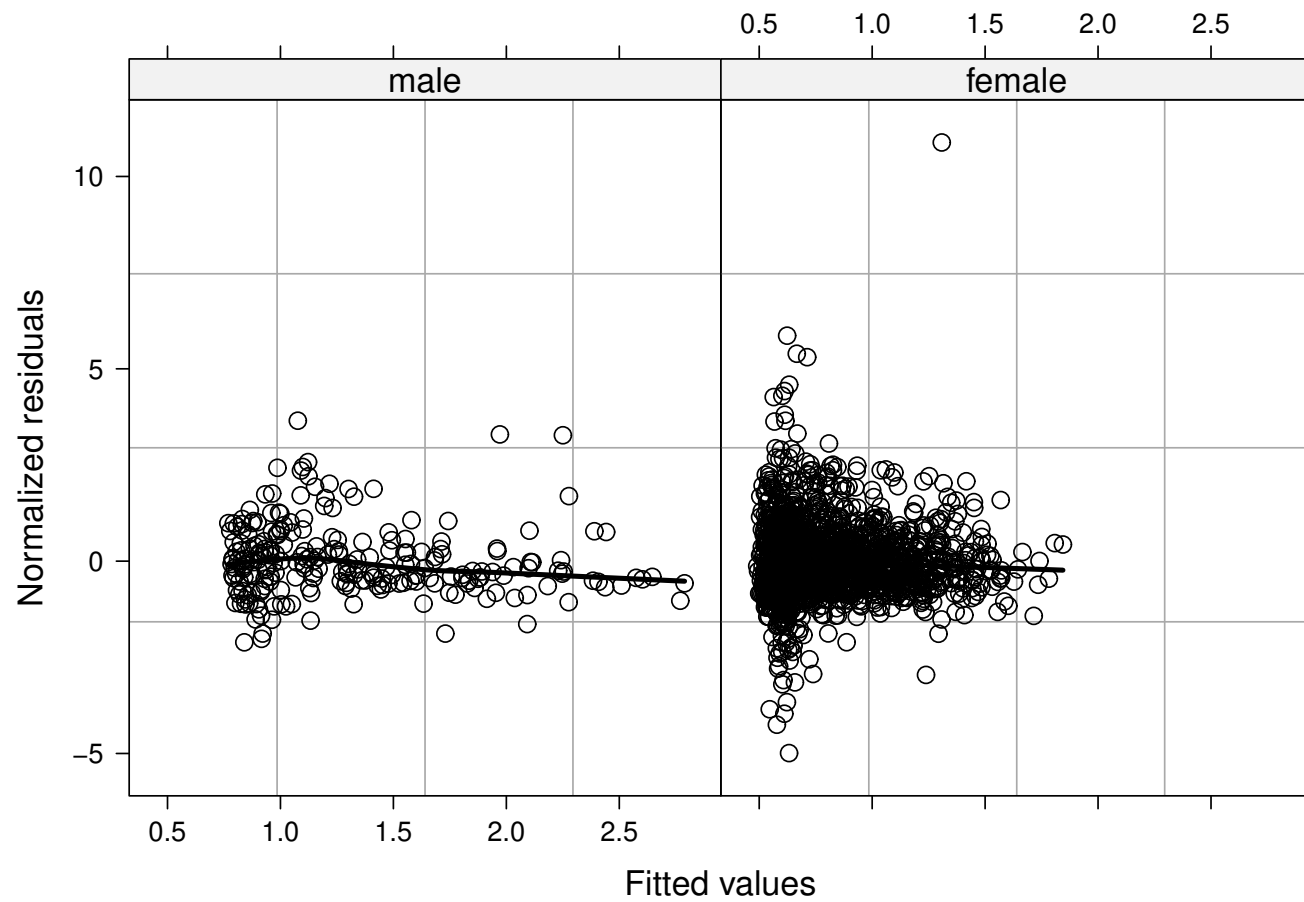\end{cases}
$$

by plotting again

  ▷ the standardized residuals versus fitted values

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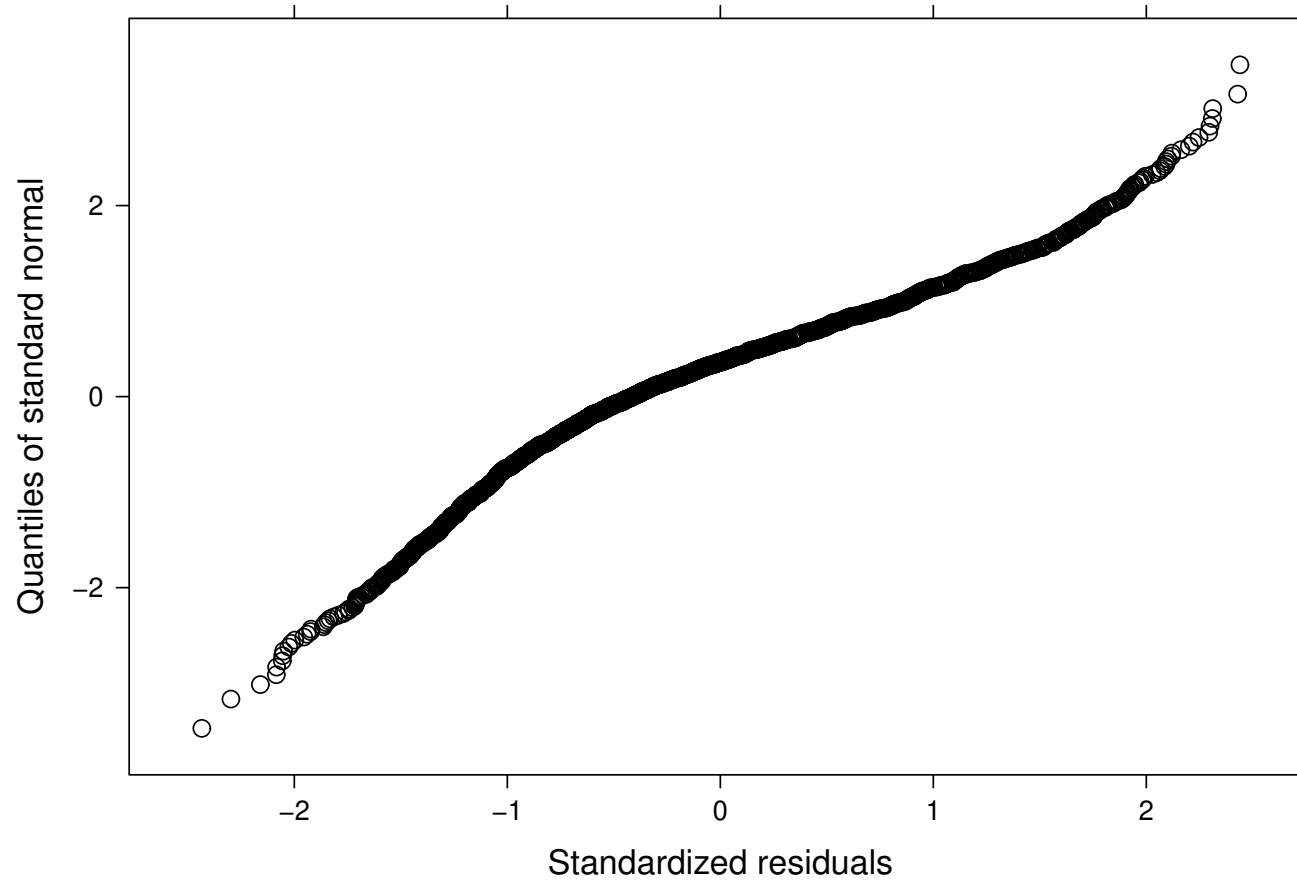
  ▷ QQ-plot of the standardized residuals

- 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 we return to make inference for the mean structure

- 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

Chapter 3

The Linear Mixed Effects Model

- In the previous chapter we focused on the *multivariate regression model*

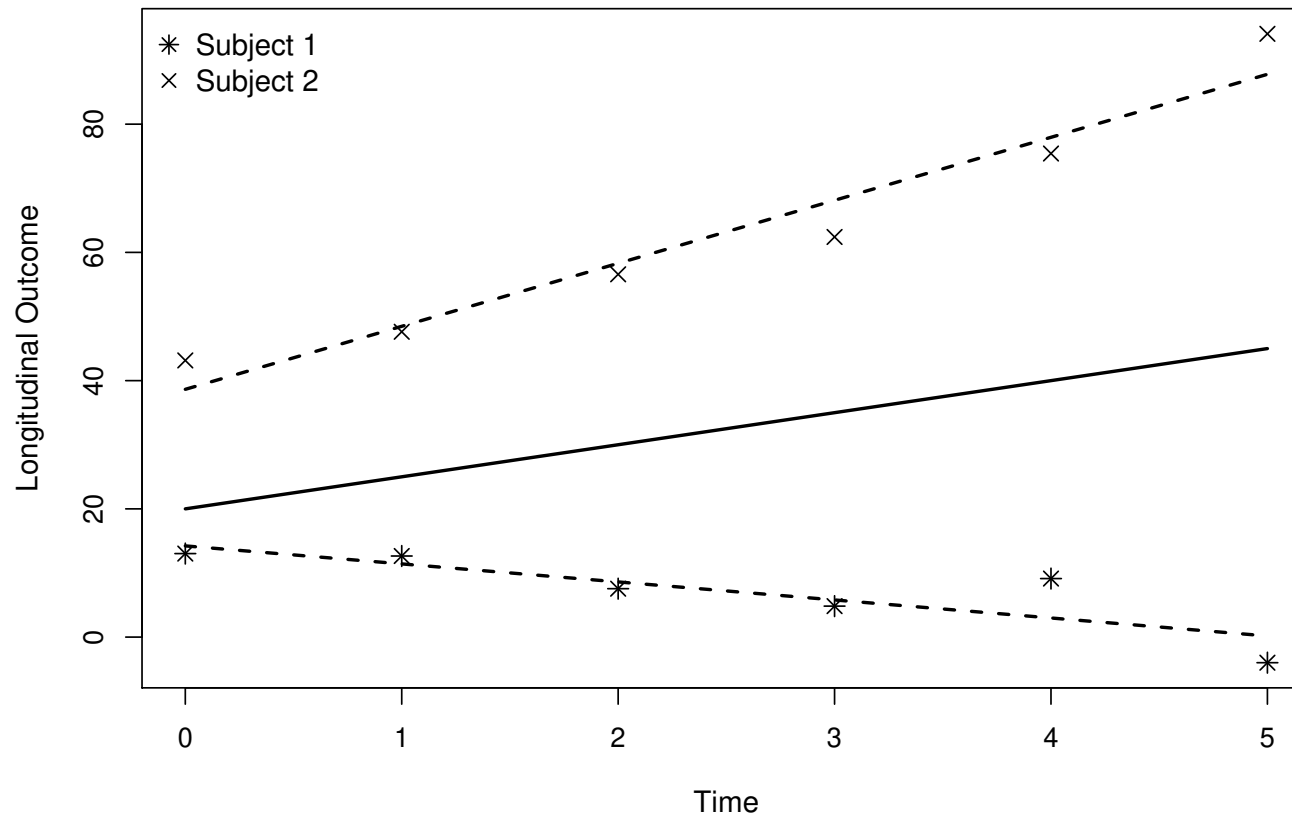$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, V_i),$$

  where

  $\triangleright$ y_i the vector of responses for the ith subject

  $\triangleright$ X_i design matrix describing structural component

  $\triangleright$ V_i covariance matrix describing the correlation structure

• **Alternative intuitive approach:** Each subject in the population has her own subject-specific mean response profile over time

- The evolution of each subject over time can be described by a linear model

$$y_{ij} = \tilde{\beta}_{i0} + \tilde{\beta}_{i1} t_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

where

▷ y_{ij} the jth response of the ith subject

▷ $\tilde{\beta}_{i0}$ is the intercept and $\tilde{\beta}_{i1}$ the slope for subject i

- **Assumption:** Subjects are randomly sampled from a population $\Rightarrow$ subject-specific regression coefficients are also sampled from a population of regression coefficients

$$\tilde{\beta}_i \sim \mathcal{N}(\beta, D)$$

- We can reformulate the model as

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \varepsilon_{ij},$$

  where

  ▷ βs are known as the *fixed effects*

  ▷ b_is are known as the *random effects*

- In accordance for the random effects we assume

$$b_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim \mathcal{N}(0, D)$$

- Put in a general form

$$
\begin{cases}
y_i &= X_i\beta + Z_i b_i + \varepsilon_i, \\[2em]
b_i &\sim \mathcal{N}(0, D), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_{n_i}),
\end{cases}
$$

with

▷ X design matrix for the fixed effects β

▷ Z design matrix for the random effects b_i

▷ b_i and ε_i are assumed independent

- Fixed and random effects:

  ▷ β_j denotes the change in the average y_i when x_j is increased by one unit

  ▷ b_i are interpreted in terms of how a subset of the regression parameters for the ith subject deviates from those in the population

- Advantageous feature: population + subject-specific predictions

  ▷ β describes mean response changes in the population

  ▷ $\beta + b_i$ describes individual response trajectories

3.2 Interpretation (cont'd)

- Example: We fit a linear mixed model for the AIDS dataset assuming

  ▷ different average longitudinal evolutions per treatment group (fixed part)

  ▷ random intercepts & random slopes (random part)

$$
\begin{cases}
\sqrt{\text{CD4}_{ij}} = \beta_0 + \beta_1 \text{Time}_{ij} + \beta_2\{\text{ddI}_i \times \text{Time}_{ij}\} + b_{i0} + b_{i1}\text{Time}_{ij} + \varepsilon_{ij}, \\
\\
b_i \sim \mathcal{N}(0, D), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)
\end{cases}
$$

- Note: We did not include a main effect for treatment due to randomization

	Value	Std.Err.	t-value	p-value
β_0	7.189	0.222	32.359	< 0.001
β_1	-0.163	0.021	-7.855	< 0.001
β_2	0.028	0.030	0.952	0.342

- No evidence of differences in the average longitudinal evolutions between the two treatments

3.2 Interpretation (cont'd)

* <u>Interaction & nonlinear terms</u>: As we have seen in the previous chapter (see pp. 59–71), often

    ▷ the effect of some predictors may be nonlinear (e.g., time effect), and/or

    ▷ the effect of some predictors on the outcome may be influenced from other predictors (e.g., different average longitudinal evolutions per treatment group)

* In such cases, we need to consider more elaborate models that contain terms to capture these features, namely

    ▷ polynomials or splines to model nonlinearities

    ▷ interaction effects

- When such terms are included in the model, the interpretation of the parameters can become quite complicated

- To understand a complex mixed model we can visualize it using **effect plots**

- ⬚ Example: ⬚ We fit a model to the PBC dataset for serum bilirubin that contains

  ▷ *fixed effects:*

  * nonlinear time effect with splines, main effect of sex, age and baseline prothrombin
  * interaction effects of sex with nonlinear time, age and baseline prothrombin

  ▷ *random effects:* nonlinear time effect

- The model has the form:

$$\log(\texttt{serBilir}_{ij}) = \beta_0 + \beta_1 N(\texttt{Time}_{ij})_1 + \beta_2 N(\texttt{Time}_{ij})_2 + \beta_3 \texttt{Female}_i + \beta_4 \texttt{Age}_i +$$
$$\beta_5 \texttt{basePro}_i + \beta_6 \{\texttt{Female}_i \times \texttt{Age}_i\} +$$
$$\beta_7 \{\texttt{Female}_i \times \texttt{basePro}_i\} + \beta_8 \{\texttt{Female}_i \times N(\texttt{Time}_{ij})_1\} +$$
$$\beta_9 \{\texttt{Female}_i \times N(\texttt{Time}_{ij})_2\} + b_{i0} + b_{i1} N(\texttt{Time}_{ij})_1 +$$
$$b_{i2} N(\texttt{Time}_{ij})_2 + \varepsilon_{ij}$$

where

▷ the terms $N(\texttt{Time}_{ij})_1$ and $N(\texttt{Time}_{ij})_2$ denote the basis for a natural spline with two degrees of freedom

▷ $b_i \sim \mathcal{N}(0, D)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

- 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 prothrombin time of 10.6 sec

  ▷ for the average age of 49 years old

  ▷ including also the corresponding 95% pointwise confidence intervals

  ▷ (in the app different ages and prothrombin times can be selected)
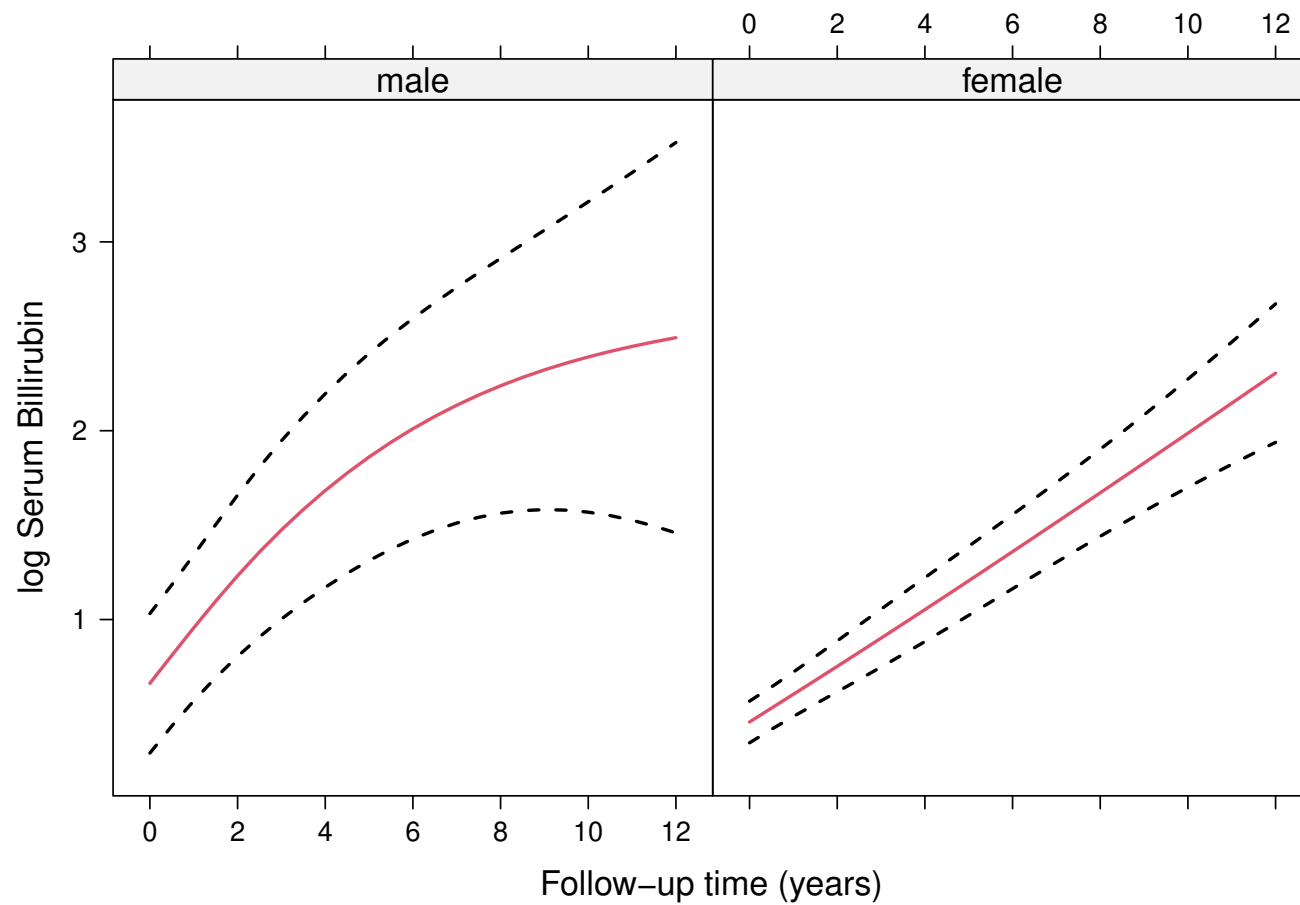
- How do the random effects capture correlation:

  ▷ Given the random effects, the measurements of each subject are independent (*conditional independence assumption*)
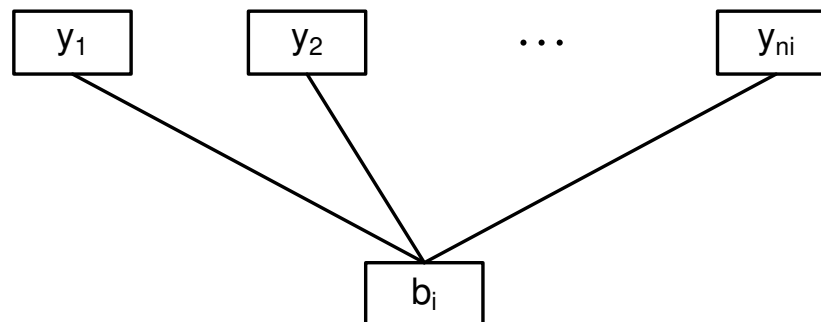
$$p(y_i \mid b_i) = \prod_{j=1}^{n_i} p(y_{ij} \mid b_i)$$

  ▷ Marginally (integrating out the random effects), the measurements of each subject are correlated

$$p(y_i) = \int p(y_i \mid b_i)\, p(b_i)\, db_i \quad \Rightarrow \quad y_i \sim \mathcal{N}(X_i\beta,\; Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i})$$

Graphical representation of the conditional independence assumption

3.3 Hierarchical vs Marginal (cont'd)

• Hence, with random effects we again model the correlations in the repeated measurements of each subject

• <u>Notes:</u> In using random effects for modeling the covariance matrix

   ▷ The more random effects we include the more flexibly we capture the correlations

   ▷ By using random effects (other than random intercept alone) we also directly allow for heteroscedasticity (i.e., non-constant variances over time)

   ▷ Nevertheless, we do assume a particular type of structure for the correlations and the variances – they are **not** allowed completely free

   ▷ Random effects work equally well with balanced or unbalanced data

- Let's try the app...

- Hierarchical formulation

  ▷ a model for y_i given b_i, and a model for b_i

  ▷ D is the covariance matrix of the random effects $\Rightarrow$ **needs to be positive definite**

- Marginal formulation

  ▷ a model for y_i, and a specific form of the marginal covariance matrix
  $V_i = Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i}$

  ▷ only V_i needs to be positive definite

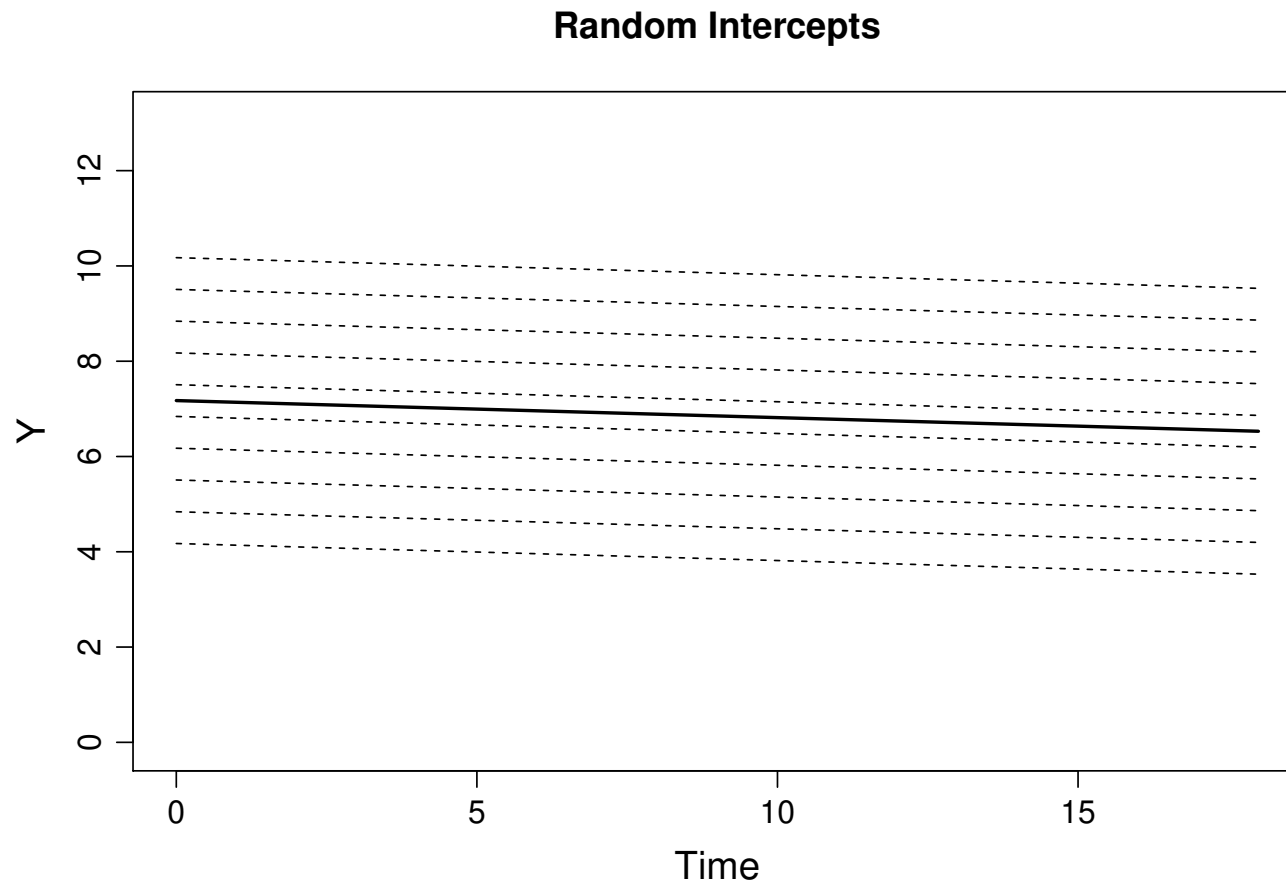 ▷ V_i **can be positive definite without** D **being positive definite**

The hierarchical model implies the marginal one,
<u>not</u> vice versa

• A simple example: Random-intercepts model

$$
\begin{cases}
y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + \varepsilon_{ij}, \\[2em]
b_{i0} \sim \mathcal{N}(0, \sigma_b^2), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2).
\end{cases}
$$

Random Intercepts

3.3 Hierarchical vs Marginal (cont'd)

• Implied marginal covariance matrix has the form

$$V_i = \sigma_b^2 1_{n_i} 1_{n_i}^\top + \sigma^2 I_{n_i}$$

it assumes

▷ constant variance $\sigma_b^2 + \sigma^2$ over time, and

▷ equal positive correlation $\rho = \sigma_b^2/(\sigma_b^2 + \sigma^2)$ between the measurements of any two time points (aka *intra-class correlation*)

▷ it is known as the *compound symmetric* covariance matrix

- Note that we could also have a compound symmetric covariance matrix with negative intra-class correlation

  ▷ such a matrix could never have come from a mixed model

Random intercepts **imply** compound symmetry
but
Compound symmetry **does not imply** random intercepts

- What are the implications of this?

- Statistical software that fit mixed models under ML actually fit the implied marginal model

  ▷ we can construct examples where two mixed models have exactly the same implied marginal model

  ▷ based on the fitted model we **cannot** say under which model the data have been generated

- We can only do it under a Bayesian approach (because there we actually fit the hierarchical model)

- Fixed effects: For known marginal covariance matrix $V_i = Z_i D Z_i^\top + \sigma^2 I_{n_i}$, the fixed effects are estimated using generalized least squares

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

- Variance Components: The unique parameters in V_i are estimated based on either maximum likelihood (ML) or restricted maximum likelihood (REML)

   ▷ REML provides unbiased estimates for the variance components in small samples

- Two-step iterative procedure

  ▷ Step 0: Set initial values for D and σ^2

  ▷ Step 1: Calculate the covariance matrix $\widehat{V}_i^{it=k}$ and following the fixed effects $\hat{\beta}^{it=k}$

  ▷ Step 2: Update $\widehat{V}_i^{it=k+1}$ using REML or ML

  ▷ Step 3: Check convergence criterion, if not satisfied return to Step 1

  Steps 1–3 are repeated until convergence is attained

- Estimation of random effects

  ▷ based on a fitted mixed model, estimates for the random effects are based on the posterior distribution:

$$p(b_i \mid y_i; \theta) \;=\; \frac{p(y_i \mid b_i; \theta)\; p(b_i; \theta)}{p(y_i; \theta)}$$

$$\propto\; p(y_i \mid b_i; \theta)\; p(b_i; \theta),$$

  in which θ is replaced by its MLE $\hat{\theta}$

- This is a whole distribution

  ▷ in the linear mixed model we have seen, this posterior distribution has a closed-form:

$$[b_i \mid y_i; \theta] \sim \mathcal{N}\left\{ DZ_i^\top V_i^{-1}(y_i - X_i\beta),\ DZ_i^\top K Z_i D \right\},$$

  with

$$K = V_i^{-1} - V_i^{-1} X_i \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i \right)^{-1} X_i^\top V_i^{-1}$$

3.4 Estimation (cont'd)

- To obtain estimates for the random effects we typically use measures of location from this posterior distribution (e.g., mean or mode)

- Due to the fact that in linear mixed models we obtain a normal distribution (in which the mean and mode coincide), we use as estimates of the random effects the means of these distributions

$$\widehat{b}_i = D Z_i^\top V_i^{-1}(y_i - X_i \beta)$$

- These estimates are called the *empirical Bayes* estimates of the random effects

- Estimates of the random effects are very useful in prediction

    ▷ in this context there is an important difference between the marginal models we have seen in Chapter 2 and the mixed models of this chapter

- In particular, the predictions from a marginal model are

$$\widehat{y}_i^{marg} = X_i\widehat{\beta}$$

whereas from the mixed model we obtain

$$\widehat{y}_i^{subj} = X_i\widehat{\beta} + Z_i\widehat{b}_i$$
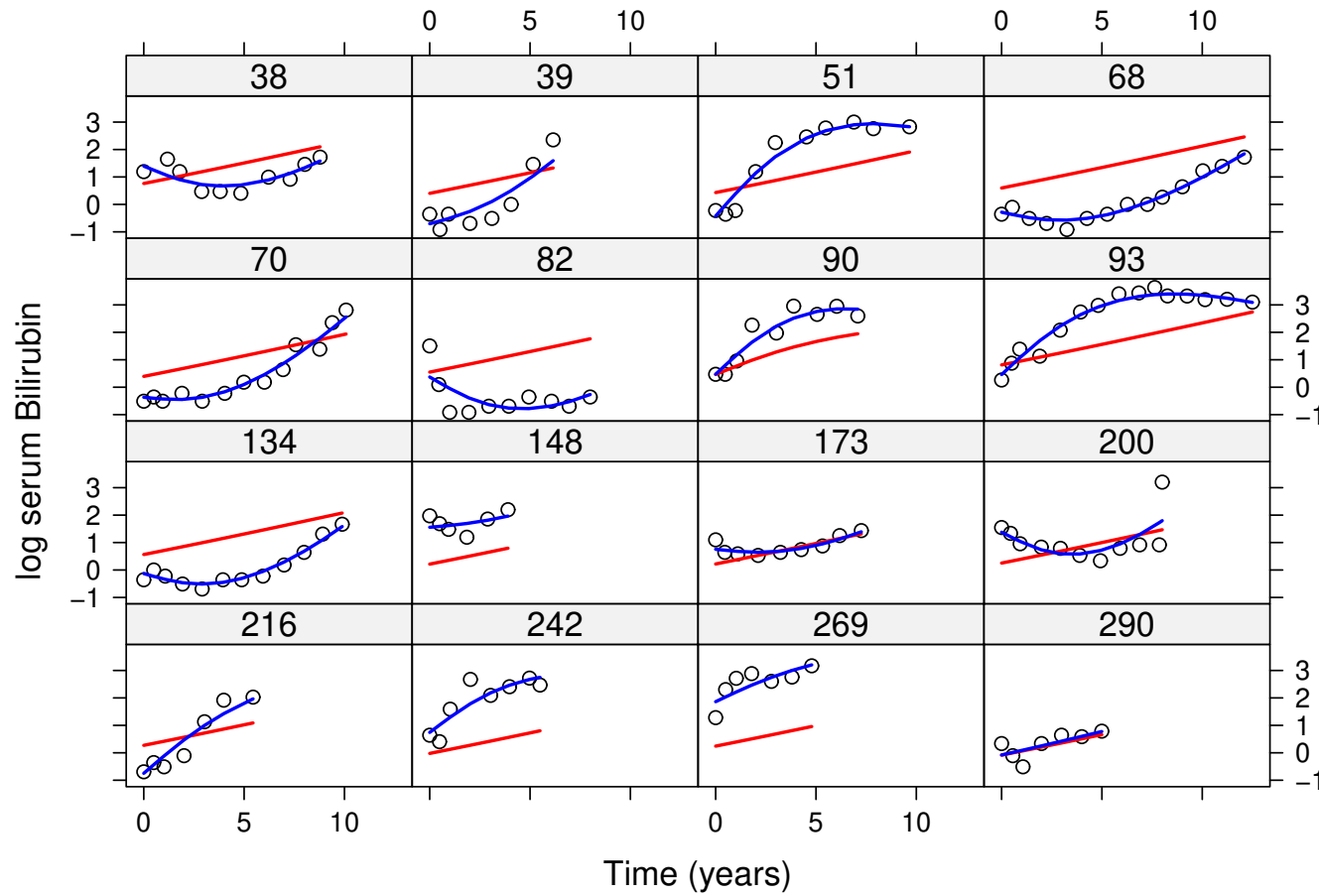
3.4 Estimation (cont'd)

- The difference is that

  ▷ from the marginal model we obtain predictions for the *'average'* patient having characteristics X_i (i.e., age, sex, etc.)

  ▷ from the mixed model we obtain predictions for the *'average'* patient that has characteristics X_i <u>and</u> observed data y_i (i.e., they have a subject-specific nature)

- The predictions $X_i\widehat{\beta} + Z_i\widehat{b}_i$ we obtain from the mixed model are called the *Best Linear Unbiased Predictions (BLUPs)*

  ▷ 'linear' because they are a linear combination of $\widehat{\beta}$ and $\widehat{b}_i$

  ▷ 'unbiased' because their average equals the true subject-specific mean

  ▷ 'best' because they have the smallest variance of all linear predictors

3.4 Estimation (cont'd)

- Example: To see an example of the difference between the marginal and subject-specific predictions, we compare the two sets of predictions for the complex linear mixed model we have seen in Section 3.2 (pp.156–159) for 16 randomly selected patients

  ▷ red lines denote the marginal predictions,

  ▷ blue lines denote the subject-specific predictions

  ▷ black circles the observed data

3.4 Estimation (cont'd)

- We clearly observe that the subject-specific predictions are much closer to the data of each individual patient than the marginal ones

3.5 Mixed-Effects Models in R

R> There are two primary packages in R for mixed models analysis:

▷ Package **nlme**

* fits linear & nonlinear mixed effects models, and marginal models for normal data
* allows for both random effects & correlated error terms
* several options for covariances matrices and variance functions

▷ Package **lme4**

* fits linear, nonlinear & generalized mixed effects models
* uses only random effects
* allows for nested and crossed random-effects designs

3.5 Mixed-Effects Models in R (cont'd)

R> The basic function to fit linear mixed models in the **nlme** package is `lme()`, and has three basic arguments

▷ `fixed`: a formula specifying the response vector and the fixed-effects structure

▷ `random`: a formula specifying the random-effects structure

▷ `data`: a data frame containing all the variables

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
⋮	⋮	⋮	⋮	⋮

R> The code used to fit the linear mixed model for the AIDS dataset (pp.153) is as follows

```
lmeFit <- lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = ~ obstime | patient)

summary(lmeFit)
```

3.5 Mixed-Effects Models in R (cont'd)

R> The same fixed-effects structure but only random intercepts

```
lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = ~ 1 | patient)
```

R> The same fixed-effects structure, random intercepts & random slopes, with a diagonal covariance matrix (using the `pdDiag()` function)

```
lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = list(patient = pdDiag(form = ~ obstime)))
```

3.5 Mixed-Effects Models in R (cont'd)

R> The basic function to fit linear mixed models in the **lme4** package is `lmer()`, and has two basic arguments

▷ `formula`: a formula specifying the response vector, the fixed- and random-effects structure

▷ `data`: a data frame containing all the variables

R> Again the data should be in the long format

R> The analogous code to fit the linear mixed model for the AIDS dataset (pp.153) is as follows

```
lmerFit <- lmer(CD4 ~ obstime + obstime:drug + (obstime | patient),
                data = aids)


summary(lmerFit)
```

R> To fit the same model but with a diagonal matrix for the random effects the call becomes:

```
lmerFit2 <- lmer(CD4 ~ obstime + obstime:drug +
                      (obstime || patient),
                 data = aids)

summary(lmerFit2)
```

3.6 Nested and Crossed Random Effects*

- In the previous examples the primary type of correlated data we have seen is longitudinal data

    ▷ correlations stems from the fact that we measure *the same* outcome repeatedly over time for each subject

- Another commonly encountered feature that induces correlation is clustering, e.g.,

    ▷ patients are clustered within hospitals

    ▷ children are clustered within schools or families

    ▷ . . .

- Example: In the Glaucoma data we have a multilevel clustered design (see pp.11)

  ▷ each location is nested within the hemifield

  ▷ each hemifield is nested within the eye

  ▷ each eye is nested within the patient

Measurements in the same cluster are expected to be **(positively) correlated**

- To account for the correlations in each level of the multilevel structure we can include level-specific random effects

- Continuing in the Glaucoma data example, we focus (for simplicity) in the higher two levels, namely the patient and the eye

  ▷ we fit a linear mixed model with a separate random effect per level

$$
\begin{cases}
\texttt{VF}_{ijk} = \beta_0 + \beta_1 \texttt{Time}_{ijk} + b_i + u_{ij} + \varepsilon_{ijk} \\
\quad b_i \sim \mathcal{N}(0, \sigma^2_{patient}), \qquad u_{ij} \sim \mathcal{N}(0, \sigma^2_{eye}), \\
\quad \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)
\end{cases}
$$

where

▷ VF_{ijk} denotes the k-th visual field sensitivity measurement for the j-th eye of the i-th patient

▷ Time_{ijk} denotes the corresponding time point this measurement was taken

▷ b_i is the random effect for the patients – the measurements of the i-th patient are correlated because all these measurements share the *same* random effect b_i

▷ u_{ij} is the random effect for the eye within the patient – the measurements of the j-th eye of the i-th patient are more correlated than the measurements of the j'-th eye because they share the *same* random effect u_{ij}

• The estimated variance components from the Glaucoma data are:

▷ $\sigma_{patient} = 4.3$

▷ $\sigma_{eye} = 5.8$

▷ $\sigma = 7.9$

• Based on these variance components we can compute the corresponding correlations, i.e.,

▷ measurements in the same eye have correlation

$$\frac{\sigma^2_{patient} + \sigma^2_{eye}}{\sigma^2_{patient} + \sigma^2_{eye} + \sigma^2} = 0.46$$

▷ and measurements from different eyes

$$\frac{\sigma^2_{patient}}{\sigma^2_{patient} + \sigma^2_{eye} + \sigma^2} = 0.16$$

• It goes without saying, that if the correlations in the data are more complex, we could include additional random effects

• Example: Continuing in the Glaucoma example, by including only random intercepts terms we assume that the correlations are constant over time

▷ as we have previously discussed, this may be a simplistic assumption for longitudinal data

• We extend the model by including a random slopes terms in the patient level, i.e.,

$$
\begin{cases}
\mathtt{VF}_{ijk} = \beta_0 + \beta_1 \mathtt{Time}_{ijk} + b_{i0} + b_{i1}\mathtt{Time}_{ijk} + u_{ij} + \varepsilon_{ijk} \\
\quad b_i \sim \mathcal{N}(0, D_{patient}), \qquad u_{ij} \sim \mathcal{N}(0, \sigma^2_{eye}), \\
\quad \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)
\end{cases}
$$

▷ now, in the patient level we have a covariance matrix D

- The estimated variance components from the Glaucoma data are:

  ▷ $\sigma_{patient,int} = 4.7$

  ▷ $\sigma_{patient,slp} = 0.4$

  ▷ $\mathsf{corr}_{patient,int-slp} = -0.4$

  ▷ $\sigma_{eye} = 5.8$

  ▷ $\sigma = 7.8$

- The examples we have seen so far in this section refer to settings in which the measurements of one level are *nested* within another level

  ▷ due to this feature, the random effects we have used in the previous examples of the Glaucoma data are called *nested random effects*

- However, there are also settings in which we have different types of groupings of measurements that are not nested

  ▷ in these cases we use *crossed random effects*

- Example: One feature of visual field sensitivity measurements is that they exhibit the so-called *Global Visit Effect* (see pp.12)

  ▷ in particular, for some visits some patients showed strangely low sensitivity to the stimuli

  ▷ in the next visit, their sensitivity levels improved

  ▷ it is not possible this low sensitivity to be due to Glaucoma because it is an irreparable disease

  ▷ hence, the low sensitivity measurements are attributed to other reasons (e.g., tiredness)

- To capture this Global Visit Effect we can include a random effect for each visit

  ▷ this random effect is **not** nested to the previously used random effects

Erasmus MC

- Hence, our model now becomes

$$
\begin{cases}
\texttt{VF}_{ijk} &= \beta_0 + \beta_1 \texttt{Time}_{ijk} + b_i + \texttt{v}_k + \varepsilon_{ijk} \\
b_i &\sim \mathcal{N}(0, \sigma^2_{patient}), \qquad \texttt{v}_k \sim \mathcal{N}(0, \sigma^2_{visit}), \\
\varepsilon_{ijk} &\sim \mathcal{N}(0, \sigma^2)
\end{cases}
$$

- The estimated variance components from the Glaucoma data are:

  ▷ $\sigma_{patient} = 5.9$

  ▷ $\sigma_{visit} = 0.8$

  ▷ $\sigma = 8.9$

• We have seen two classes of models for longitudinal data, namely

▷ *Marginal Models*

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

▷ *Conditional Models*

$$\begin{cases} y_i = X_i\beta + Z_i b_i + \varepsilon_i, \\ \\ b_i \sim \mathcal{N}(0, D), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_{n_i}) \end{cases}$$

- It is also possible to combine the two approaches and obtain a linear mixed model with correlated error terms

$$
\begin{cases}
y_i = X_i\beta + Z_i b_i + \varepsilon_i, \\[2em]
b_i \sim \mathcal{N}(0, D), \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma_i),
\end{cases}
$$

where, as in marginal models, we can consider different forms for Σ_i

- The corresponding marginal model is of the form

$$
y_i \sim \mathcal{N}(X_i\beta, \; Z_i D Z_i^\top + \Sigma_i)
$$

3.7 Mixed Models with Correlated Errors (cont'd)

- Features

  ▷ both b_i and Σ_i try to capture the correlation in the observed responses y_i

  ▷ this model does not assume conditional independence

- Choice between the two approaches is to a large extent philosophical

  ▷ *Random Effects*: trajectory of a subject dictated by time-independent random effects $\Rightarrow$ the shape of the trajectory is an inherent characteristic of this subject

  ▷ *Serial Correlation*: attempts to more precisely capture features of the trajectory by allowing subject-specific trends to vary over time

Often in practice it is **not** possible to include both a serial correlation term and many random effects because of numerical problems

- Example: In the AIDS dataset we investigate the fit of a mixed model with exponential serial correlation and increasing number of random effects – in particular:

  ▷ Model I: random intercepts

  ▷ Model II: random intercepts & random slopes

  the fixed-effects part includes linear and quadratic slopes and their interaction with treatment

	Model I	Model II
Intercept	7.173	7.214
Time_{ij}	-0.247	-0.251
Time_{ij}^2	0.007	0.007
$\text{ddI}_i \times \text{Time}_{ij}$	0.186	0.154
$\text{ddI}_i \times \text{Time}_{ij}^2$	-0.013	-0.010

• We observe small differences in the estimated fixed effects

	Model I	Model II
ϕ	2.29	0.52
95% CI	(1.62; 3.23)	(0.08; 3.45)

- However, we observe a more profound effect in the estimated parameter of the exponential serial correlation structure

  ▷ as we include more random effects, less information is available for estimating the serial correlation structure – note length of 95% CIs

- *Numerical problems:*

  ▷ The model is fitted with the exponential serial correlation structure,

  ▷ but if you instead tried the Gaussian serial correlation structure, then the models do not appropriately converge (Hessian matrix of the MLEs is not positive-definite)

3.8 Time-Varying Covariates*

- Up to now we have only included in mixed models covariates, which were fixed from baseline (except of course the time variable)

- However, often we may also be interested in assessing how a longitudinal outcome is associated with a covariate whose value changes over time

  ▷ such covariates are called *time-varying covariates*

- Example: In the PBC dataset we are interested in the effect of prothrombin time on serum bilirubin – prothrombin time has also been collected longitudinally during follow-up

---

3.8 Time-Varying Covariates* (cont'd)

- The handling of time-varying covariates poses some *important challenges*:

  1. Not always the longitudinal outcome and the time-varying covariate are collected at the same time points

  2. The longitudinal outcome at a particular time point t may depend not only on the value of the covariate at the same time point but also at other time points

  3. There are two types of time-varying covariates, *endogenous* and *exogenous*

     ▷ a time-varying covariate is *exogenous* if its distribution at time t is conditionally independent of all preceding outcomes

     ▷ a time-varying covariate is *endogenous* if it is not exogenous

3.8 Time-Varying Covariates* (cont'd)

- The formal definitions of *exogenous* and *endogenous* time-varying covariates are:

$$p\{x_i(t) \mid \mathcal{H}_i^Y(t), \mathcal{H}_i^X(t)\} = p\{x_i(t) \mid \mathcal{H}_i^X(t)\}$$

$$p\{x_i(t) \mid \mathcal{H}_i^Y(t), \mathcal{H}_i^X(t)\} \neq p\{x_i(t) \mid \mathcal{H}_i^X(t)\}$$

where

- ▷ $\mathcal{H}_i^Y(t) = \{y_i(t_{i1}), \ldots, y_i(t_{ik}); t_{ik} \leq t\}$ denotes the set of longitudinal measurements up to time t

- ▷ $\mathcal{H}_i^X(t) = \{x_i(t_{i1}), \ldots, x_i(t_{ik}); t_{ik} \leq t\}$ denotes the set of covariate measurements up to time t

- These features complicate postulating an appropriate model with such covariates

- A procedure to follow when working with time-varying covariates

  ▷ Determine if the covariate is *endogenous* or *exogenous*

   * if it is exogenous, then
     - we can proceed by postulating a standard mixed (or marginal) model, and
     - the longitudinal outcome at time t can only be associated with past covariate measurements, i.e., $\mathcal{H}_i^X(t)$

   * if it is endogenous, then more complicated types of analysis are required (joint models or marginal structural models) that fall outside the scope of this course

▷ Next, determine how to link the time-varying covariate to the longitudinal outcome (association structure)

* the longitudinal outcome at t is associated to the covariate at which time points (the same, previous time points, etc.)

* <u>Note:</u> If the scientific interest is focused on a particular type of association structure but in reality the longitudinal outcome is differently associated to the time-varying covariate, then the estimated association of interest may be diluted (biased) unless a specific type of analysis is followed (a marginal model with independent error terms, i.e., linear regression and corrected standard errors using the sandwich estimator)

▷ Depending on the chosen association structure in the previous step, and if the time-varying covariate is not measured at the same time points as the longitudinal outcome, then a form of interpolation may be required

- Example: In the PBC dataset we are interested in the effect of prothrombin time on serum bilirubin

$$\log(\texttt{serBilir}_{ij}) = \beta_0 + \beta_1 N(\texttt{Time}_{ij})_1 + \beta_2 N(\texttt{Time}_{ij})_2 + \beta_3 \texttt{Female}_i + \beta_4 \texttt{Age}_i +$$
$$\beta_5 \texttt{Prothr}_{ij} + b_{i0} + b_{i1} N(\texttt{Time}_{ij})_1 + b_{i2} N(\texttt{Time}_{ij})_2 + \varepsilon_{ij}$$

the covariance matrix of the random effects is assumed to be diagonal

	Value	Std.Err.	t-value	p-value
β_0	0.347	0.366	0.948	0.343
β_1	1.772	0.139	12.738	< 0.001
β_2	1.266	0.197	6.422	< 0.001
β_3	-0.233	0.184	-1.263	0.207
β_4	-0.000	0.006	-0.080	0.936
β_5	0.036	0.008	4.675	< 0.001

- Log serum bilirubin at time t is strongly related with the prothrombin time at the same time point – a unit increase of prothrombin time at follow-up time t increases the expected log serum bilirubin at the same follow-up time by 0.036

- We continue on the same example, but now we allow the log serum bilirubin at time t to be associated with the prothrombin time at previous time points as well – in particular:

$$\log(\texttt{serBilir}_{ij}) = \beta_0 + \beta_1 N(\texttt{Time}_{ij})_1 + \beta_2 N(\texttt{Time}_{ij})_2 + \beta_3 \texttt{Female}_i + \beta_4 \texttt{Age}_i +$$
$$\beta_5 \texttt{CumProthr}_{ij} + b_{i0} + b_{i1} N(\texttt{Time}_{ij})_1 + b_{i2} N(\texttt{Time}_{ij})_2 + \varepsilon_{ij}$$
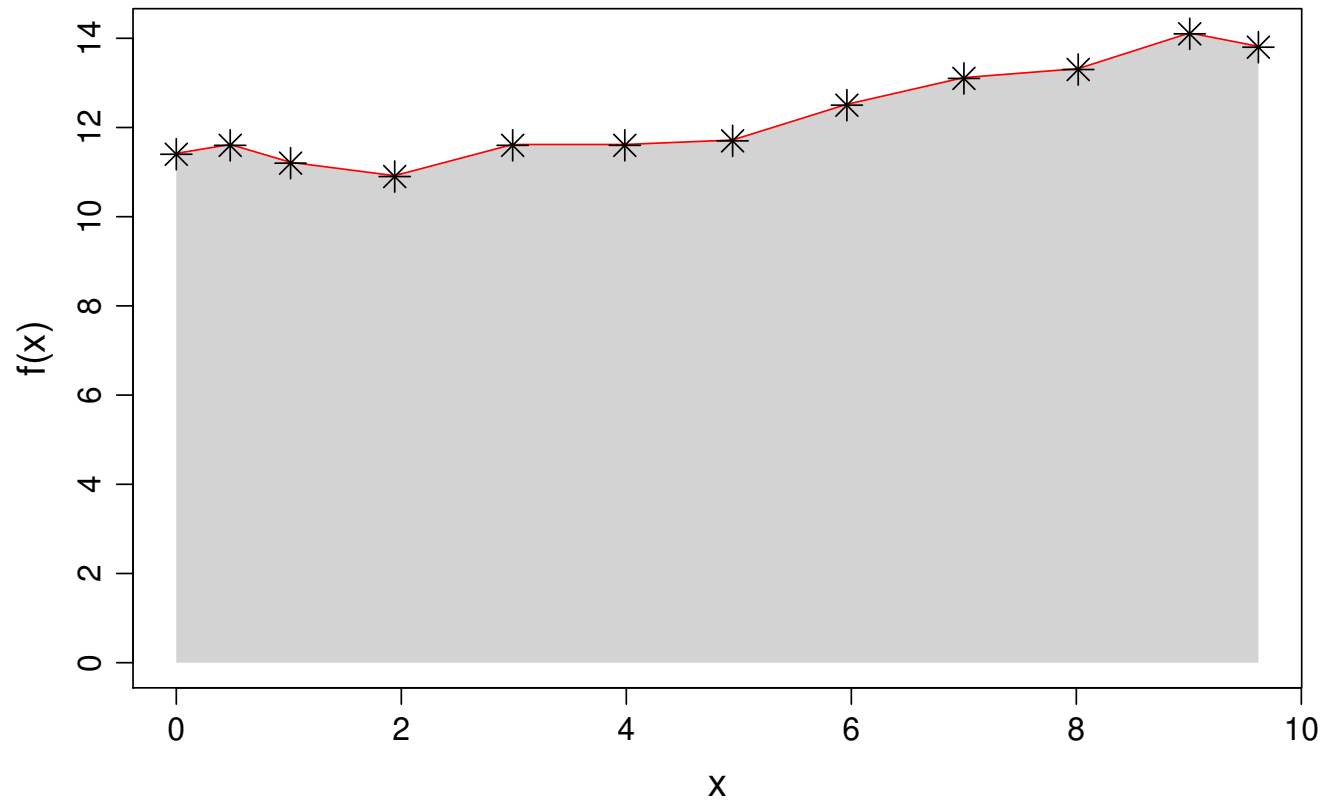
the covariance matrix of the random effects is assumed to be diagonal

- $\texttt{CumProthr}_{ij}$ denotes the cumulative effect of prothrombin time

  ▷ for Patient 21 and at two different follow-up times this effect is:
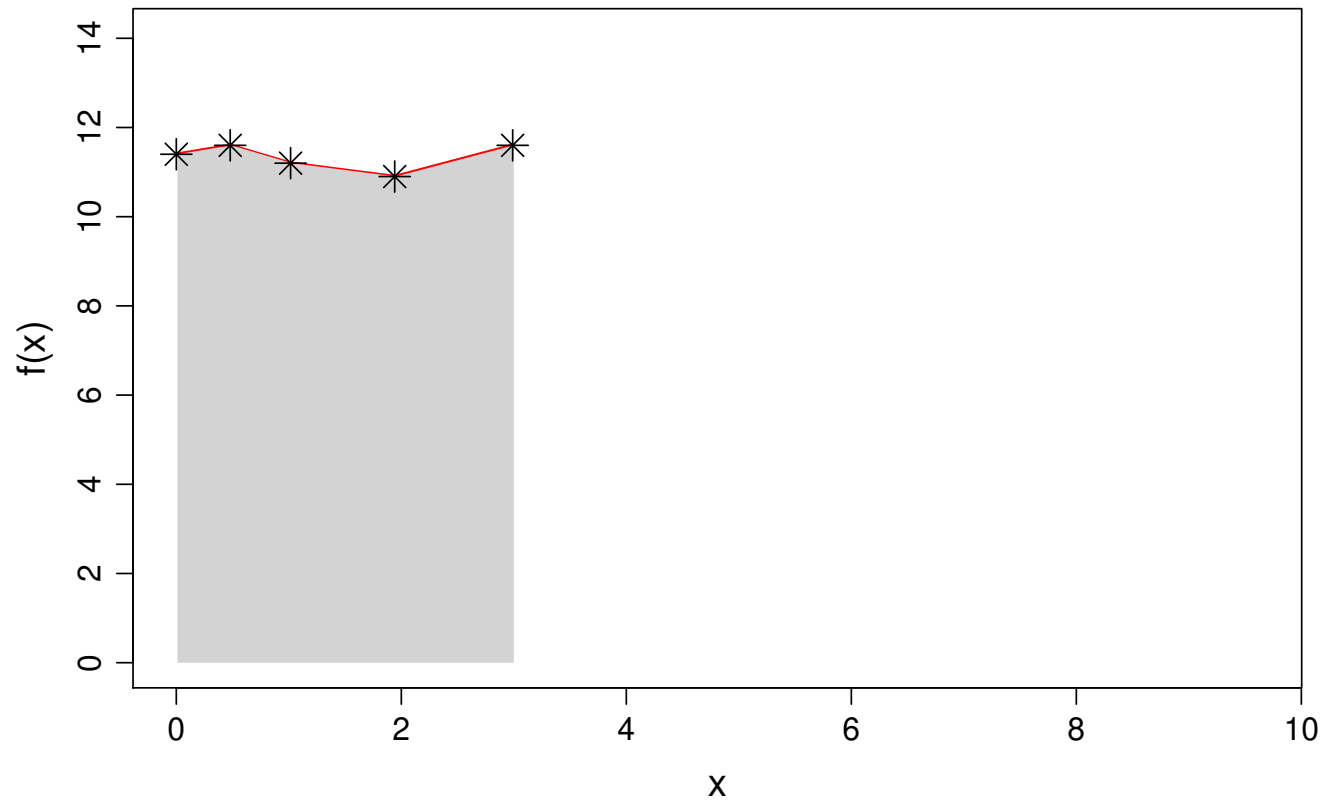
	Value	Std.Err.	t-value	p-value
β_0	0.728	0.361	2.015	0.044
β_1	1.700	0.159	10.686	< 0.001
β_2	1.256	0.210	5.974	< 0.001
β_3	-0.245	0.187	-1.311	0.191
β_4	-0.000	0.006	-0.007	0.995
β_5	0.009	0.004	2.462	0.014

- Log serum bilirubin at time t is strongly related with the cumulative prothrombin time up to the same time point – a unit increase of the cumulative prothrombin time up to follow-up time t increases the expected log serum bilirubin at the same follow-up time by 0.009

3.9 Model Building

* Mixed models consist of two parts, namely

  ▷ *fixed effects* that describe how specific covariates influence the average longitudinal evolutions

  ▷ *random effects* that describe how specific regression coefficients deviate from the overall mean described by the fixed effects

     * the random effects also model the correlations in the repeated measurements

* Interest can either be

  ▷ on the fixed-effects part alone (e.g., does treatment influence the average evolutions) or

  ▷ on both parts (e.g., to obtain subject specific predictions)

3.9 Model Building (cont'd)

- The general model building strategy we have seen in the previous chapter for marginal models also applies in the case of mixed models – more specifically:

  1. Put all the covariates of interest in the fixed-effects part, considering possible nonlinear terms and/or interactions between them – **do NOT** remove the ones that are not significant

  2. Then select an appropriate random-effects structure that adequately describes the correlations in the repeated measurements
     * typically we start from random intercepts and include each time an additional random effect term to see if we improve the fit (i.e., random slopes, quadratic random slopes, etc.)
     * 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 nonlinear & interaction terms

3.10 Hypothesis Testing

- Similarly to the marginal models of Chapter 2, in mixed models 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 again split the types of hypothesis tests in two parts:

  ▷ first, describe how can we choose the appropriate covariance matrix, and

  ▷ second, focus on hypothesis testing for the mean part of the model

- **Hypothesis testing for** $V_i = Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i}$: Assuming the same mean structure, we can fit a series of mixed models and choose the one that best describes the covariances

- In general, we distinguish between two cases

  ▷ comparing two mixed models with *nested* covariance matrices

  ▷ comparing two mixed 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

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

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

where

> ▷ $\ell(\hat{\theta}_0)$ the value of the log-likelihood function under the null hypothesis, i.e., the special case model

> ▷ $\ell(\hat{\theta}_a)$ the value of the log-likelihood function under the alternative hypothesis, i.e., the general model

> ▷ p denotes the number of parameters being tested

- **<u>Note:</u>** Provided that the mean structure in the two models is the same, we can either compare the REML of ML likelihoods of the models (preferable is REML)

Though, there is a **technical** complication when we compare nested mixed models for which one model has more random effects than the other

- To illustrate the issue, consider the hypothesis test between the random intercepts and the random intercepts & random slopes models

  ▷ random intercepts model

$$y_{ij} = X\beta + b_{i0} + \varepsilon_{ij}, \qquad b_{i0} \sim \mathcal{N}(0, \sigma_{b_0}^2)$$

  ▷ random intercepts & random slopes model

$$y_{ij} = X\beta + b_{i0} + b_{i1}t + \varepsilon_{ij}, \qquad b_i = [b_{i0}, b_{i1}] \sim \mathcal{N}(0, D)$$

  with

$$D = \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_{01}} \\ \sigma_{b_{01}} & \sigma_{b_1}^2 \end{bmatrix}$$

- Hence, the hypotheses to be tested are

$$H_0: \quad \sigma_{b_1}^2 = \sigma_{b_{01}} = 0$$

$$H_a: \quad \sigma_{b_1}^2 \neq 0 \text{ or } \sigma_{b_{01}} \neq 0$$

- What is the problem? The null hypothesis for $\sigma_{b_1}^2$ is on the boundary of its corresponding parameter space

  ▷ statistical tests derived from standard ML theory assume the H_0 is an interior point of the parameter space

  ▷ **the classical asymptotic χ^2 distribution for the likelihood ratio test statistic does not apply**

- For simple settings (as the one above), it has been proposed to use a mixture of χ^2 distributions to derive p-values, namely

  ▷ 50% from the χ^2 distribution with degrees of freedom the number of parameters being tested, and

  ▷ 50% from the χ^2 distribution with degrees of freedom the number of parameters which are not on the boundary under H_0

- Nonetheless, it has been suggested that this solution does not always work satisfactorily

  ▷ e.g., see package **RLRsim** in R and the references therein

- **Example:** In the AIDS dataset we compare two mixed models with linear and quadratic slopes in the fixed effects, and in the random effects

  ▷ M_1 : random intercepts & linear random slopes

  ▷ M_2 : random intercepts, linear random slopes & quadratic random slopes

- Hence, the covariance matrices of the random effects under the two models are

$$M_1 : D = \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_{01}} \\ \sigma_{b_{01}} & \sigma_{b_1}^2 \end{bmatrix} \quad \text{and} \quad M_2 : D = \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_{01}} & \sigma_{b_{02}} \\ \sigma_{b_{01}} & \sigma_{b_1}^2 & \sigma_{b_{12}} \\ \sigma_{b_{02}} & \sigma_{b_{12}} & \sigma_{b_2}^2 \end{bmatrix}$$

• And, the hypotheses being tested are

$$H_0 : \quad \sigma_{b_2}^2 = \sigma_{b_{02}} = \sigma_{b_{12}} = 0$$

$$H_a : \quad \text{at least one different from zero}$$

• The likelihood ratio test gives:

	df	logLik	LRT	p-value	Mixture p-value
M_1	9	-3573.88			
M_2	12	-3570.71	6.34	0.0961	0.0690

- About the two p-values

  ▷ The first p-value is based on the classic χ^2 distribution with degrees of freedom the number of parameters being tested, i.e., in this case 3

  ▷ The second p-value is based on the mixture of χ^2 distributions with 3 degrees of freedom (i.e., the classic one) and 2 degrees of freedom (the number of parameters not on the boundary under H_0), respectively

- We observe that the classic p-value is more conservative

  ▷ as we have seen in the previous section (see pp.216), when choosing the appropriate random effects we should be more liberal, and hence the mixture of χ^2 distribution is to be preferred

- When we have **non-nested** models we **cannot** use standard tests anymore

  ▷ the alternative in this case is to use the information criteria AIC or BIC

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

- $\boxed{\text{Example:}}$ In the PBC dataset we want to compare two mixed models with a spline effect of time and its interaction with sex in the fixed effects, and in the random effects

  ▷ M_1 : random intercepts & linear random slopes, with an unstructured matrix for these random effects

  ▷ M_2 : random intercepts, & nonlinear random slopes with splines, with a diagonal matrix for these random effects

  In the fixed-effects part and in the random-effects part of model M_2 : the splines are natural cubic splines with 2 internal knots

- These models are not nested and hence to compare them we use the AIC and BIC values

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

	df	logLik	AIC	BIC
M_1	10	-1522.38	3064.75	3120.45
M_2	10	-1438.53	2897.06	2952.76

  ▷ Both AIC and BIC suggest that the model with the nonlinear random slopes is better than the model with the linear random slopes

- <u>Notes:</u> Hypothesis testing for the covariance matrix V_i

  ▷ The aforementioned procedures assume that the fixed-effects structure of the mixed models to be compared are the same
    * under this assumption we can compare mixed models fitted with the restricted maximum likelihood (REML) method
    * otherwise the models should be fitted with maximum likelihood (ML)

  ▷ 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

3.10 Hypothesis Testing (cont'd)

- **Hypothesis testing for the regression coefficients** β: 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

- Everything works in the same manner as we have seen for marginal models in Chapter 2 (see pp.108–111)

  ▷ hence, we are not going to repeat the details here

- ☐ Example: We have fitted the following model to the Prothro dataset:

$$
\left\{
\begin{aligned}
\texttt{pro}_{ij} ={} & (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})N(\texttt{Time}_{ij})_1 + (\beta_2 + b_{i2})N(\texttt{Time}_{ij})_2 + \\
& (\beta_3 + b_{i3})N(\texttt{Time}_{ij})_3 + \beta_4\texttt{predn}_i + \\
& \beta_5\{\texttt{predn}_i \times N(\texttt{Time}_{ij})_1\} + \beta_6\{\texttt{predn}_i \times N(\texttt{Time}_{ij})_2\} + \\
& \beta_7\{\texttt{predn}_i \times N(\texttt{Time}_{ij})_3\} + \varepsilon_{ij} \\
\\
& b_i \sim \mathcal{N}(0, D) \qquad D \text{ is a diagonal matrix}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)
\end{aligned}
\right.
$$

▷ The terms $N(\texttt{Time}_{ij})_1$, $N(\texttt{Time}_{ij})_2$ and $N(\texttt{Time}_{ij})_3$ denote the basis for a natural cubic spline with three degrees of freedom to model possible nonlinearities in the time effect

- We are interested in

  ▷ the main effect of treatment,

  ▷ the overall effect of time, and

  ▷ the overall effect of treatment (i.e., main effect + interactions)

- Under the postulated model the main effect of treatment is given by parameter β_4, i.e.,

$$
\begin{aligned}
H_0 : \quad & \beta_4 = 0 \\
H_a : \quad & \beta_4 \neq 0
\end{aligned}
$$

- The output of the model gives: . . .

	Value	Std.Err.	t-value	p-value
β_0	72.357	1.435	50.423	< 0.001
β_1	-12.131	3.953	-3.069	0.002
β_2	31.954	3.445	9.274	< 0.001
β_3	34.015	4.706	7.228	< 0.001
β_4	-4.154	2.057	-2.019	0.044
β_5	14.621	5.679	2.575	0.010
β_6	-7.809	5.040	-1.549	0.121
β_7	-3.253	7.177	-0.453	0.650

- Hence, a significant treatment effect at baseline (strange!)

  ▷ the t-value in the output is the estimated coefficient divided by its standard error

- For the overall effect of time, we are interested in the hypothesis:

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_a: \quad \text{at least one coefficient different from 0}$$

- To test this hypothesis we can use an F-test but appropriately constructing the contrasts matrix

- We obtain

F-value	df_1	df_2	p-value
23.555	6	1939	< 0.0001

▷ Hence, a significant overall time effect

- For the overall treatment effect, we obtain the hypothesis:

$$H_0: \quad \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_a: \quad \text{at least one coefficient different from 0}$$

- This cannot be tested with an F-test because of technical reasons

  ▷ the denominator degrees of freedom are not the same for the main effect and the terms involving time

- As an alternative we can use the likelihood ratio test

  ▷ i.e., we compare the model we fitted with the model that only has the nonlinear effect of time in the fixed effects

• The likelihood ratio test gives

	df	logLik	AIC	BIC	LRT	p-value
without Treatment	9	-13240.53	26499.06	26553.02		
with Treatment	13	-13229.80	26485.59	26563.53	21.47	0.0003

▷ Hence, we obtain a significant overall treatment effect

3.11 Residuals

- As we have similarly done for marginal models in Chapter 2, before extracting conclusions from mixed models, we will first need to validate the underlying assumptions they make

- To do this we can use the residuals of the model

- In the setting of mixed models we have two types of residuals

  ▷ *Marginal residuals:* These are based on the implied marginal model behind a linear mixed model (see pp.160)

  ▷ *Conditional residuals:* These are based on the hierarchical representation of the mixed model and utilize the empirical Bayes estimates of the random effects (see pp.174)

- The exact definitions are as follows:

  ▷ *Marginal residuals:*

$$
\begin{cases}
y_i & = X_i\beta + \varepsilon_i^*, \quad \varepsilon_i^* \sim \mathcal{N}(0, Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i}) \\
\\
r_i^{marg} & = y_i - X_i\widehat{\beta}
\end{cases}
$$

  ▷ These residuals predict the marginal errors ε_i^*

  ▷ They can be used to

  - investigate misspecification of the mean structure $X_i\beta$
  - validate the assumptions for the within-subjects covariance structure $Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i}$

▷ *Conditional residuals*

$$
\begin{cases}
y_i &= X_i\beta + Z_i b_i + \varepsilon_i, \quad b_i \sim \mathcal{N}(0, D), \;\; \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_{n_i}) \\
\\
r_i^{cond} &= y_i - X_i\widehat{\beta} - Z_i\widehat{b_i}
\end{cases}
$$

▷ These residuals predict the conditional errors ε_i

▷ They can be used to

    * investigate misspecification of the hierarchical mean structure $X_i\beta + Z_i b_i$

    * validate the assumptions for the within-subjects variance structure σ^2

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

$$
\begin{cases}
\texttt{pro}_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})N(\texttt{Time}_{ij})_1 + (\beta_2 + b_{i2})N(\texttt{Time}_{ij})_2 + \\
\qquad (\beta_3 + b_{i3})N(\texttt{Time}_{ij})_3 + \beta_4\texttt{predn}_i + \beta_5\{\texttt{predn}_i \times N(\texttt{Time}_{ij})_1\} + \\
\qquad \beta_6\{\texttt{predn}_i \times N(\texttt{Time}_{ij})_2\} + \beta_7\{\texttt{predn}_i \times N(\texttt{Time}_{ij})_3\} + \varepsilon_{ij}, \\
\\
b_i \sim \mathcal{N}(0, \mathsf{diag}\{D\}), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)
\end{cases}
$$

$N(\cdot)$ denotes a natural cubic spline basis

by plotting

▷ the standardized marginal residuals versus fitted values per treatment group

▷ the standardized conditional residuals versus fitted values per treatment group

▷ QQ-plot of the standardized marginal residuals
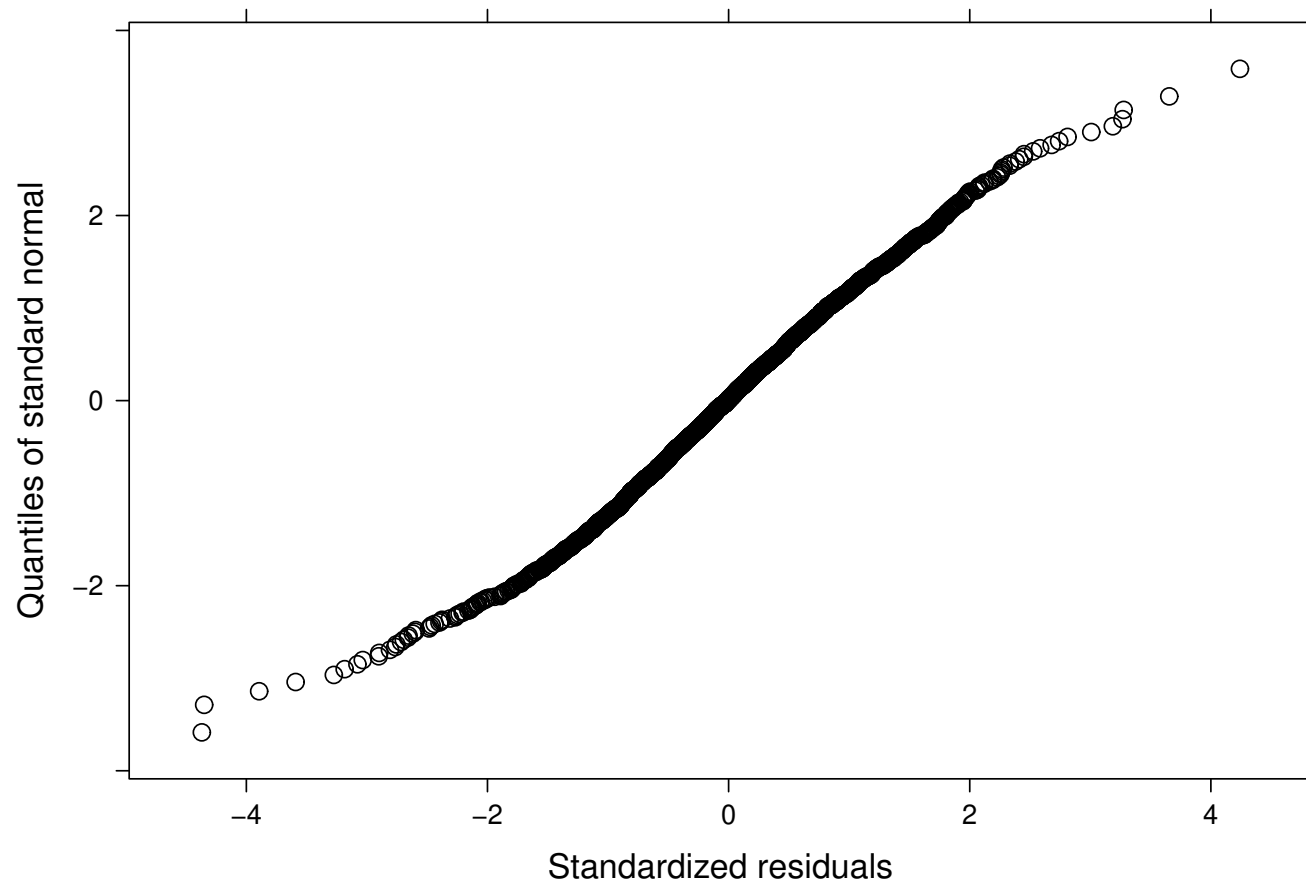
▷ QQ-plot of the standardized conditional residuals
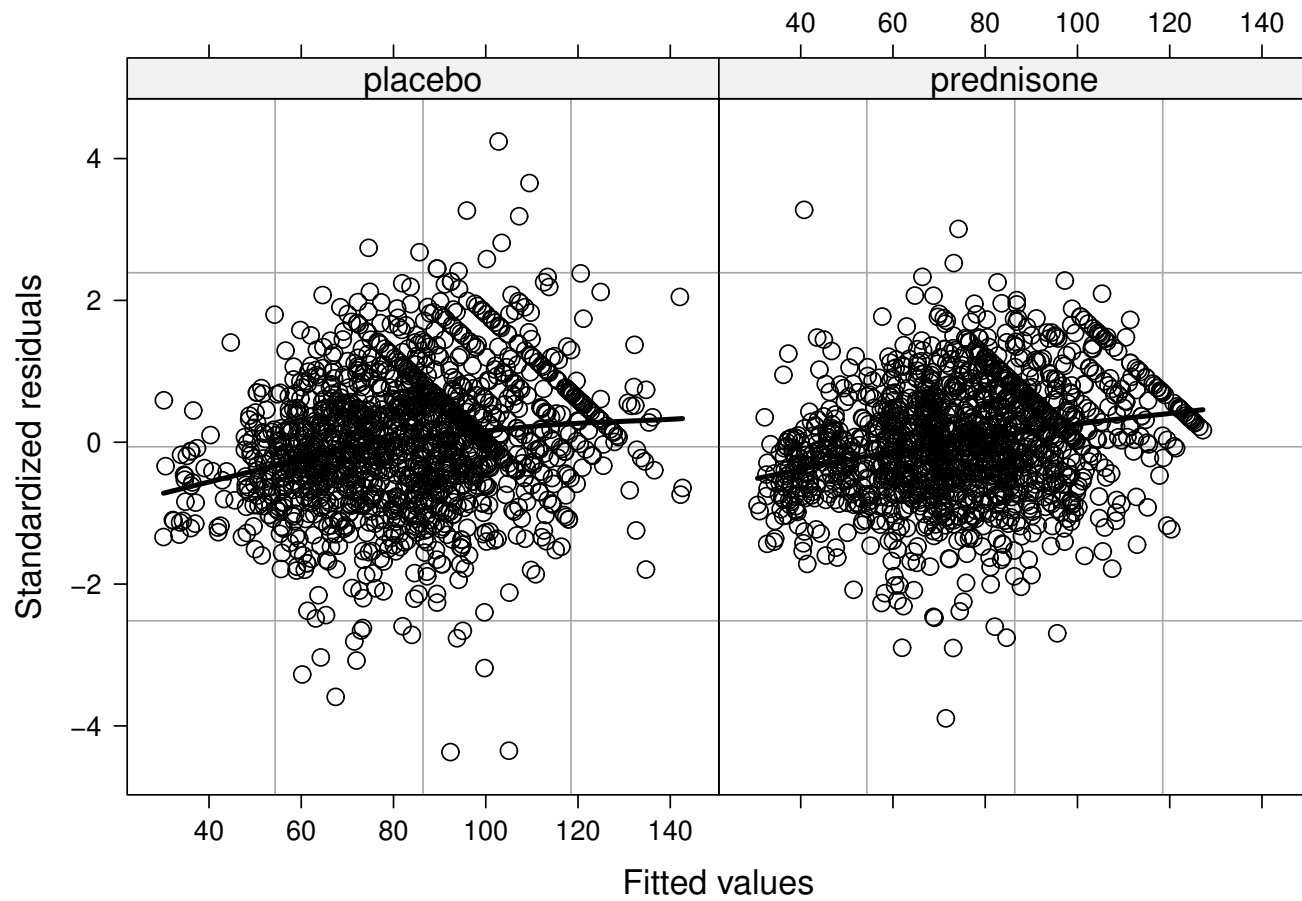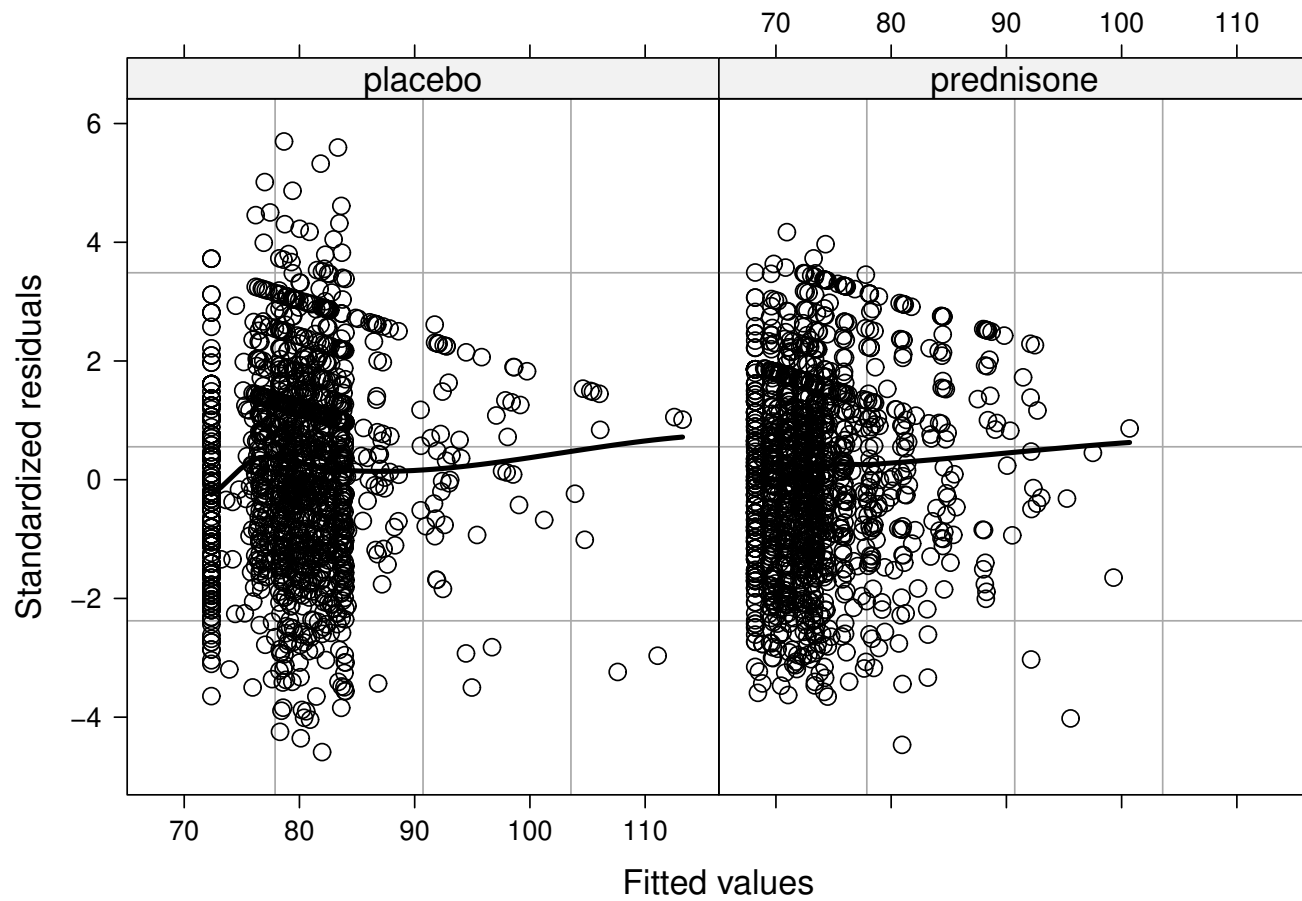
- Observations

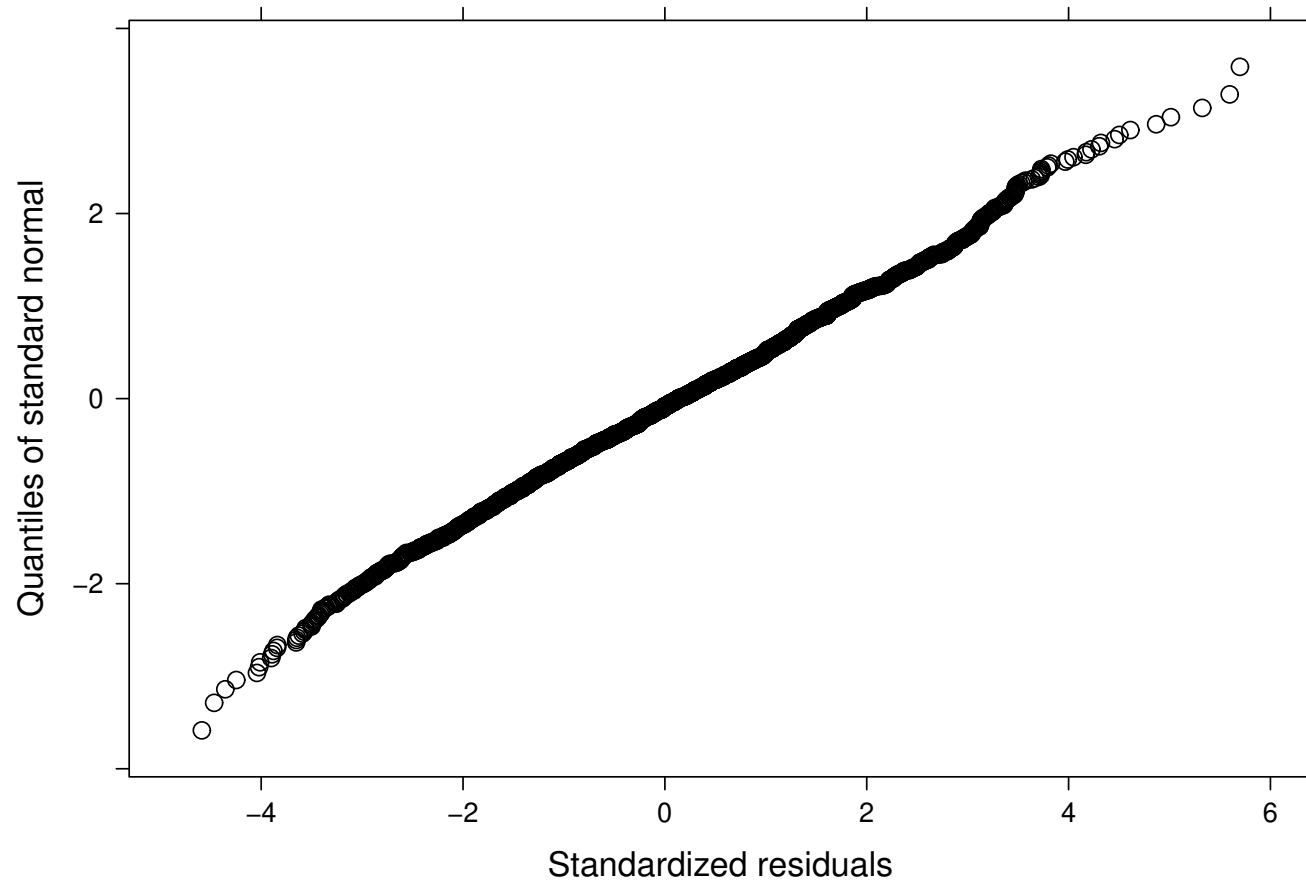 ▷ the plots of the residuals versus the fitted values do show a slightly systematic behavior

  ▷ the QQ-plots do not show big discrepancies from normality

3.12 Review of Key Points

- Mixed effects models constitute an alternative modeling framework for analyzing grouped/clustered data

  ▷ basic idea: sample units in the same group/cluster share the same random effects

  ▷ the random effects are *unobserved* variables that induce correlation

- From a practical viewpoint, mixed models provide a more flexible framework to model correlations when

  ▷ we have unbalanced data and/or

  ▷ the correlation structure has a complicated form (e.g., multilevel designs)

- The random effects can be estimated using empirical Bayes methodology

  ▷ mixed models provide subject-specific predictions that are more accurate than marginal predictions

- Mixed models can be extended to include correlated error terms

  ▷ this is in the same spirit as the marginal models of Chapter 2

  ▷ however, this extension often makes the model computationally unstable

- 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

  ▷ marginal and conditional residuals available

Chapter 4

Marginal Models for Discrete Data

4.1 Review of Generalized Linear Models

- So far we have concentrated on continuous/normal repeated measurements data

  ▷ serum bilirubin and serum cholesterol in the PBC dataset

  ▷ CD4 cell counts in the AIDS dataset

  ▷ prothrombin time in the Prothro dataset

However often we may want to analyze other types of repeatedly measured outcomes that are not normally distributed
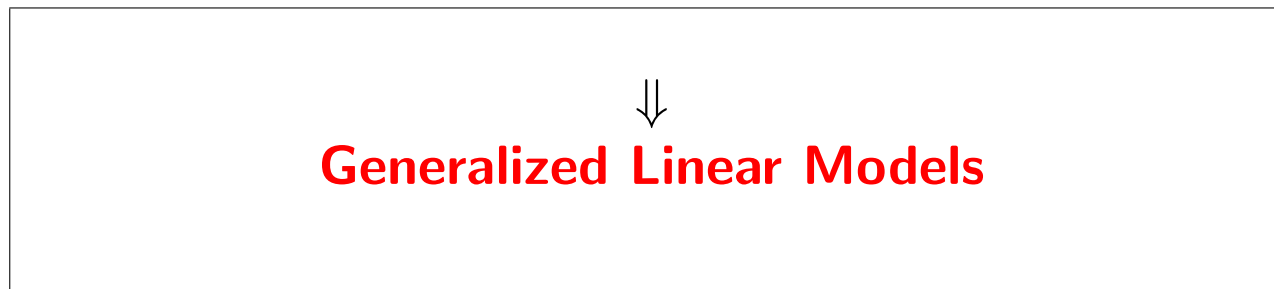
4.1 Review of Generalized Linear Models (cont'd)

- Examples:

  ▷ in colon cancer studies the iFOBT test (presence or not of blood in stool) is used to monitor patients ⇒ *dichotomous data*

  ▷ after a heart transplantation patients report their quality of life in frequent intervals, with the categories 'Poor', 'Moderate', 'Good' and 'Very Good' ⇒ *ordinal data*

  ▷ in asthma studies often interest is in the number of asthma attacks patients have in a period of time ⇒ *count data*

  ▷ . . .

4.1 Review of Generalized Linear Models (cont'd)

- Suppose we have a dichotomous outcome Y measured *cross-sectionally*

  ▷ ⬚Example:⬚ The serum cholesterol levels from the PBC dataset at baseline (i.e., time $t = 0$) that are higher than 210 mg/dL

- 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 cholesterol levels higher than 210 mg/dL?

$$\Downarrow$$

Generalized Linear Models

- Reminder: The General Linear Model for continuous data

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

- There are two issues with applying this model for the dichotomous outcome
  $y_i = I(\texttt{serChol}_i > 210)$ ($I(\cdot)$ is the indicator function: $I(A) = 1$ if A is true & $I(A) = 0$ when A is false)

  1. y_i does not follow a normal distribution, it follows a *Bernoulli* distribution

  2. the mean of the Bernoulli distribution is π_i, *the probability*, of having serum cholesterol higher than the threshold – the mean in the linear regression model is
     $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$

4.1 Review of Generalized Linear Models (cont'd)

- Say we use a naive strategy and consider a linear regression model for $y_i = I(\texttt{serChol}_i > 210)$ – the mean would be

$$\pi_i = E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

- What is the problem?

  ▷ π is a probability and is restricted to values between 0 and 1

  ▷ $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ can take values below 0 and above 1, i.e., in $(-\infty, +\infty)$

- We have to bring π in the scale $(-\infty, +\infty) \rightarrow$ a classic option is:

$$0 < \qquad \pi \qquad < 1$$

$$0 < \quad \frac{\pi}{1 - \pi} \quad < +\infty$$

$$-\infty < \log \frac{\pi}{1 - \pi} < +\infty$$

• This gives rise to the *logistic regression model*

$$\text{log odds of success} = \log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

$$\text{odds of success} = \frac{\pi_i}{1 - \pi_i} = \exp\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)$$

$$\text{probability of success} = \pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

which respects the constraint that π takes values between 0 and 1

4.1 Review of Generalized Linear Models (cont'd)

- A unit change in X_1 from x to $x + 1$ (while all other covariates are held fixed) corresponds to

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x + \ldots + \beta_p x_{ip}$$

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 (x + 1) + \ldots + \beta_p x_{ip}$$

- Thus,

$$\beta_1 = \log \frac{\pi_i}{1 - \pi_i} - \log \frac{\pi_i}{1 - \pi_i} = \log \left\{ \frac{\pi_i}{1 - \pi_i} \middle/ \frac{\pi_i}{1 - \pi_i} \right\}$$

$$\exp(\beta_1) = \frac{\pi_i}{1 - \pi_i} \middle/ \frac{\pi_i}{1 - \pi_i}$$

- <u>Notes:</u>

  ▷ The relationship between log odds of success and the covariates is linear

  ▷ The relationship between π and the covariates is non-linear

  $\Rightarrow$ Interpretation of parameters is different than in linear regression models

- For dichotomous X_1

    $\Rightarrow \beta_1$ is the log odds ratio of 'success' between the two levels of X_1 given that all other covariates remain constant

    $\Rightarrow \exp(\beta_1)$ is the odds ratio between the two levels of X_1 given that all other covariates remain constant

- For continuous X_1

    $\Rightarrow \beta_1$ is change in log odds of 'success' for a unit change in X_1 given that all other covariates remain constant

    $\Rightarrow \exp(\beta_1)$ is the odds ratio for a unit change in X_1 given that all other covariates remain constant

* Relationships

$$\exp(\beta_1) = \begin{cases} = 1, & \text{the two odds are the same} \\ > 1, & \text{increased odds of success} \\ < 1, & \text{decreased odds of success} \end{cases}$$

* As π increases

  ▷ odds of success increases

  ▷ log odds of success increases

- Example: In the PBC dataset we are interested in investigating how the probability of serum cholesterol higher than 210 mg/dL is associated with age, sex and the treatment the patients received – the model has the form

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 \texttt{Age}_i + \beta_2 \texttt{Female}_i + \beta_3 \texttt{D-penicil}_i$$

where $\pi_i = \Pr(\texttt{serChol}_i > 210)$

	Value	Std.Err.	z-value	p-value
β_0	3.739	1.306	2.862	0.004
β_1	-0.038	0.020	-1.879	0.060
β_2	0.313	0.563	0.556	0.578
β_3	0.512	0.420	1.218	0.223

▷ $\beta_0 = 3.7$ is the log odds of excess levels of serum cholesterol for a male patient in the control group who is 0 years old

▷ $\beta_1 = -0.04$ is the log odds ratio for a unit increase in age for patients of the same sex who receive the same treatment

▷ $\beta_2 = 0.3$ is the log odds ratio of females versus males of the same age who receive the same treatment

4.2 Generalized Estimating Equations

• We return our focus on repeated measurements data, namely, repeated categorical data

   ▷ **we need to account for the correlations**

• <u>Reminder:</u> In the marginal models for continuous multivariate data (Chapter 2) we took account of the correlations by *incorporating a correlation matrix in the error terms*

• For categorical data it is not straightforward to do that because there are no clear multivariate analogues of the univariate distributions

   ▷ we will do something similar, **not** in the *error terms* but in the *score equations*

4.2 Generalized Estimating Equations (cont'd)

- Liang and Zeger (1986, Biometrika) made the following important contribution

  ▷ The parameters of Generalized Linear models are estimated using the *maximum likelihood* approach

  ▷ *Key idea:* Finding the top of the log-likelihood mountain is equivalent to finding the parameter values for which the slope of the mountain is flat (i.e., zero)

  ▷ The slope of the log-likelihood mountain is given by the score vector

$$S(\beta) = \sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} (Y_i - \mu_i)$$

where in standard logistic regression

$\triangleright$ μ_i the mean of Y_i, e.g., for dichotomous data $\mu_i = \pi_i$

$\triangleright$ V_i is a diagonal matrix with the variance of Y_i, e.g., for dichotomous data $V_i = \text{diag}\{\pi_i(1 - \pi_i)\}$

• The idea of Liang and Zeger was to replace the diagonal matrix V_i with a full covariance matrix

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

where

$\triangleright$ $A_i = \text{diag}\{\text{var}(Y_i)^{1/2}\}$ a diagonal matrix with the standard deviations

$\triangleright$ $R_i(\alpha)$ a 'working' assumption for the pairwise correlations

4.2 Generalized Estimating Equations (cont'd)

- If the assumed mean structure μ_i is correctly specified, then

$$\widehat{\beta} \sim \mathcal{N}\{\beta, \mathsf{var}(\widehat{\beta})\}$$

where

$$\boxed{\mathsf{var}(\widehat{\beta}) = V_0^{-1} V_1 V_0^{-1}} \quad \text{is called the } \textbf{Sandwich} \text{ or } \textbf{Robust} \text{ estimator}$$

with

$$\underbrace{V_0 = \sum_i \frac{\partial \mu_i}{\partial \beta^\top} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}}_{\textbf{bread}} \quad \text{and} \quad \underbrace{V_1 = \sum_i \frac{\partial \mu_i}{\partial \beta^\top} V_i^{-1} \mathsf{var}(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta}}_{\textbf{meat}}$$

- **Sandwich/Robust** vs **Naive/Model Based** standard errors

  ▷ software often also report the **Naive/Model Based** standard errors

  ▷ these standard errors assume that the working correlation matrix is correctly specified

  ▷ the **Sandwich/Robust** corrects for a possible misspecification of the correlation structure

  * though at the expense of power

GEE is not a likelihood-based approach (i.e., a model)

$$\Downarrow$$

it is an estimation method

- No assumptions for the joint distribution of repeated measurements
  $\Rightarrow$ *Semi-parametric approach*

- Three components

  1. Model for mean response $E(Y_i) = \mu_i$, e.g., binary data $E(Y_i) = \pi_i$

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

2. Variance of $Y_i \Rightarrow$ follows from GLM assumption for each measurement, e.g., binary data

$$\text{var}(Y_i) = \phi \pi_i (1 - \pi_i)$$

with ϕ a scale parameter that models over-dispersion

3. Pairwise correlations $\Rightarrow$ we make a "working" assumption that possibly depends on parameters to be estimated

The mean and the correlations are **separately** defined!
This is in contrast to the GLMMs we will see in the next chapter

<div style="border:1px solid">

Interest is primarily in the βs, the covariance structure is considered as "nuisance"

$$\Downarrow$$

Assumptions for the variance and correlation are <u>not</u> supposed to be correct

</div>

- This has implications for

  ▷ Hypothesis testing

   ⇒ Likelihood ratio test or score test not applicable

   ⇒ The Wald test can be used

  ▷ Performance when we have missing data

- Interpretation of β is the same as in classic GLMs

    ▷ β_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 PBC dataset we are interested in the effect of treatment on the average longitudinal evolutions for the probability of serum cholesterol higher than 210 mg/dL – we fit the GEE model with

    ▷ different average longitudinal evolutions per treatment group ($X\beta$ part)

    ▷ exchangeable working correlation matrix

- The model has the form:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \texttt{Time}_{ij} + \beta_2 \texttt{D-penicil}_i + \beta_3 \{\texttt{Time}_{ij} \times \texttt{D-penicil}_i\}$$

	Value	Std.Err.	z-value	p-value
β_0	2.058	0.250	67.658	< 0.001
β_1	-0.114	0.044	6.798	0.009
β_2	0.166	0.341	0.237	0.626
β_3	0.003	0.054	0.004	0.949

- We found no indication of a difference between the two treatments groups and the odds of having excess cholesterol levels during follow-up

- Interpretation of parameters (note that we have an interaction term)

  ▷ $\exp(\beta_1) = 0.9$ is the odds ratio for a year increase for patients receiving placebo

  ▷ $\exp(\beta_2) = 1.2$ is the odds ratio of D-penicil to placebo at baseline

  ▷ $\exp(\beta_3) = 1.003$ is the relative difference between the odds ratios for a year increase in the two treatment groups

    * the odds ratio for a year increase in patients receiving D-penicil is 0.3% higher than the corresponding odds ratio of the placebo patients

- As we have previously seen, to effectively communicate complex models we can use effect plots

- Example: In the PBC dataset we are interested in the probability of having excess serum cholesterol levels

  ▷ we allow for nonlinear time and baseline age effects using natural cubic splines with 3 degrees of freedom
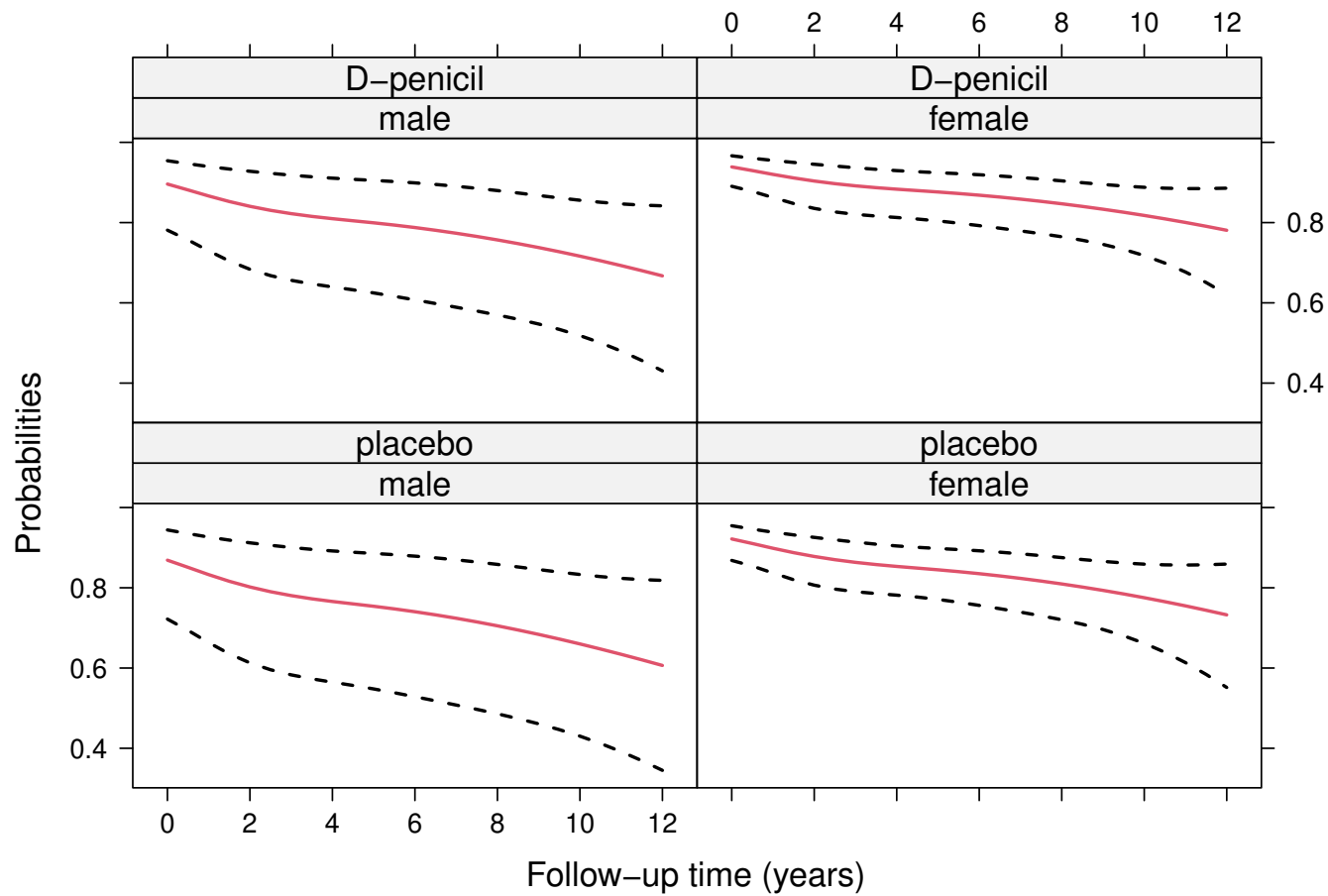
  ▷ we also correct for treatment and gender

- The following two figures depict the relationship between

    ▷ the log odds of excess serum cholesterol level

    ▷ the probability of excess serum cholesterol level

    for male and female patients receiving either the active drug or placebo, who are 49 years old (in the app different ages can be selected)
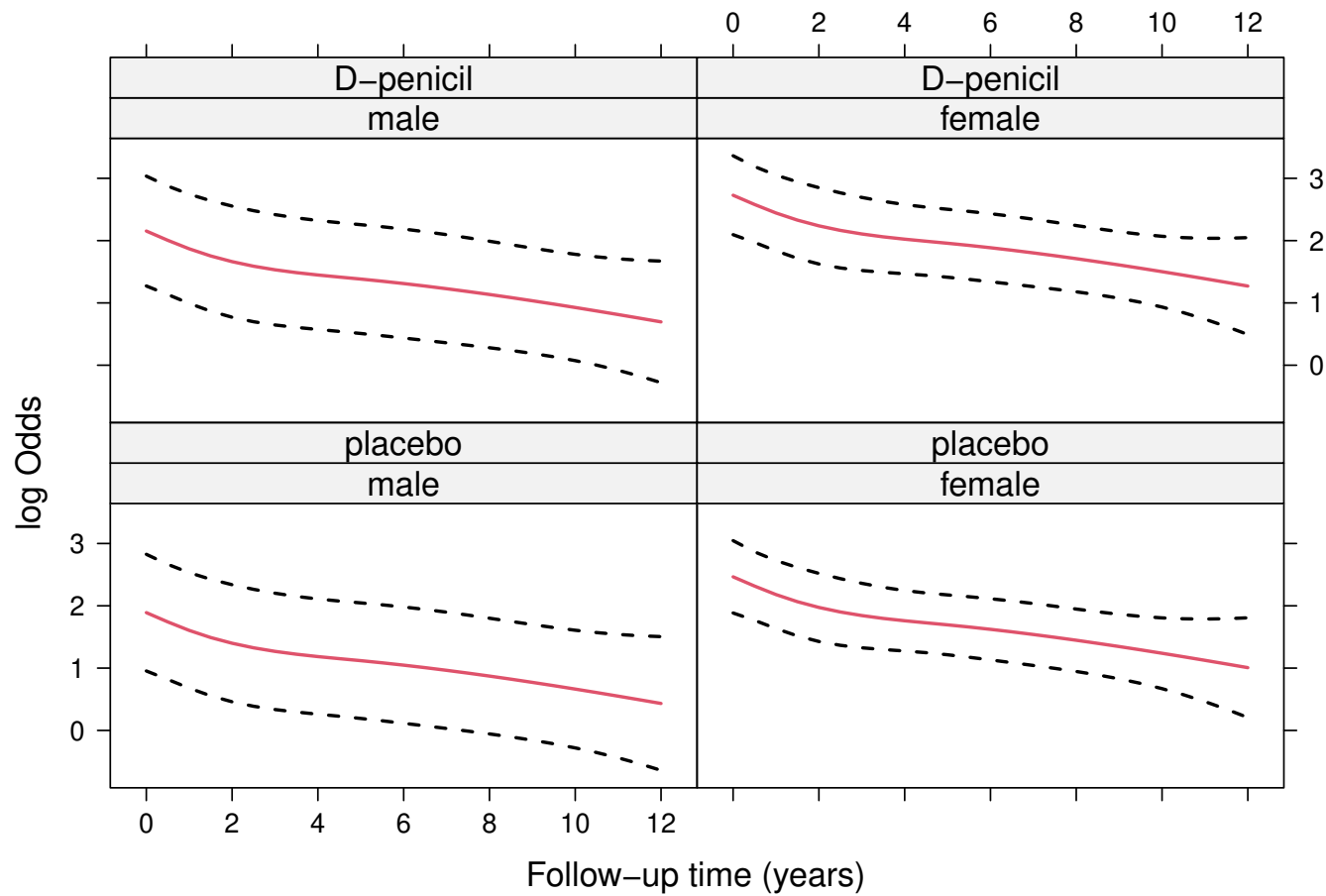
4.4 Generalized Estimating Equations in R

R> In R there are two main packages for GEE analysis, namely **gee** and **geepack** – in this course we will only use **geepack**

 ▷ The main function to fit GEEs is `geeglm()` – this has similar syntax as the `glm()` function of base R that fits GLMs

R> Main arguments of `geeglm()`

 ▷ `formula`: An R formula specifying the response variable and the predictors

 ▷ `family`: a description of the error distribution and link function to be used in the model

 ▷ `id`: the variable denoting which measurements belong to the same group (e.g., to the same subject)

R> Main arguments of `geeglm()`

▷ `data`: a data frame that contains all these variables; **important**: the rows of this data frame must be ordered with respect to `id`, and the rows within the same `id` should be ordered with respect to time

▷ `corstr`: the assumed working correlation matrix; options are `"independence"`, `"exchangeable"`, `"ar1"`, `"unstructured"`, and `"userdefined"`

4.4 Generalized Estimating Equations in R (cont'd)

R> The following code fits a GEE model for serum cholesterol from the PBC dataset with an exchangeable working correlation matrix

```
geeFit <- geeglm(serCholD ~ year * drug, family = binomial(),
                 data = pbc2, id = id, corstr = "exchangeable")

summary(geeFit)
```

4.5 Working Correlation Matrix

• As we have seen, the GEEs are a *semiparametric* approach

  ▷ using the *sandwich estimator* we obtain valid inferences *even if* the working correlation matrix is misspecified

• Hence, one could wonder: Why not always use the sandwich estimator in order not to have to care about the correlation structure?

  ▷ in other words, why do we need this course?

<div style="border:1px solid black; padding:1em; text-align:center; color:red; font-weight:bold;">
The sandwich estimator does have important limitations!
</div>

4.5 Working Correlation Matrix (cont'd)

- About the sandwich estimator:

  ▷ is based on asymptotic arguments $\Rightarrow$ it only works for big samples

  ▷ more specifically, it works better under the following settings
    * when the number of subjects is considerably larger than the number of repeated measurements,
    * balanced designs in which all subjects provide measurements at the same time points
    * we do not have too many covariates (especially continuous)

- **Therefore**, the choice of the working correlation matrix is of particular importance

  ▷ this is why all statistical software offer several alternative options for this matrix

• Some standard options are

▷ Independence

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

* repeated measurements are assumed uncorrelated
* even though an *unrealistic* assumption, in some circumstances it is an appropriate route to follow (see e.g., pp.209)

▷ Exchangeable

$$
\begin{bmatrix}
1 & \rho & \rho & \rho \\
\rho & 1 & \rho & \rho \\
\rho & \rho & 1 & \rho \\
\rho & \rho & \rho & 1
\end{bmatrix}
$$

* constant correlation over time

* more appropriate for short time intervals

▷ First-order autoregressive

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

* correlations decrease over time

* more appropriate for longer time intervals

4.5 Working Correlation Matrix (cont'd)

▷ General/Unstructured

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_4 & \rho_5 \\ \rho_2 & \rho_4 & 1 & \rho_6 \\ \rho_3 & \rho_5 & \rho_6 & 1 \end{bmatrix}$$

* the most flexible correlation structure
* it can only be fitted (i.e., without numerical problems) with balanced data and big sample sizes (rule of thumb: when $n \gg p(p-1)/2$, with n denoting the number of subjects and p the number of repeated measurements)

- Example: A very low CD4 count (less than 150 cells/mm^3) is an indicator for opportunistic infections

  ▷ In the following analysis we dichotomize the CD4 cell counts from the AIDS dataset using this threshold

  ▷ We fit GEEs for this dichotomous response and only the categorical version of the time as covariate – for the working correlation matrix we assume "Independence", "Exchangeable", "AR1" and "Unstructured"

  ▷ We compare parameter estimates and standard errors (model-based and sandwich)

- The model has the form:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1\{\texttt{Time}_{ij} = 2\} + \beta_2\{\texttt{Time}_{ij} = 6\} + \beta_3\{\texttt{Time}_{ij} = 12\} +$$
$$\beta_4\{\texttt{Time}_{ij} = 18\}$$

where

▷ $\pi_{ij} = \mathsf{Pr}(\texttt{CD4}_{ij} < 150)$

▷ $\{\texttt{Time}_{ij} = 2\}$ denotes the dummy variable for month 2, $\{\texttt{Time}_{ij} = 6\}$ the dummy variable for month 6, and so on

- The estimated regression coefficients under the four GEEs are:

	Ind.	Exch.	AR1	Unstr.
β_0	1.474	1.474	1.477	1.533
β_1	-0.043	0.013	0.029	0.026
β_2	0.151	0.295	0.332	0.291
β_3	-0.110	0.465	0.415	0.660
β_4	0.541	1.173	0.652	-0.240

▷ We observe considerable differences in the magnitude of some parameters

- The estimated standard errors under the four GEEs are:

	Sandwich				Model-based			
	Ind.	Exch.	AR1	Unstr.	Ind.	Exch.	AR1	Unstr.
$s.e.(\beta_0)$	0.119	0.119	0.118	0.139	0.119	0.127	0.126	0.129
$s.e.(\beta_1)$	0.106	0.100	0.102	0.105	0.178	0.099	0.077	0.123
$s.e.(\beta_2)$	0.140	0.139	0.138	0.146	0.194	0.114	0.113	0.139
$s.e.(\beta_3)$	0.146	0.178	0.174	0.243	0.204	0.136	0.148	0.074
$s.e.(\beta_4)$	0.524	0.627	0.315	1.031	0.545	0.423	0.343	NaN

▷ We also observe differences in the magnitudes of the estimated standard errors

▷ These are smaller in the sandwich than in the model-based estimator

- How to choose the most appropriate working correlation matrix?

  ▷ unfortunately, there are no generally accepted formal tests for choosing the working correlation matrix

  ▷ the choice **should not** be based on the grounds of statistical significance

  ▷ consider appropriate choices based on the features of the data
    * for balanced data with big sample size $\Rightarrow$ Unstructured
    * unbalanced data $\Rightarrow$ Exchangeable or AR1
    * when more than one options plausible $\Rightarrow$ sensitivity analysis, *report all results*

Similarly to what we have seen in Chapter 2 and 3, **a prerequisite is that the mean structure is correctly specified**

4.6 Hypothesis Testing

- Having fitted a GEE model often scientific interest lies in testing specific hypotheses

- Due to the fact that GEE models are semiparametric models, we can **only** employ Wald tests to test the hypothesis of interest

  ▷ score and likelihood ratio tests are not available

- In addition, in standard GEEs, these tests are only available for the mean parameters (i.e., the regression coefficients) and not the parameters of the working correlation matrix

- For individual parameters the Wald test has the form:

$$\widehat{\beta} \Big/ s.e.(\widehat{\beta}) \sim \mathcal{N}(0, 1)$$

  where

  ▷ $\widehat{\beta}$ denotes the estimated regression coefficient, and

  ▷ $s.e.(\widehat{\beta})$ the sandwich or model-based standard error of this regression coefficient

- We have seen an example of these tests in the analysis of excess cholesterol levels in the PBC dataset (see pp.276–277)

- When interest is more than one parameters, then we use the multivariate version of the Wald test, i.e.,

$$
\begin{aligned}
H_0: \quad & L\beta = 0 \\
H_a: \quad & L\beta \neq 0
\end{aligned}
$$

  where L is the contrasts matrix of interest

- Example: We extend the GEE model for the AIDS dataset we have seen in pp.292

  ▷ time is treated as categorical variable and we also take interactions with treatment

  ▷ we are interested in the overall treatment effect

  ▷ the working correlation matrix is assumed to have an AR1 structure

- The models under the null and alternative hypothesis have the form:

$$
\begin{cases}
H_0 : \log \dfrac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \{\texttt{Time}_{ij} = 2\} + \beta_2 \{\texttt{Time}_{ij} = 6\} + \\
\qquad\qquad \beta_3 \{\texttt{Time}_{ij} = 12\} + \beta_4 \{\texttt{Time}_{ij} = 18\} \\[2em]
H_a : \log \dfrac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \{\texttt{Time}_{ij} = 2\} + \beta_2 \{\texttt{Time}_{ij} = 6\} + \\
\qquad\qquad \beta_3 \{\texttt{Time}_{ij} = 12\} + \beta_4 \{\texttt{Time}_{ij} = 18\} + \beta_5 \texttt{ddI}_i + \\
\qquad\qquad \beta_6 \big[\texttt{ddI}_i \times \{\texttt{Time}_{ij} = 2\}\big] + \beta_7 \big[\texttt{ddI}_i \times \{\texttt{Time}_{ij} = 6\}\big] + \\
\qquad\qquad \beta_8 \big[\texttt{ddI}_i \times \{\texttt{Time}_{ij} = 12\}\big] + \beta_9 \big[\texttt{ddI}_i \times \{\texttt{Time}_{ij} = 18\}\big]
\end{cases}
$$

• Hence, the parameters we wish to test are:

$$H_0: \quad \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$
$$H_a: \quad \text{at least one coefficient is different from zero}$$

• The Wald test gives:

▷ W = 8.5

▷ df = 5

▷ p-value = 0.131

Hence, no evidence of a treatment effect

- Note that we could also apply a multivariate test not only with a contrast matrix L, but also with a design matrix of interest X

- Example: We fit a GEE model for excess serum cholesterol levels in which we include the following terms

  ▷ nonlinear effect of time with natural cubic splines with 3 degrees of freedom

  ▷ main effect of treatment

  ▷ interaction of treatment with nonlinear effect of time

  ▷ nonlinear effect of baseline age with natural cubic splines with 3 degrees of freedom

- As we have previously discussed, when we include nonlinear terms parameters do not have a straightforward interpretation

- But we could still test meaningful hypothesis

  ▷ more specifically, we are interested in testing for a treatment effect at year 7 for patients who are 49 years old

- Let X_1 denote the design matrix of the placebo patient at year 7 and age 49, and X_2 the analogous design matrix for the patient who receives D-penicillamine

- Hence, we want to test

$$H_0 : \quad X_1\beta = X_2\beta$$
$$H_a : \quad X_1\beta \neq X_2\beta$$

which is equivalent to

$$H_0 : \quad (X_1 - X_2)\beta = 0$$
$$H_a : \quad (X_1 - X_2)\beta \neq 0$$

- The Wald test gives:

  ▷ Estimated difference $= -0.2$ (s.e. $= 0.4$)

  ▷ W $= 0.18$

  ▷ df $= 1$

  ▷ p-value $= 0.675$

4.7 Review of Key Points

- Generalized Estimating Equations: What they are

  ▷ extension of the marginal models we have seen in Chapter 2, in the setting of categorical data

  ▷ Three components: (i) a model for the mean, (ii) a model for the variance, & (iii) an assumption for the pairwise correlations

  ▷ semiparametric approach $\Rightarrow$ no assumptions for the distribution of the data

- Generalized Estimating Equations: Features

  ▷ important to correctly specify the mean

  ▷ sandwich estimator protects against misspecification of the working correlation but works (satisfactorily) under specific settings

- Generalized Estimating Equations: Features

  ▷ only Wald tests available

  ▷ strict assumptions with respect to incomplete data

Chapter 5

Mixed Models for Discrete Data

- The previous chapter focused on the framework of Generalized Estimating Equations

  ▷ this can be seen as the extension of the marginal models for continuous data of Chapter 2 to the setting of categorical longitudinal responses

- In this chapter we will see the analogue of linear mixed models for categorical data

$$\Downarrow$$

Generalized Linear Mixed Models (GLMMs)

- The intuitive idea behind GLMMs is the same as in linear mixed models, i.e.,

  ▷ the correlation between the repeated categorical measurements is induced by unobserved random effects

  ▷ in other words: the categorical longitudinal measurements of a subject are correlated because all of them share the *same* unobserved random effect (**conditional independence assumption**)

Graphical representation of the conditional independence assumption
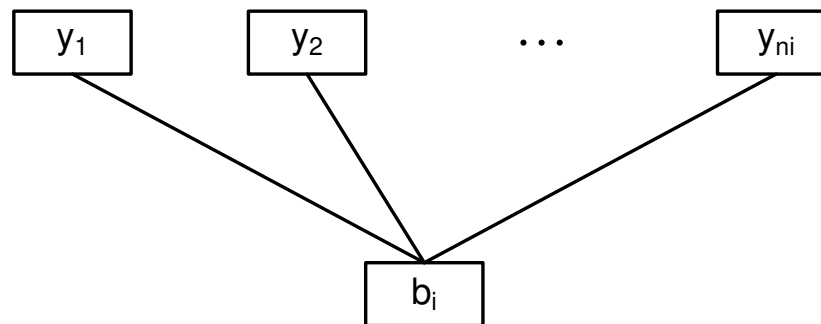
- Similarly to Chapter 4, we will focus on clustered dichotomous/binary data

  ▷ nonetheless, the same ideas and issues also apply to other categorical responses (e.g., Poisson, ordinal data, multinomial data, etc.)

- Suppose we have a binary outcome y_{ij}

$$
y_{ij} = \begin{cases} 1, & \text{if subject } i \text{ has a positive response at measurement } j \\ 0, & \text{if subject } i \text{ has a negative response at measurement } j \end{cases}
$$

- The generic mixed model for y_{ij} is a *Mixed-Effects Logistic Regression* and has the form:

$$
\begin{cases}
\log \dfrac{\pi_{ij}}{1 - \pi_{ij}} = x_{ij}^{\top}\beta + z_{ij}^{\top}b_i \\[2em]
b_i \sim \mathcal{N}(0, D)
\end{cases}
$$

where

▷ $\pi_{ij} = \Pr(y_{ij} = 1)$ the probability of a positive response

▷ x_{ij} a vector of fixed-effects covariates, with corresponding regression coefficients β

▷ z_{ij} a vector of random-effects covariates, with corresponding regression coefficients b_i

- More formally, we have the following three-part specification

  1. Conditional on the random effects b_i, the responses y_{ij} are independent and have a Bernoulli distribution with mean $E(y_{ij} \mid b_i) = \pi_{ij}$ and variance $\text{var}(y_{ij} \mid b_i) = \pi_{ij}(1 - \pi_{ij})$

  2. The conditional mean of y_{ij} depends upon fixed and random effects via the following expression:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = x_{ij}^\top \beta + z_{ij}^\top b_i$$

  3. The random effects follow a multivariate normal distribution with mean zero and variance-covariance matrix D

- <u>Notes:</u> On the definition of GLMMs

    ▷ The three-part specification of GLMMs corresponds to a full specification of the distribution of the outcome y_{ij} – this is in contrast to the GEE approach, which is a semi-parametric method

    ▷ The mean and correlation structures are simultaneously defined using random effects
    $\Rightarrow$ As we will see next, this has direct and important implications with respect to the interpretation of the parameters

5.2 Interpretation

- Example: In the AIDS dataset, a very low CD4 count (less than 150 cells/mm^3) is an indicator for opportunistic infections

  ▷ In the following analysis we dichotomize the CD4 cell counts from the AIDS dataset using this threshold

  ▷ We fit a mixed effects logistic regression with
    * *fixed effects:* time, treatment and their interaction
    * *random effects:* random intercepts

- The model has the form:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \texttt{Time}_{ij} + \beta_2 \texttt{ddI}_i + \beta_3 \{\texttt{Time}_{ij} \times \texttt{ddI}_i\} + b_i, \quad b_i \sim \mathcal{N}(0, \sigma_b^2)$$

	Value	Std.Err.	z-value	p-value
β_0	5.477	0.368	14.896	< 0.001
β_1	0.141	0.041	3.410	0.001
β_2	-0.650	0.508	-1.279	0.201
β_3	-0.027	0.056	-0.484	0.628
σ_b	5.447			

- Interpretation of the fixed-effects coefficients in the mixed-effects logistic regression model

  ▷ e.g., e^{β_1} does **not** have the interpretation of the *average* odds ratio for a month increase in follow-up

- Let's see why

  ▷ say that we compare two patients at different follow-up times who both took ddC, Patient i at month m and Patient i' at month $m + 1$

  ▷ the equation of the model for Patient i is:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \{ \text{Time}_{ij} = m \} + b_i$$

▷ the equation of the model for Patient i' is:

$$\log \frac{\pi_{i'j}}{1 - \pi_{i'j}} = \beta_0 + \beta_1 \{\texttt{Time}_{ij} = m + 1\} + b_{i'}$$

▷ hence, the corresponding odds ratio is:

$$\text{log odds ratio:} \quad \log \frac{\pi_{i'j}}{1 - \pi_{i'j}} - \log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_1 + (b_{i'} - b_i) \Rightarrow$$

$$\text{odds ratio:} \quad \frac{\pi_{i'j}/(1 - \pi_{i'j})}{\pi_{ij}/(1 - \pi_{ij})} = \exp\{\beta_1 + (b_{i'} - b_i)\} \neq \exp(\beta_1)$$

- Hence, the interpretation of e^{β_1} is not the odds ratio for unit increase of `Time` averaged over subjects, but rather for subjects with *the same random-effect value*

- To illustrate this graphically, we depict the relationship between time and the probability of low CD4 cell counts. . .
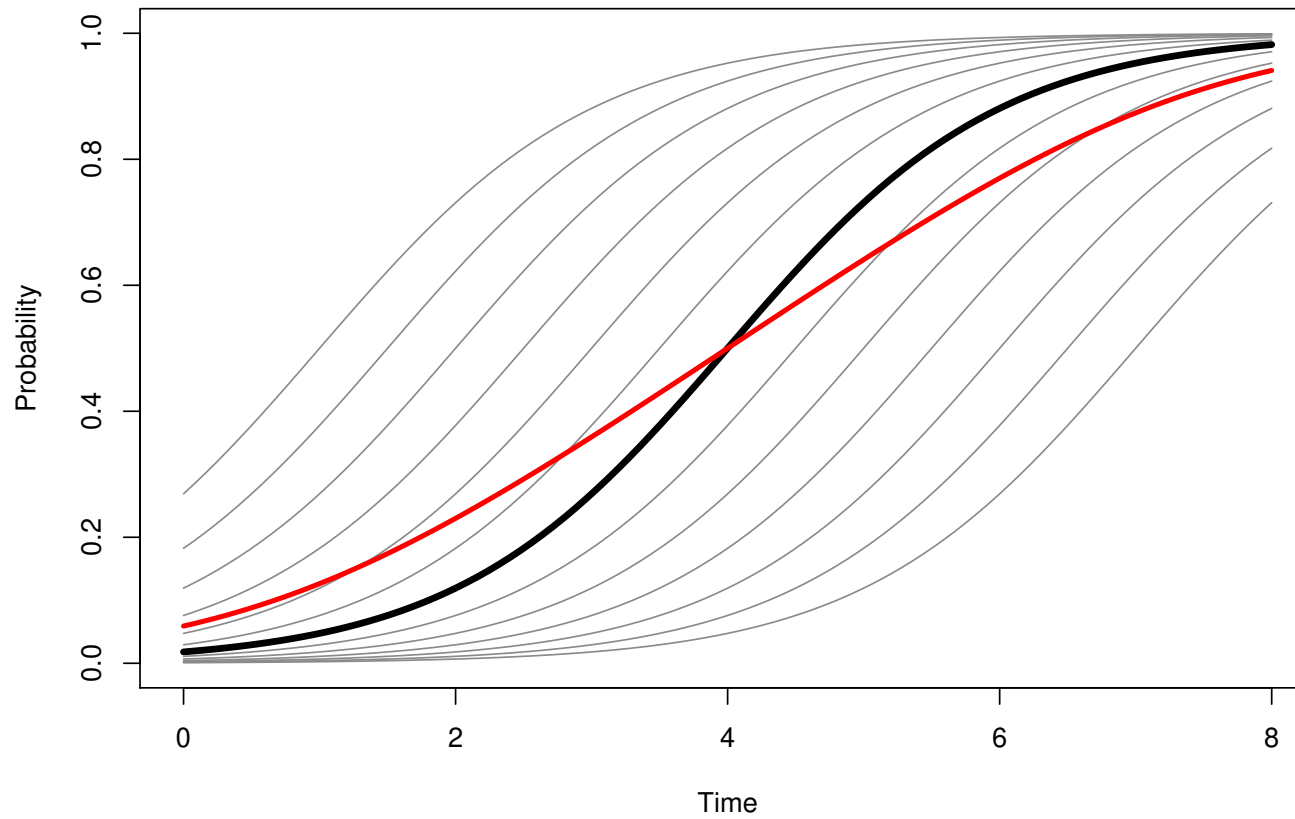
▷ the grey lines depict 13 random subjects with increasing random effects

▷ the black line corresponds to the subject with $b_i = 0$ (i.e., the mean individual)
$\Rightarrow$ This line is actually $1/[1 + \exp\{-(\beta_0 + \beta_1 \mathtt{Time}_{ij})\}]$

▷ the red line that crosses the 13 lines denotes the average longitudinal evolution of the probability of low CD4 cells counts across subjects

Why is this happening? ...

to connect the probabilities with covariates we need the *non-linear* *logit function*
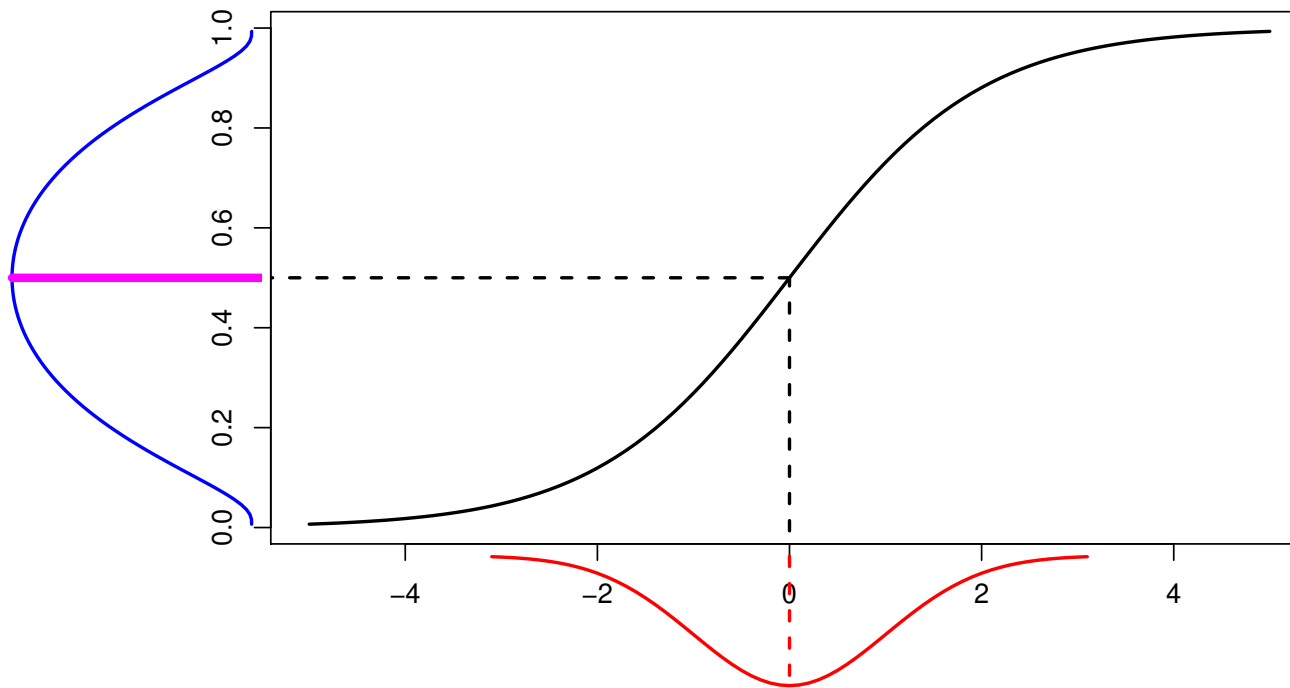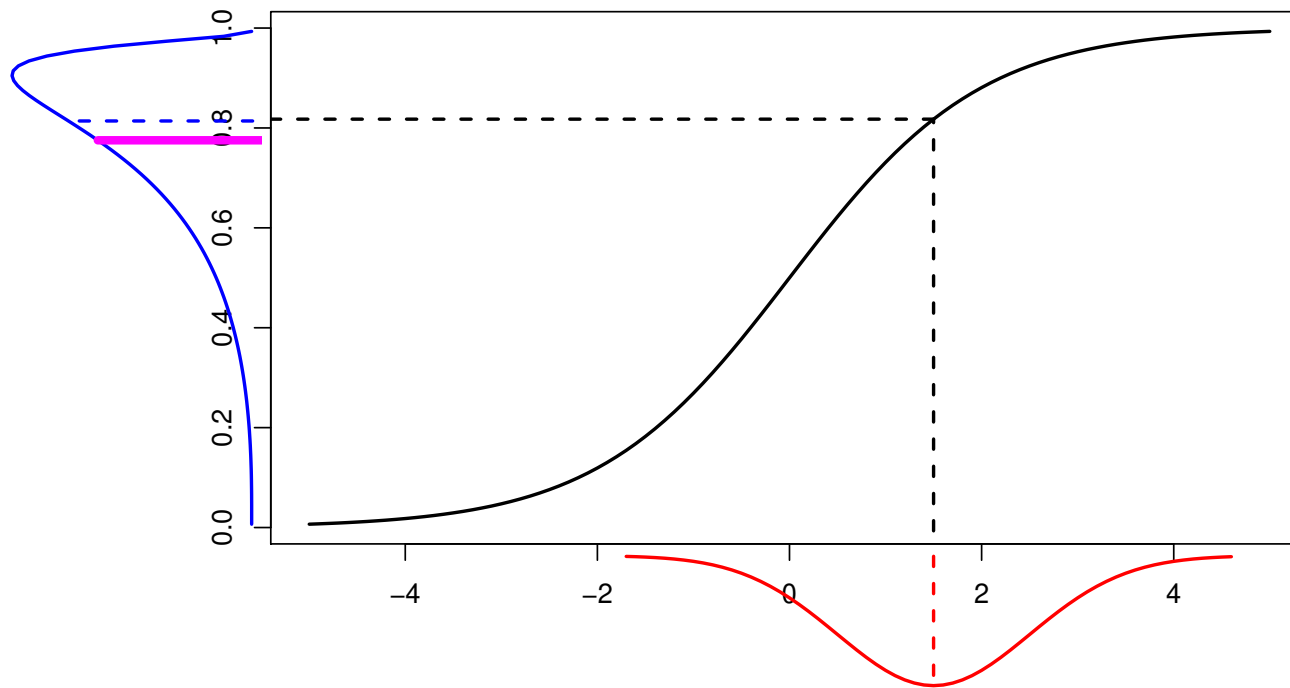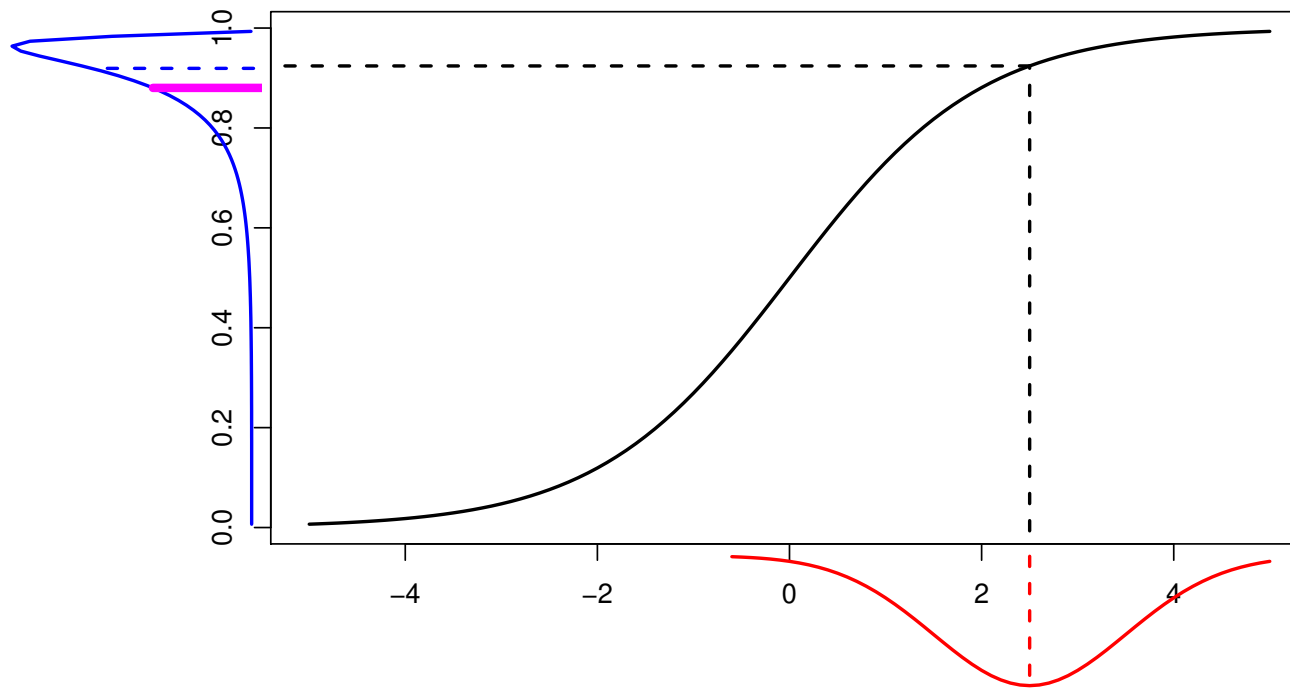
- We did not have this problem in the case of the linear mixed model because we did not have a link function

  ▷ or to put it more precisely, the link function was the identity $g(x) = x$

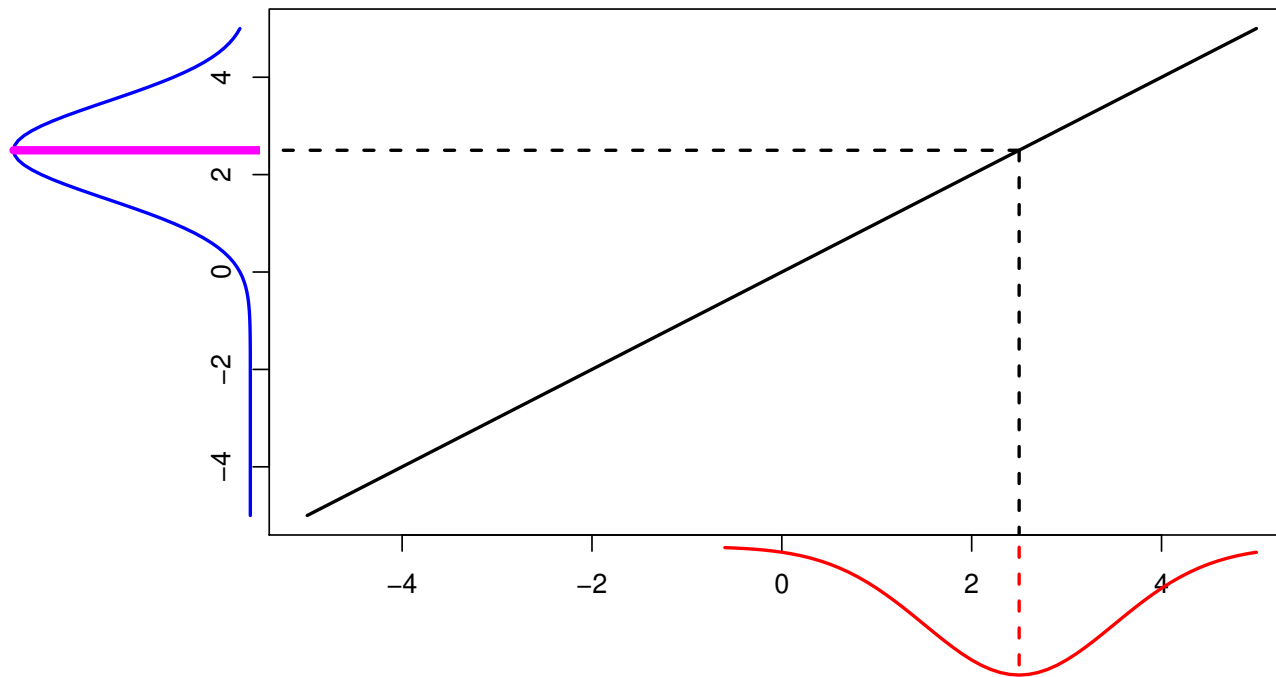- Let's see graphically again why for linear mixed models we do <u>not</u> have the same problem ...
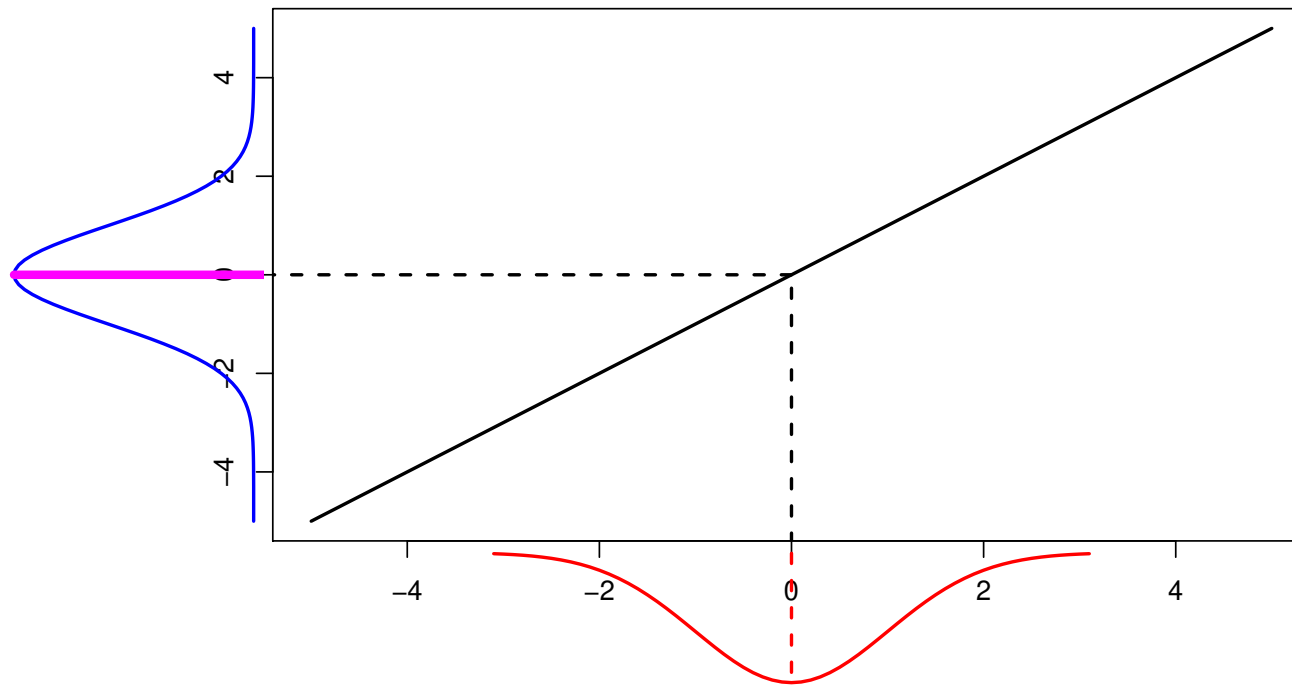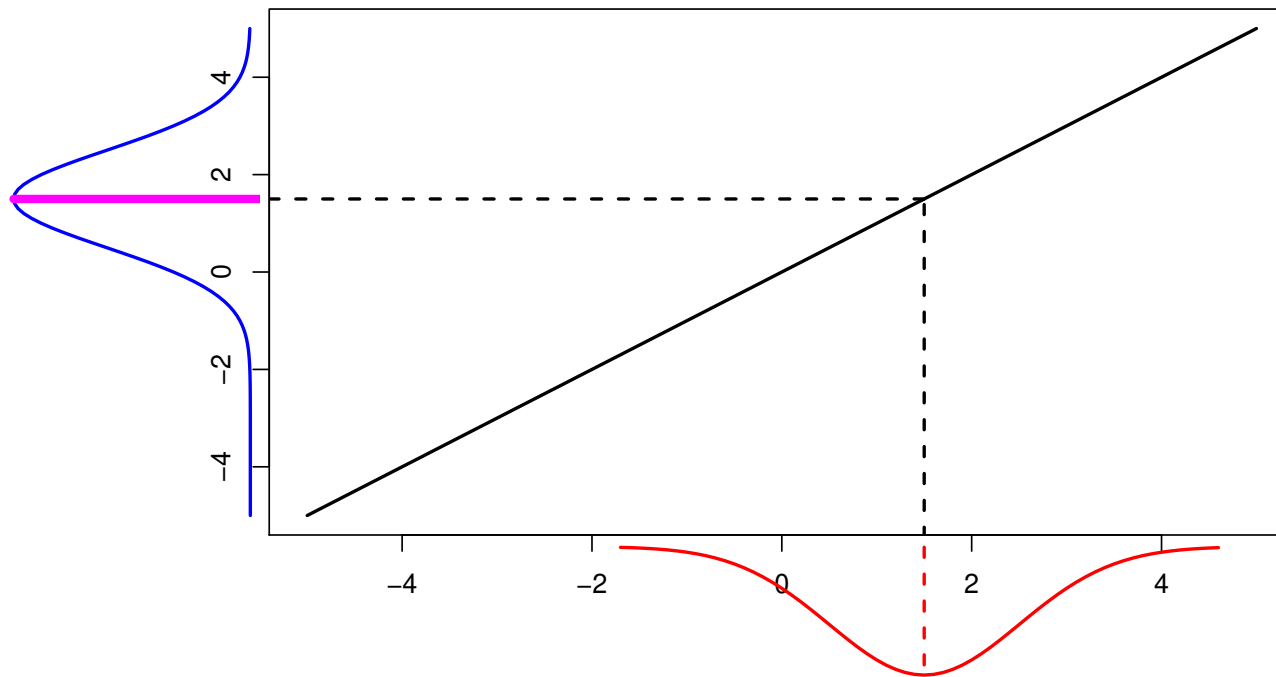
- To summarize:

  ▷ The fixed-effects regression coefficients are interpreted in terms of the effects of covariates on changes in an *individual's* transformed mean response, while holding the remaining covariates fixed

  ▷ Because the components of the fixed effects β, have interpretations that depend upon holding b_i (the i-th subject's random effects) fixed, they are often referred to as *subject-specific* regression coefficients

  ▷ As a result, GLMMs are most useful when the main scientific objective is to make inferences about individuals rather than population averages

  ▷ Population averages are the targets of inference in marginal models (i.e., GEE)

Hence, contrary to the marginal and mixed effects model for continuous data (Chapters 2 & 3), the regression coefficients from marginal models for discrete data **do not** have the same interpretation as the corresponding coefficients from mixed effects models

- **Nonetheless,** for the special case of random intercepts, there is a closed-form expression to obtain the marginal regression coefficients from the subject-specific ones, i.e.,

$$\beta^M = \frac{\beta^{SS}}{\sqrt{1 + 0.346\sigma_b^2}}$$

where

▷ β^M denotes the marginal coefficients

▷ β^{SS} denotes the subject-specific coefficients

▷ σ_b^2 denotes the variance of the random intercepts

- **However,** this formula only works for random intercepts, and cannot be used in models with more random effects

- An alternative solution has been recently proposed to overcome this issue,

  ▷ in particular the marginal coefficients β^M are obtained as the solution to the equation

$$\beta^M = (X^\top X)^{-1} X^\top \mathsf{logit}(\pi^M)$$

  where

  * X is the design matrix of the fixed effects
  * $\pi_i^M = \int \mathsf{expit}(x_i^\top \beta^{SS} + z_i^\top b_i) p(b_i)\, db_i$ are the marginal probabilities derived from the mixed model and the subject-specific coefficients

- The marginal probabilities can be obtained using a Monte Carlo sampling procedure

  ▷ for each row of the design matrix X, we generate a random sample of subjects with random effects values coming from the normal distribution $\mathcal{N}(0, \widehat{D})$, where $\widehat{D}$ denotes the estimated covariance matrix of the random effects

  ▷ for each of these simulated subjects we calculate the probability of $Y = 1$

  ▷ we take as an estimate of the marginal probability the mean of the calculated probabilities of the simulated subjects

- $\boxed{\text{Example:}}$ We continue on the previous example from the AIDS dataset (see pp.316) and we compute the corresponding marginal regression coefficients

	Subject-specific				Marginal			
	Value	Std.Err.	z-value	p-value	Value	Std.Err.	z-value	p-value
β_0	5.477	0.368	14.896	< 0.001	1.588	0.135	11.802	< 0.001
β_1	0.141	0.041	3.410	0.001	0.046	0.014	3.278	0.001
β_2	-0.650	0.508	-1.279	0.201	-0.199	0.157	-1.269	0.204
β_3	-0.027	0.056	-0.484	0.628	-0.010	0.019	-0.509	0.611
σ_b	5.447							

5.2 Interpretation (cont'd)

* We observe considerable differences between the two sets of parameters

  ▷ the subject-specific odds ratio for a unit increase over time for a specific `ddC` patients is 1.15 (95% CI: 1.06; 1.25),

  ▷ whereas the corresponding marginal odds ratio averaged over all `ddC` patients equals 1.047 (95% CI: 1.019; 1.076)

  ▷ note that the lower limit of the 95% CI for the subject-specific odds ratio almost equals the upper limit of the 95% CI for the marginal odds ratio
  $\Rightarrow$ *the confidence intervals just overlap*

- As we have previously seen, effect plots can be used to effectively communicate complex models

  ▷ especially in GLMMs, these plots also can be used to depict the marginal average evolutions (i.e., even if the fixed effects coefficients have a subject-specific interpretation, we can still calculate the marginal means)

- Example: In the PBC dataset we are interested in the probability of having high serum cholesterol levels

  ▷ we include the main effects of time, drug, age & sex

  ▷ the interaction effect between time and drug, and the interaction effect between age and sex

- In the following figure we depict the log odds ratio as a function of time, separately for each combination of the randomized treatment and sex, for the average/median patient

 ▷ average/median patient is the one with random effects values equal to zero
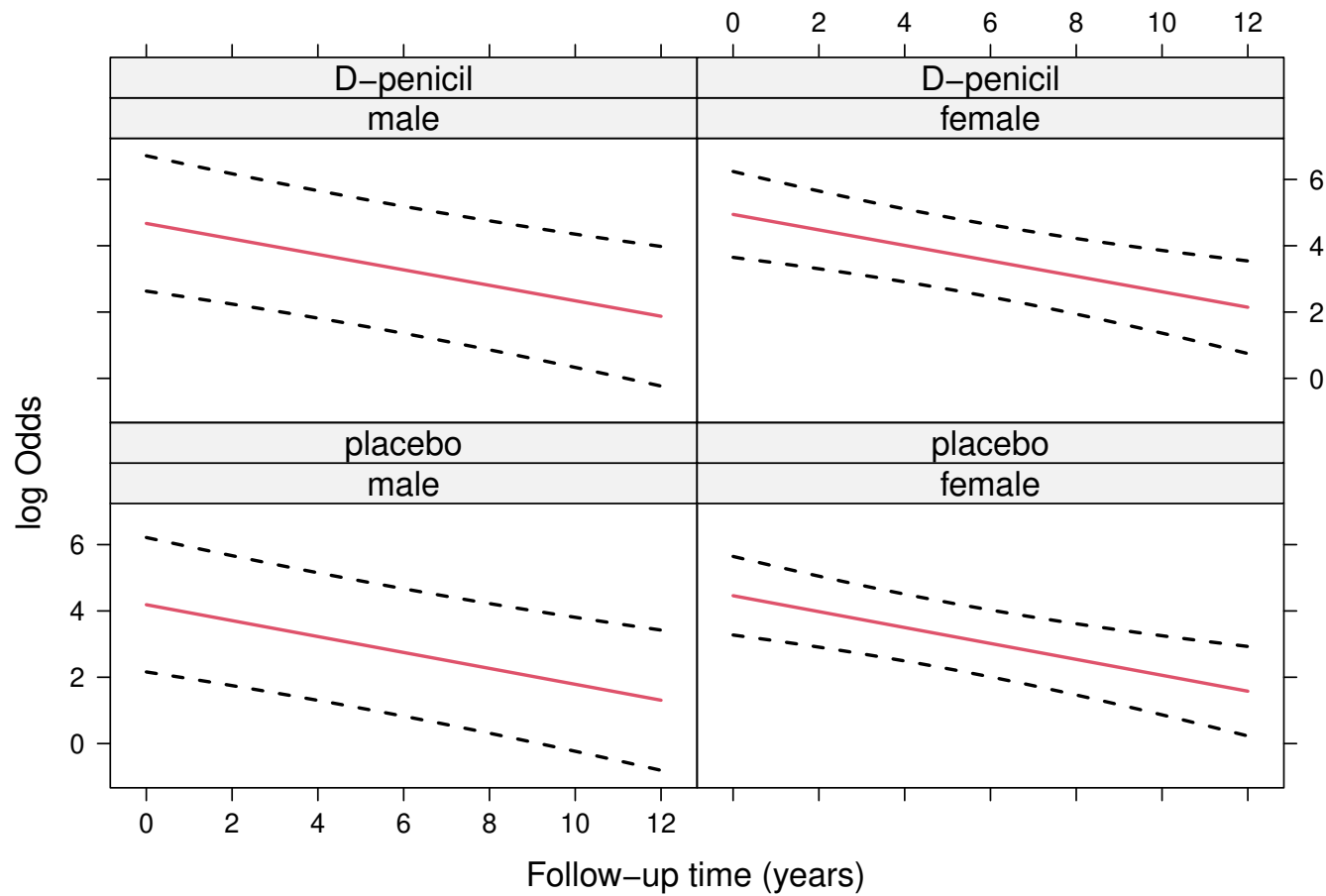
- In the following figure we depict the marginal probabilities, as a function for time, separately for each combination of randomized treatment and sex

    ▷ contrary to the previous figure, these probabilities are averaged over the patients

    ▷ they are obtained using the marginalized coefficients, presented in pp.334

    ▷ to see the difference, the blue line corresponds to the probabilities of the average patient
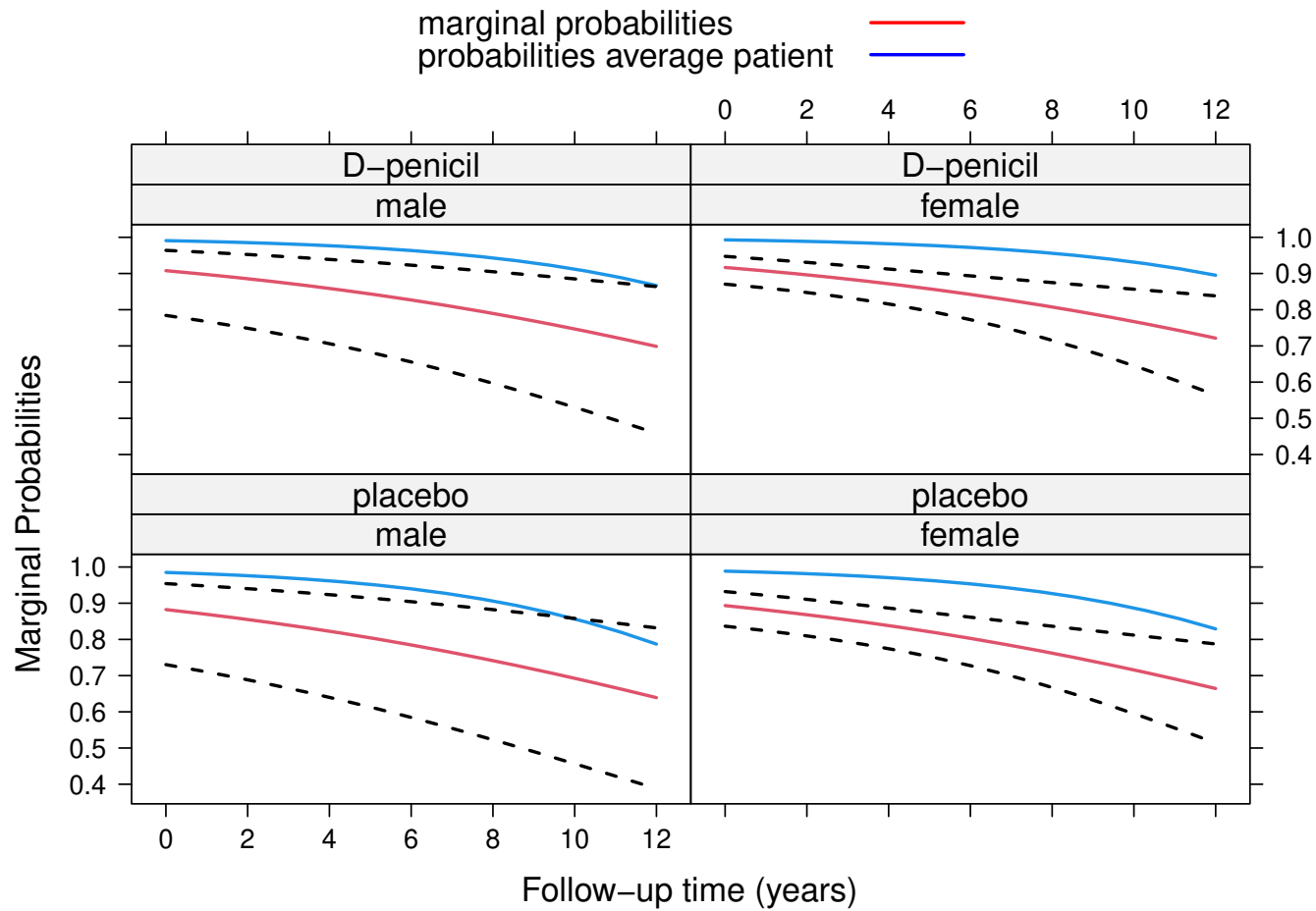
- Calculation of 95% confidence intervals for the estimated marginal probabilities is based on an additional Monte Carlo scheme

- The estimation of GLMMs is based on the same principles as in marginal and mixed models for continuous data

  ▷ i.e., we have a full specification of the distribution of the data (contrary to GEE), and hence we can use *maximum likelihood*

- Nevertheless, there is an important complication in GLMMs

The fitting of GLMMs is a computationally challenging task!

- Even though the nature of this problem is of rather computational/technical nature, we will need to discuss it in more detail ...

- What is the problem?

  ▷ The log-likelihood expression for GLMMs has the same form as in linear mixed models (see pp.160)

$$\ell(\theta) = \sum_{i=1}^{n} \log \int p(y_i \mid b_i; \theta)\, p(b_i; \theta)\, db_i$$

where θ are the parameters of the model

- In linear mixed effects models both terms in the integrand

  $\triangleright\ p(y_i \mid b_i; \theta)$

  $\triangleright\ p(b_i; \theta)$

  are densities of (multivariate) normal distributions, and also because y_i and b_i are linearly related

---

In linear mixed effects models the integral in the log-likelihood expression **has a closed-form solution** (i.e., we can compute it on paper)

- In GLMMs the two terms of the integrand denote densities of different distributions – e.g., in mixed effects logistic regression

  ▷ $p(y_i \mid b_i; \theta) \Rightarrow$ Bernoulli distribution

  ▷ $p(b_i; \theta) \Rightarrow$ multivariate normal distribution

The implication is that

<div style="border:1px solid">

In GLMMs the same integral does not have a closed-form solution

</div>

• To overcome this problem two general types of solutions have been proposed in the literature

▷ *Approximation of the integrand:* this entails approximating the product inside the integral (i.e., $\{p(y_i \mid b_i; \theta)p(b_i; \theta)\}$) by a multivariate normal distribution for which the integral has a closed-form solution

* Penalized Quasi Likelihood (PQL)
* Laplace approximation

▷ *Approximation of the integral:* this entails approximating the whole integral (i.e., $\int p(y_i \mid b_i; \theta)p(b_i; \theta)db_i$) by a sum

* Gaussian Quadrature & adaptive Gaussian Quadrature
* Monte Carlo & MCMC (Bayesian approach)

> **From the two alternatives, methods that rely on approximation of the integral have been shown to be superior**

- Though they are (much) more computationally demanding – they have a parameter that controls the accuracy of the approximation:

  ▷ in Gaussian quadrature rules it is the number of quadrature points (*adaptive Gaussian quadrature with 1 point is equivalent to the Laplace approximation*)

  ▷ in Monte Carlo/MCMC approaches it is the number of samples

- Example: We continue on the AIDS example, but we now treat the time variable as a factor (i.e., categorical) – the model has the form:

$$
\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1\{\text{Time}_{ij} = 2\} + \beta_2\{\text{Time}_{ij} = 6\} + \beta_3\{\text{Time}_{ij} = 12\} + \\
\beta_4\{\text{Time}_{ij} = 18\} + b_i
$$

where

- $\pi_{ij} = \text{Pr}(\text{CD4}_{ij} < 150)$

- $\{\text{Time}_{ij} = 2\}$ denotes the dummy variable for month 2, $\{\text{Time}_{ij} = 6\}$ the dummy variable for month 6, and so on

- We have fitted this model using

  ▷ PQL

  ▷ Laplace approximation (adaptive Gaussian quadrature with 1 point)

  ▷ adaptive Gaussian quadrature with 7, 10, 15 and 21 points

- The following figure depicts the estimated fixed effect coefficients under each approximation with corresponding 95% CIs
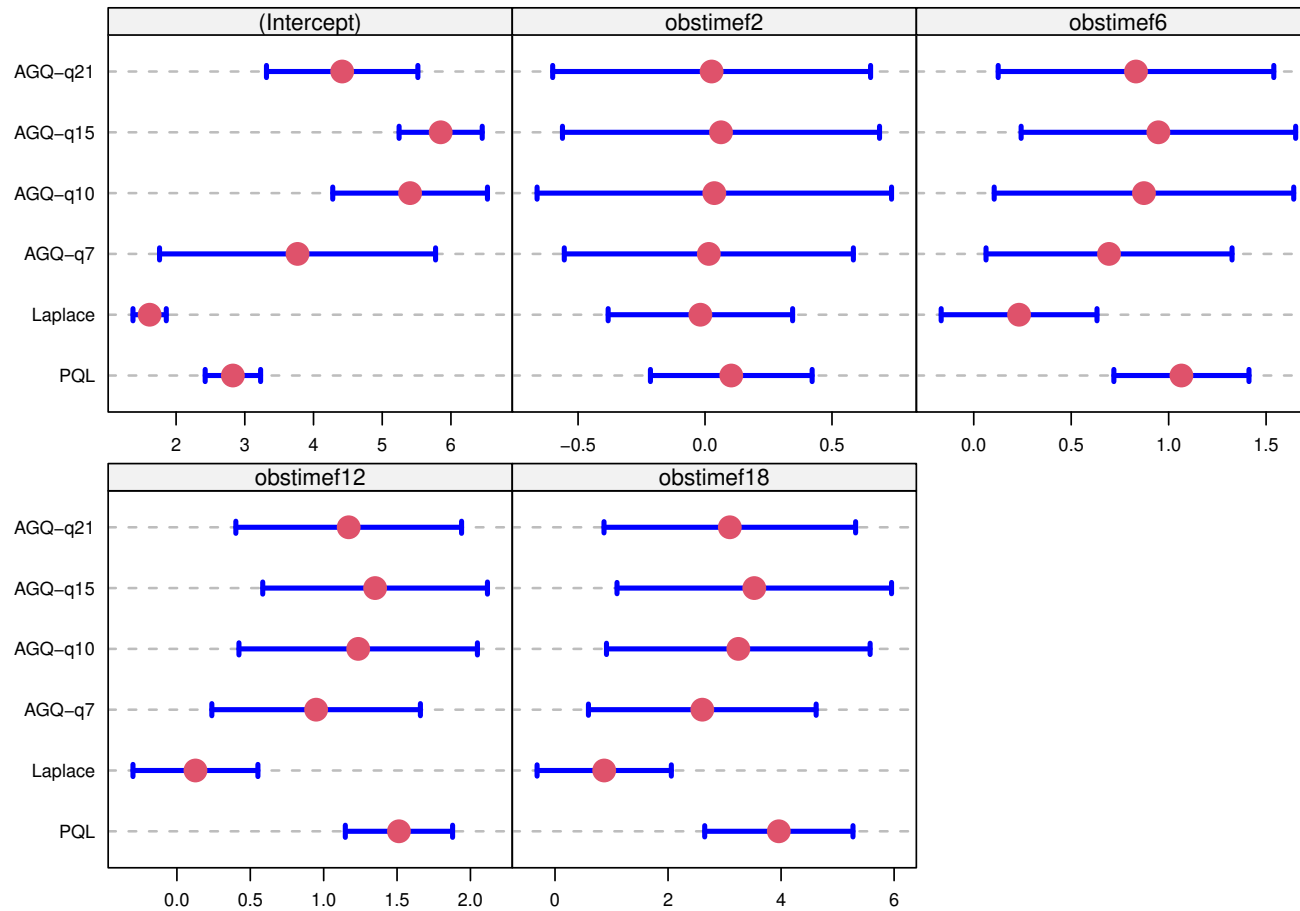
5.3 Estimation (cont'd)

* We observe considerable differences between

  ▷ PQL & Laplace (approximation of the integrand), and

  ▷ adaptive Gaussian quadrature (approximation of the integral)

* In general, PQL and Laplace will work better as the data get more 'continuous', i.e.,

  ▷ in Bernoulli data as the number of repeated measurements increases *considerably*

  ▷ in Binomial data as the number of trials increases

  ▷ in Poisson data as the rate increases

- Estimation of the random effects proceeds in a similar manner as in linear mixed models (see pp.172–179)

  ▷ based on a fitted mixed model, estimates for the random effects are based on the posterior distribution:

$$p(b_i \mid y_i; \theta) = \frac{p(y_i \mid b_i; \theta)\, p(b_i; \theta)}{p(y_i; \theta)}$$

$$\propto\ p(y_i \mid b_i; \theta)\, p(b_i; \theta),$$

  in which θ is replaced by its MLE $\hat{\theta}$

5.3 Estimation (cont'd)

- This is a whole distribution

  ▷ to obtain estimates for the random effects we typically use measures of location from this posterior distribution (e.g., mean or mode)

  ▷ as an estimate of the dispersion of the random effect we use the variance of the local curvature around the mode of the posterior distribution

- Contrary to linear mixed models in which this distribution has a closed-form, in GLMMs for categorical responses this is not the case

  ▷ calculation of the above mentioned measures of location and dispersion is achieved using numerical algorithms

5.4 GLMMs in R

R> In R there are several packages to fit GLMMs – in this course we are going to see **lme4** and **GLMMadaptive**

- The function that fits GLMMs in **lme4** is `glmer()` – this has similar syntax as the `lmer()` function that fits linear mixed models, namely

  ▷ `formula`: a formula specifying the response vector, the fixed- and random-effects structure

  ▷ `data`: a data frame containing all the variables

  ▷ `family`: a `family` object specifying the distribution of the outcome and the link function

  ▷ `nAGQ`: the number of quadrature points

R>   The following code fits a mixed effects logistic regression for abnormal serum cholesterol from the PBC dataset with random intercepts and 15 quadrature points for the adaptive Gauss-Hermite rule

```
glmmFit <- glmer(serCholD ~ year * drug + (1 | id),
                 family = binomial(), data = pbc2, nAGQ = 15)

summary(glmmFit)
```

`R>` The function that fits GLMMs in **GLMMadaptive** is `mixed_model()`; its basic arguments are

▷ `fixed`: a formula specifying the response vector, and the fixed-effects part of the model

▷ `random`: a formula specifying the random-effects part

▷ `data`: a data frame containing all the variables

▷ `family`: a `family` object specifying the distribution of the outcome and the link function

▷ `nAGQ`: the number of quadrature points

`R>` More info & examples available at:
`https://drizopoulos.github.io/GLMMadaptive/`

R> To fit the same model for the PBC data as we did above with `glmer()` the code is:

```
glmmFit <- mixed_model(fixed = serCholD ~ year * drug,
    random = ~ 1 | id, family = binomial(), data = pbc2, nAGQ = 15)

summary(glmmFit)
```

5.4 GLMMs in R (cont'd)

- Differences between `glmer()` (package **lme4**) and `mixed_model()` (package **GLMMadaptive**)

  ▷ `glmer()` only provides the adaptive Gaussian quadrature rule for the random intercepts case, whereas `mixed_model()` uses this integration method with several random terms.

  ▷ `mixed_model()` currently only handles a single grouping factor for the random effects, i.e., you cannot fit nested or crossed random effects, whereas such designs can be fitted with `glmer()`

  ▷ `mixed_model()` can fit zero-inflated Poisson and negative binomial data, allowing for random effects in the zero part

5.5 Model Building

- Model building for GLMMs proceeds in the same manner as for linear mixed models, i.e.,

  ▷ we start with an elaborate specification of the fixed-effects structure that contains all the variables we wish to study, and potential nonlinear and interactions terms

  ▷ following we build-up the random-effects structure, starting from random intercepts, next including also random slopes, quadratic slopes, etc.
    * in each step we perform likelihood ratio tests to see whether including the additional random effect improves the fit of the model

  ▷ having chosen the random-effects structure, we return to the fixed effects and check whether the specification can be simplified
    * again we first start by testing the complex terms (i.e., interactions and nonlinear terms), and then we continue to drop explanatory variables, if required

5.5 Model Building (cont'd)

- **Nevertheless,** quite often, and especially for dichotomous data, extending the random-effects structure may lead to numerical/computational problems

  ▷ this is because dichotomous data contain the least amount of information

- Hence, for dichotomous data and when we have few to moderate number of repeated measurements per subject, we often can only fit random intercepts models

5.6 Hypothesis Testing

- Having fitted a GLMM with maximum likelihood, testing of either the fixed- or random-effects structure proceeds in a similar manner as in linear mixed models

- **Important difference:** in GLMMs we do not have REML we always work with full maximum likelihood

  ▷ when we want to test the random-effects, the fixed-effects structure is also allowed to be different (though comparing nested models is a requirement for using the standard tests)

- Example: In the PBC dataset and for the dichotomous longitudinal outcome high serum cholesterol levels (defined as before as above the threshold of 210 mg/dL), we fit a model that postulates

    ▷ *fixed effects:*

    * main effects of time, treatment, and sex
    * interaction effects between time and treatment, and between drug and sex

    ▷ *random effects:* random intercepts

    We are interested in testing whether the model can be simplified by dropping the interaction terms

- The models under the two hypotheses are:

$$
\begin{cases}
H_0: \ \log \dfrac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \texttt{Time}_{ij} + \beta_2 \texttt{D-penicil}_i + \beta_3 \texttt{Female}_i + b_i \\[2em]
H_a: \ \log \dfrac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0 + \beta_1 \texttt{Time}_{ij} + \beta_2 \texttt{D-penicil}_i + \beta_3 \texttt{Female}_i + \\[1em]
\qquad\qquad \beta_4 \{\texttt{Time}_{ij} \times \texttt{D-penicil}_i\} + \beta_5 \{\texttt{Female}_i \times \texttt{D-penicil}_i\} + b_i
\end{cases}
$$

where $\pi_{ij} = \Pr(\texttt{serChol}_{ij} > 210)$

• With respect to coefficients:

$$\begin{cases} H_0 : \ \beta_4 = \beta_5 = 0 \\ H_a : \ \text{at least one different from 0} \end{cases}$$

	df	logLik	AIC	BIC	LRT	p-value
H_0		-353.59	717.17	735.76		
H_a	2	-353.33	720.66	746.68	0.52	0.7726

• The result suggests that the interaction terms do not seem to improve the fit of the model

- Similarly to previous chapters, when we want to test non-nested models we can use information criteria, i.e., the AIC or the BIC

- GLMMs are the analogue of linear mixed models for categorical data

  ▷ we include random effects in the linear predictor to account for the correlations in the outcomes belonging to the same group/cluster

- Features of GLMMs

  ▷ these models provide a complete specification of the distribution of the grouped/longitudinal outcome – contrary to GEE, which is a semi-parametric method

  ▷ interpretation of parameters is conditional on the random effects – contrary to GEE, which provide coefficients with a marginal interpretation

- Features of GLMMs

  ▷ estimation of GLMMs is more complex, and requires careful choice of numerical algorithms

  ▷ **they provide valid inferences under MAR – contrary to GEE, which only provide valid inferences under MCAR**

- Model building and hypothesis testing works in the same way as in the previous models we have seen

Chapter 6

Statistical Analysis with Incomplete Grouped Data

6.1 Missing Data in Longitudinal Studies

> **A major challenge for the analysis of grouped/longitudinal data is missing data**

▷ Even though studies are designed to collect data on every subject at a set of prespecified follow-up times

▷ Subjects often miss some of their planned measurements for a variety of reasons

6.1 Missing Data in Longitudinal Studies (cont'd)

- Implications of missingness:

    ▷ we collect less data than originally planned $\Rightarrow$ *loss of efficiency*

    ▷ not all subjects have the same number of measurements $\Rightarrow$ *unbalanced datasets*

    ▷ missingness may depend on outcome $\Rightarrow$ *potential bias*

- For the handling of missing data, we introduce the missing data indicator

$$
r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}
$$

- We obtain a partition of the complete response vector y_i

    ▷ observed data y_i^o, containing those y_{ij} for which $r_{ij} = 1$

    ▷ missing data y_i^m, containing those y_{ij} for which $r_{ij} = 0$

- We can have different patterns of missing data. . .

Subject	Visits				
	1	2	3	4	5
1	x	x	x	x	x
2	x	x	x	?	?
3	?	x	x	x	x
4	?	x	?	x	?

▷ Subject 1: Completer

▷ Subject 2: dropout

▷ Subject 3: late entry

▷ Subject 4: intermittent

- When the focus is only on dropout the notation can be simplified

  ▷ Discrete dropout time: $r_i^d = 1 + \sum_{j=1}^{n_i} r_{ij}$ (ordinal variable)

  ▷ Continuous dropout time: T_i^* denotes the time to dropout

- Focusing on dropout only is justifiable because often intermittent missing data can be considered MAR (definition of MAR follows...)

6.2 Missing Data Mechanisms

- To describe the probabilistic relation between the measurement and missingness processes Rubin (1976, Biometrika) has introduced three mechanisms

- *Missing Completely At Random (MCAR)*: The probability that responses are missing is unrelated to both y_i^o and y_i^m

$$p(r_i \mid y_i^o, y_i^m) = p(r_i)$$

- Examples

  ▷ subjects go out of the study after providing a pre-determined number of measurements

  ▷ laboratory measurements are lost due to equipment malfunction

- Features of MCAR:

  ▷ The observed data y_i^o **can** be considered a random sample of the complete data y_i

  ▷ We can use any statistical procedure that is valid for complete data
  * sample averages per time point
  * linear regression, ignoring the correlation (consistent, but not efficient)
  * t-test at the last time point
  * ...

- *Missing At Random (MAR)*: The probability that responses are missing is related to y_i^o, but is unrelated to y_i^m

$$p(r_i \mid y_i^o, y_i^m) = p(r_i \mid y_i^o)$$

- Examples

  ▷ study protocol requires patients whose response value exceeds a threshold to be removed from the study

  ▷ physicians give rescue medication to patients who do not respond to treatment
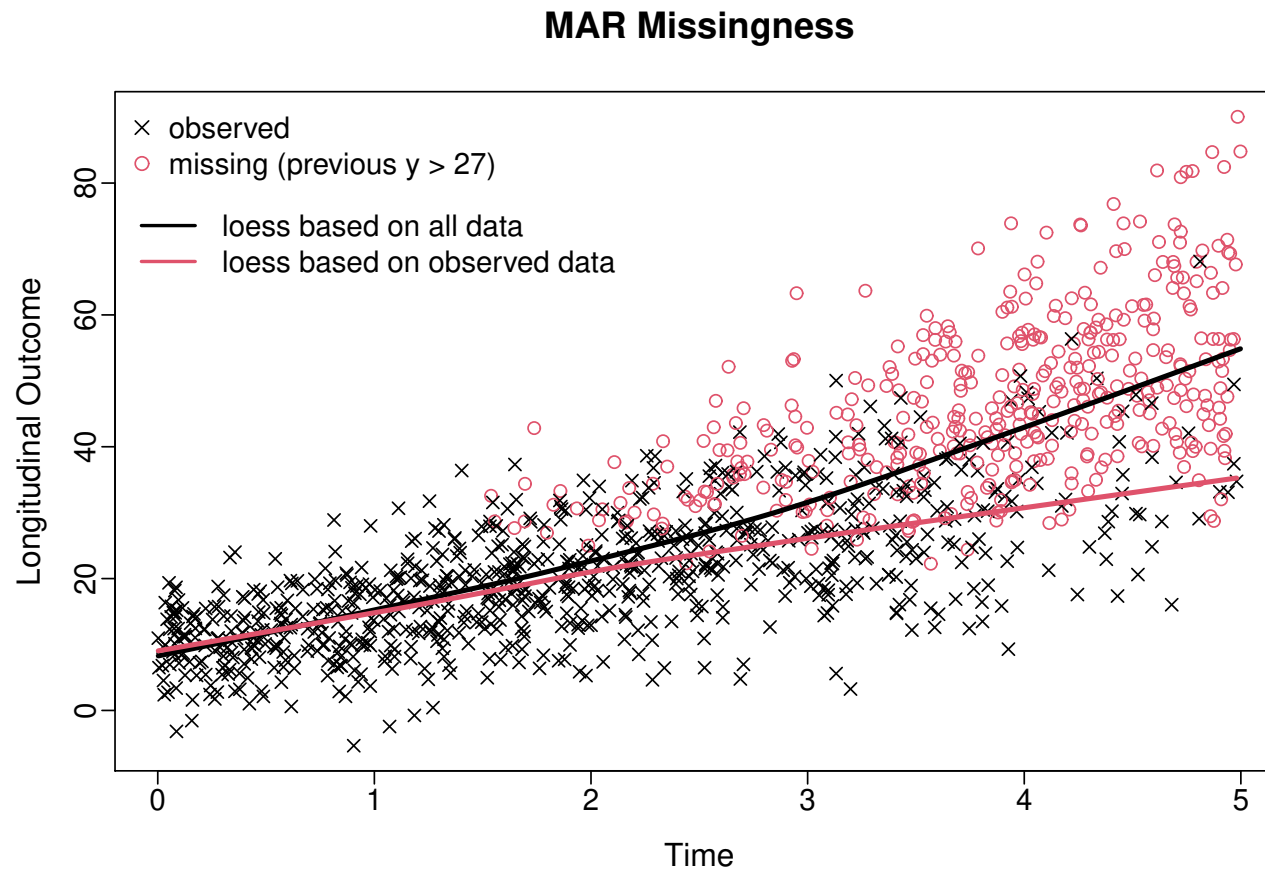
6.2 Missing Data Mechanisms (cont'd)

- Features of MAR:

  ▷ The observed data **cannot** be considered a random sample from the target population

  ▷ Not all statistical procedures provide valid results

| Not valid under MAR	Valid under MAR
sample marginal evolutions	sample subject-specific evolutions
methods based on moments, such as GEE	likelihood based inference
multivariate models with misspecified correlation structure	multivariate models with correctly specified correlation structure
marginal residuals	subject-specific residuals

MAR Missingness

MAR Missingness
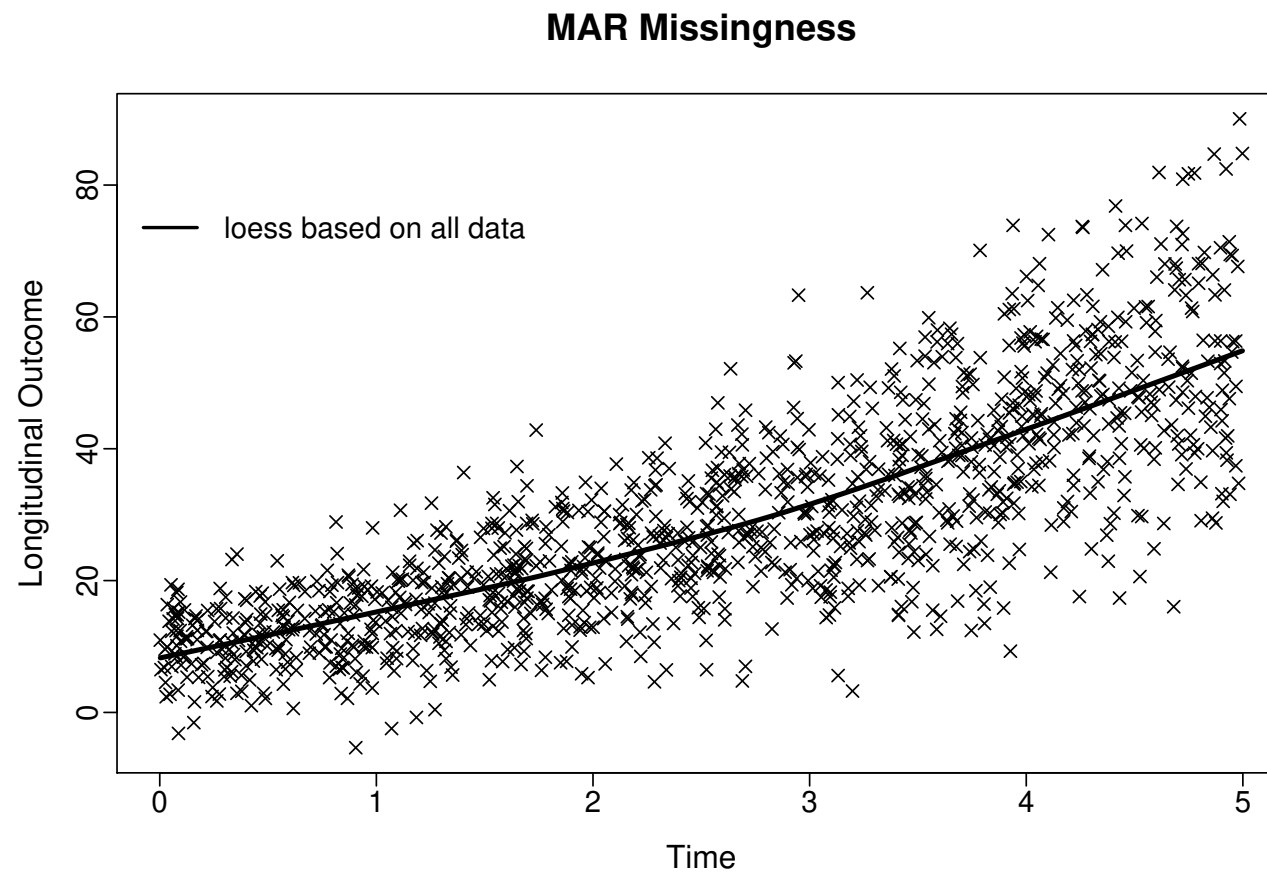
* To illustrate the important implications of incomplete data, let's return to the residuals plots we have seen in Chapter 2 (pp.138–142)

* We had fitted the following model to the PBC dataset

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

and the scatterplot of the standardized residuals versus fitted values was
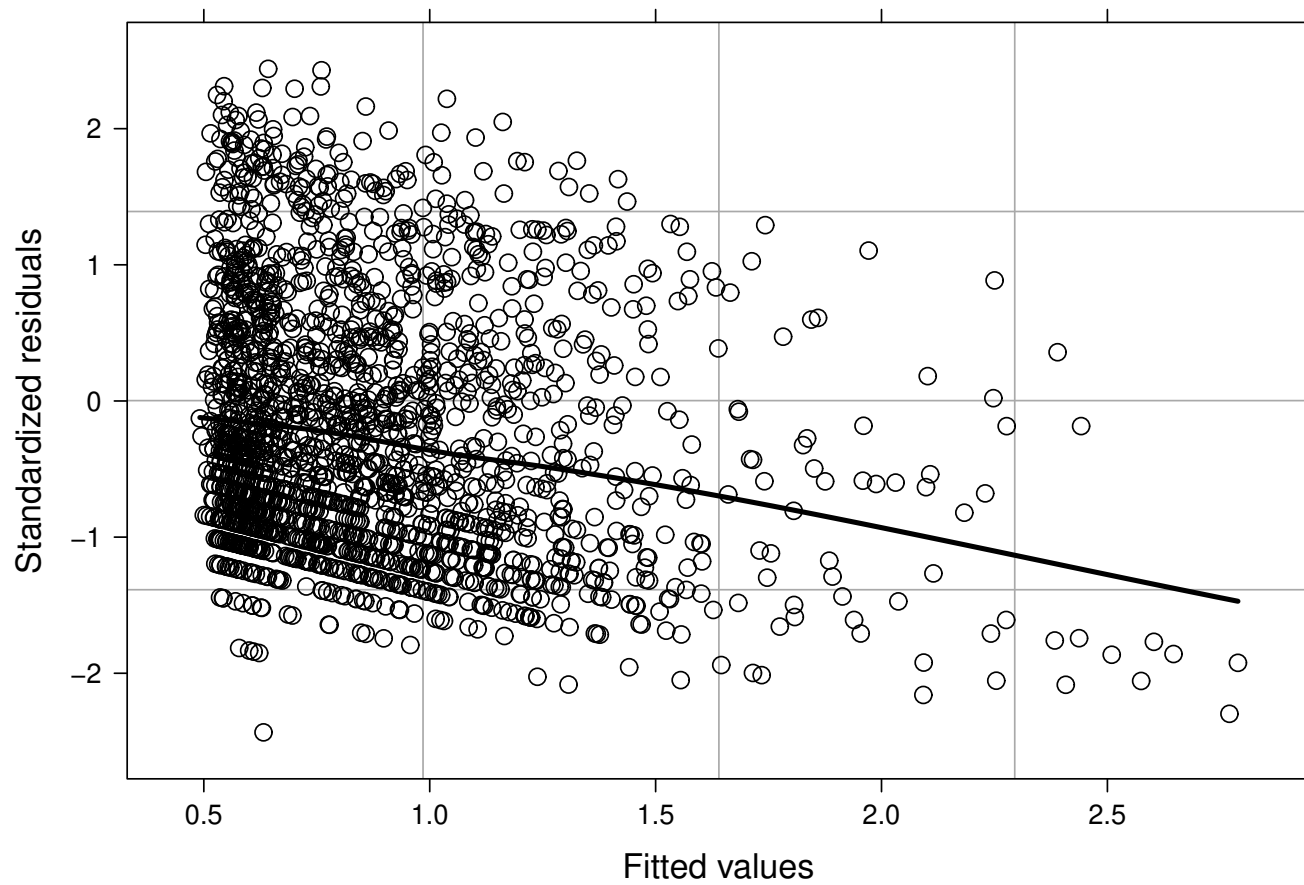
- We see a clear systematic trend

- What's the problem?

  ▷ is this really a model misspecification, or

  ▷ is it an artefact of missing data?

- Why we say that:

  ▷ patients with high serum bilirubin levels have higher chance of dropping out

  ▷ the model will account for that and give as average longitudinal evolution the average of patients who did not drop out (i.e., observed evolutions), and the patients who did drop out (i.e., unobserved evolutions)

- However, the residuals are calculated based on the observed data alone

- Hence, even if the model is correct, we could still see systematic trends because of dropout

With MAR incomplete data standard residuals plots may show misleading systematic trends

- *Missing Not At Random (MNAR)*: The probability that responses are missing is related to y_i^m, and possibly also to y_i^o

$$p(r_i \mid y_i^m) \quad \text{or} \quad p(r_i \mid y_i^o, y_i^m)$$

- Examples

  ▷ in studies on drug addicts, people who return to drugs are less likely than others to report their status

  ▷ in longitudinal studies for quality-of-life, patients may fail to complete the questionnaire at occasions when their quality-of-life is compromised

- Features of MNAR

  ▷ The observed data **cannot** be considered a random sample from the target population

  ▷ Only procedures that explicitly model the joint distribution $\{y_i^o, y_i^m, r_i\}$ provide valid inferences $\Rightarrow$ **analyses which are valid under MAR will not be valid under MNAR**

6.2 Missing Data Mechanisms (cont'd)

> **We cannot tell from the data at hand whether the missing data mechanism is MAR or MNAR**

<u>Note:</u> We can distinguish between MCAR and MAR

- *Missing Covariate Depended*: The probability that responses are missing is related to covariates x

$$p(r_i \mid x_i, y_i^o, y_i^m) = p(r_i \mid x_i)$$

- Examples

  ▷ in a study on hypertensive patients, overweight patients are inclined not to have their blood pressure measured, and BMI is related with blood pressure

- Features of Missing Covariate Depended

  ▷ If we do not include the covariates that drive the missingness process in the regression model for the longitudinal outcome Y, and these covariates are associated with Y, then we obtain an MNAR mechanism

6.3 Analysis with Incomplete Data

- We have seen what the implications of missingness are and how they complicate matters

- To this end, several approaches have been proposed to account for missing data

  ▷ **depending on the missing data mechanism, not all of them provide valid results!**

- **Complete Cases Analysis**

  ▷ **General idea:** Restrict analyses to only those subjects for which all measurements are observed

  ▷ **Advantages:**
    * very simple to implement
    * standard software can be used

  Disadvantages:
    * substantial loss of information
    * valid inferences only when missingness is completely unrelated to the outcome (i.e., MCAR)

- **Last Observation Carried Forward (LOCF)**

  ▷ **General idea:** Any missing value is replaced by the last observed value

  ▷ **Advantages:**

  * very simple to implement
  * standard software can be used

  Disadvantages:

  * extremely strong assumption that a subject's measurement stays at the same level as soon as he/she is not observed
  * even if the mechanism is MCAR, LOCF may not provide valid results
  * overestimates precision

• **Unconditional Mean Imputation**

▷ **General idea:** Each missing outcome y_{ij}^m is replaced by the average of the observed measurements at the j-th occasion

▷ **Advantages:**

* very simple to implement
* standard software can be used

Disadvantages:

* can only be implemented with balanced designs
* it provides valid results only under MCAR
* overestimates precision

● **Conditional Mean Imputation**

▷ **General idea:** The vector y_i^m of missing observations for the i-th subject is replaced by its prediction, conditional on the vector y_i^o of observed observations for that subject

* we specify a model for y_i^m conditional on y_i^o and parameters θ – often this model will result from a full specification of the marginal model $y_i = (y_i^o, y_i^m)$

* we fit the model to the completers and obtain estimates $\widehat{\theta}$ for the parameters

* based on this fitted model we can calculate predictions for the missing observations, i.e.,

$$\widehat{y}_i^m = E(y_i^m \mid y_i^o, \widehat{\theta})$$

6.3 Analysis with Incomplete Data (cont'd)

• **Conditional Mean Imputation**

▷ **Advantages:**

  * less strict assumptions that the previously mentioned approaches

  Disadvantages:

  * requires programming for its implementation
  * overestimates precision

- **Multiple Imputation**

  ▷ A common issue in all aforementioned imputation techniques was the *overestimation of precision*
  $\Rightarrow$ no correction was made for the uncertainty introduced from imputing the missing observations

  ▷ **General idea:** To propagate this uncertainty we impute not only once but *multiple* times from the conditional distribution $p(y_i^m \mid y_i^o, \widehat{\theta})$

    * M completed datasets are formed

    * we perform the same analysis in each

    * we pool the estimated parameters using Rubin's formulas

- **Multiple Imputation**

  ▷ **Advantages:**

    * correctly propagates uncertainty due to incomplete data

    * valid under MAR

    * allows for different types of analysis (e.g., concentrate at a specific time point – cross-sectional analysis)

  Disadvantages:

    * not available for grouped/clustered data in all software

6.3 Analysis with Incomplete Data (cont'd)

- **Full Specification of the Outcome Distribution**

  ▷ **General idea:** Use a model for the joint distribution of the responses – this includes the models we have seen in Chapters 2, 3, & 5 (but not the GEE approach of Chapter 4)

  ▷ **Advantages:**
     * no requirement to impute data
     * available in all standard software
     * valid results under MCAR and MAR

  Disadvantages:
     * not valid results under MNAR

- **Missing Not At Random Models**

  ▷ **General idea:** When the missing data mechanism is MNAR, we need to define a model for the joint distribution of the longitudinal outcome $\{y_i^o, y_i^m\}$ and the missingness outcome r_i

  * Three model families have been proposed
    – *selection models*
    – *pattern mixture models*
    – *shared parameter models*

- **Missing Not At Random Models**

  * Selection models use the decomposition

  $$p(y_i^o, y_i^m, r_i) = p(y_i^o, y_i^m)\, p(r_i \mid y_i^o, y_i^m)$$

  * These models postulate that the probability of dropping out is directly related on the missing longitudinal outcomes

6.3 Analysis with Incomplete Data (cont'd)

- **Missing Not At Random Models**

  * Pattern mixture models use the decomposition

  $$p(y_i^o, y_i^m, r_i) = p(y_i^o, y_i^m \mid r_i)\, p(r_i)$$

  * These models postulate that we have a different specification of the longitudinal model per dropout pattern (e.g., completers show different average evolutions than subjects who dropout earlier on)

- **Missing Not At Random Models**

  * Shared parameter models use the decomposition

  $$p(y_i^o, y_i^m, r_i) = \int p(y_i^o, y_i^m \mid b_i) \, p(r_i \mid b_i) \, p(b_i) \, db_i$$

  $$= \int p(y_i^o, \mid b_i) \, p(r_i \mid b_i) \, p(b_i) \, db_i$$

  * These models postulate that the characteristics of the longitudinal profile of a subject (described by the random effects) dictate the chance of dropping out

6.3 Analysis with Incomplete Data (cont'd)

- **Missing Not At Random Models**

  ▷ **Advantages:**

  * no requirement to impute data
  * provide valid results under MNAR

  Disadvantages:

  * only some of them available in software
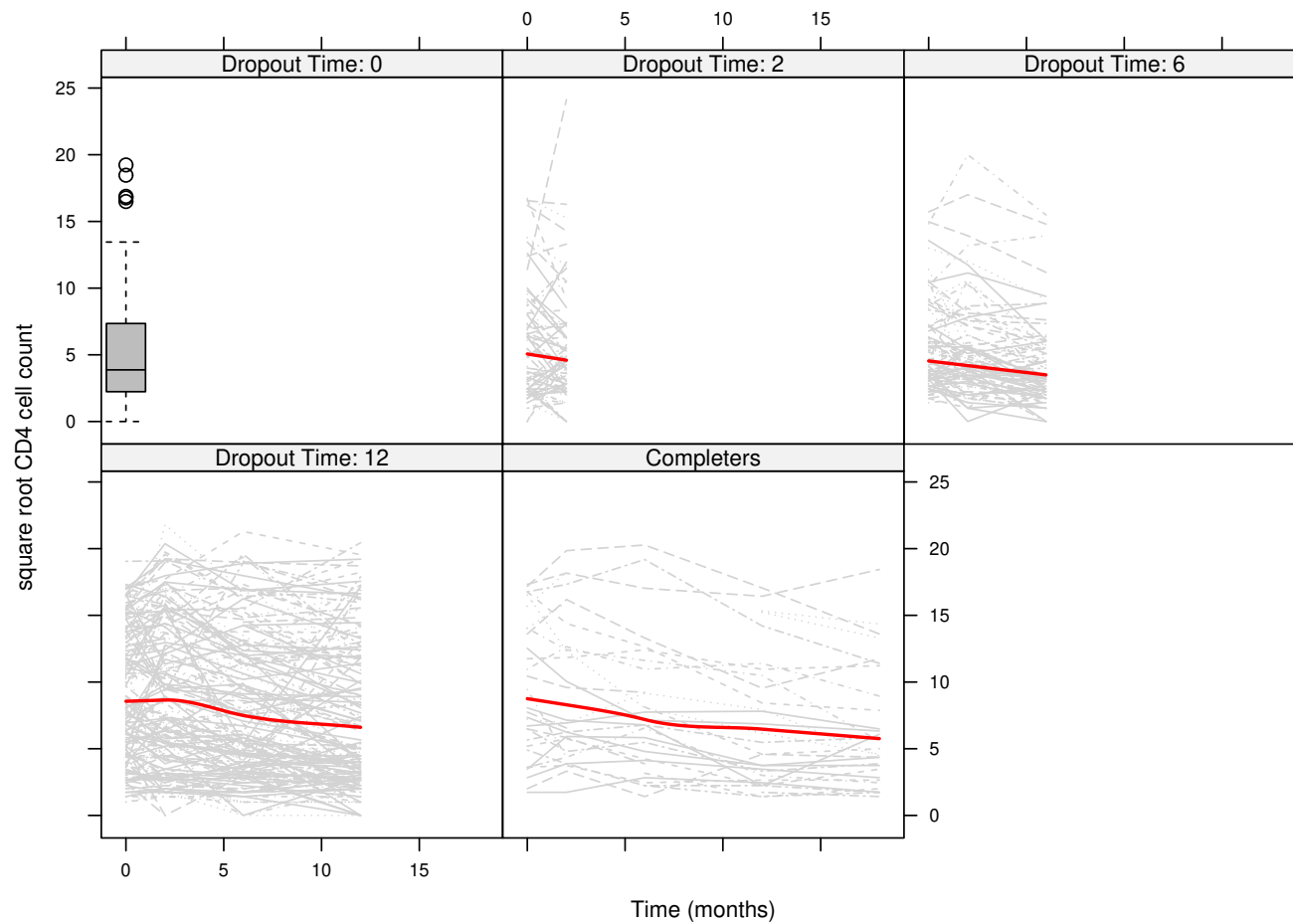 * difficult to fit
  * require sensitivity analysis

- Example: In the AIDS data set we have a considerable amount of missing data

	Missing Data per Month				
	0	2	6	12	18
Freq.	0	99	157	241	433
%	0.0	21.2	33.6	51.6	92.7

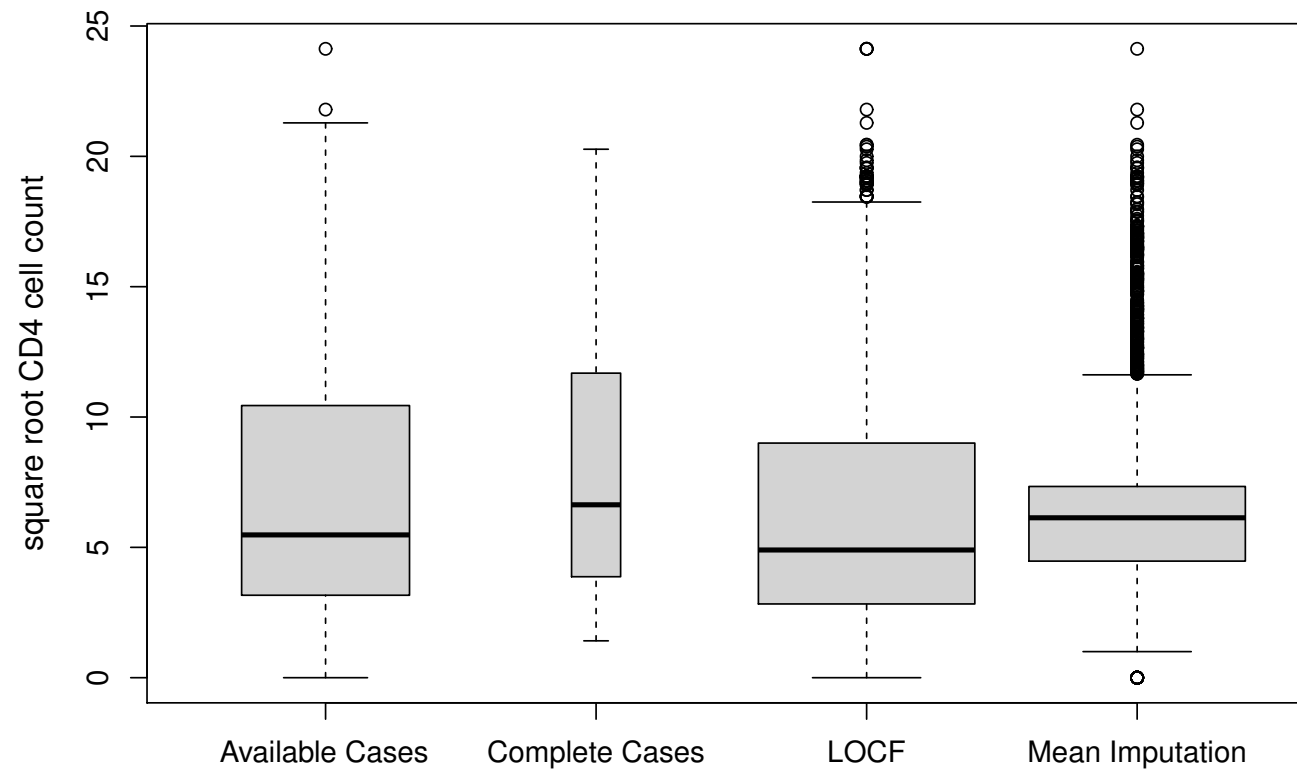- The sample evolutions of the square root CD4 cell counts per dropout pattern have the form

6.3 Analysis with Incomplete Data (cont'd)

- We are interested in a mixed model with the following specification

  ▷ fixed effects:
  - * main effects: time, AZT and previous opportunistic infection
  - * interaction effects: time with AZT, and time with previous opportunistic infection

  ▷ random effects: random intercepts & random slopes

- We will compare the MAR analysis (using a linear mixed model and all available cases) with

  ▷ complete cases analysis

  ▷ last observation carried forward analysis

  ▷ Mean Imputation analysis

- The following boxplots illustrate the distribution of square root CD4 cell counts under the different strategies

6.3 Analysis with Incomplete Data (cont'd)
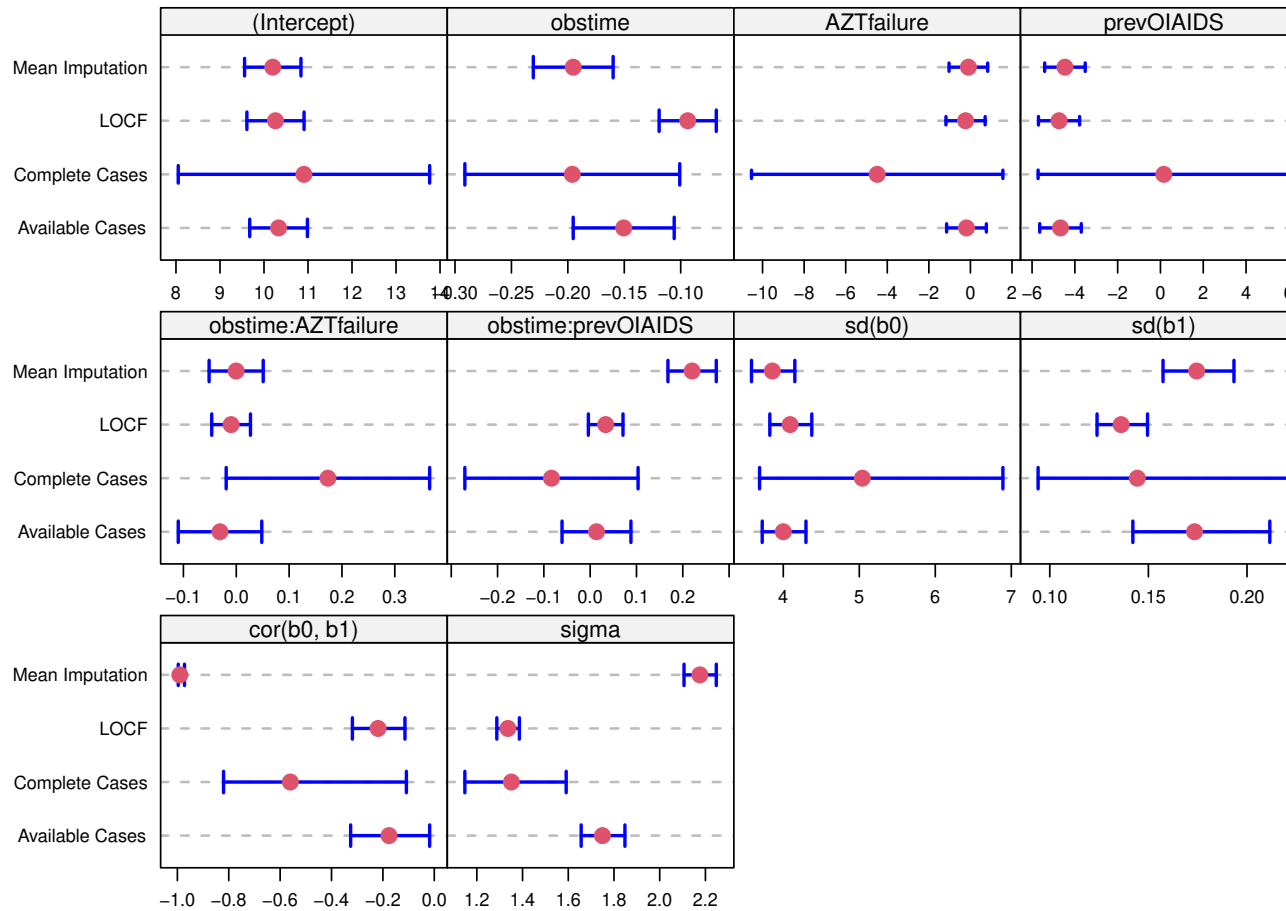
- The following figure illustrates the estimated coefficients and the corresponding 95% confidence intervals from the four mixed models fitted to the different versions of the square root CD4 cell counts variable

6.3 Analysis with Incomplete Data (cont'd)

- We observe considerable differences between the different approaches with respect to both

  ▷ parameter estimates

  ▷ standard errors (width of the confidence intervals)

The manner one decides to handle incomplete data can have a profound effect in the derived results

6.4 Summary

- It is now universally recognized that the default type of statistical analysis should provide valid results under MAR

- Hence, whenever we have missing data in the outcome, it is advisable to employ a full-likelihood approach based on the models we have seen for continuous and categorical responses
  $\Rightarrow$ **No need for (multiple) imputation**

- *However,* to be protected we need an appropriate specification of the joint distribution of the data

6.4 Summary (cont'd)

- This encompasses both the mean and the covariance/correlation structure
  $\Rightarrow$ **do not favor simpler covariance matrices if the p-value is just non-significant**

- When we also have missing data in the covariates a Multiple Imputation approach should be employed

- In general, when dealing with incomplete data it is advisable to perform a **Sensitivity Analysis**

  ▷ check how results change under logical alterations of your model

- Missing data pose an important complication in the analysis of clustered/grouped data

  ▷ loss of efficiency

  ▷ potential bias

- Need to carefully consider the reasons why data are missing

  ▷ MCAR $\Rightarrow$ missingness not related to the outcome

  ▷ MAR $\Rightarrow$ missingness related to the *observed* part of the outcome

  ▷ MNAR $\Rightarrow$ missingness related to the *unobserved* part of the outcome

• Standard analysis should be one that provides valid results under (at least) MAR

▷ full & flexible specification of the distribution of the data

▷ weighted GEE

▷ sensitivity analysis

• MNAR setting $\Rightarrow$ difficult to handle in practice

▷ in some cases, including important covariates may alleviate the problem

▷ missing covariate depended

Chapter 7

Closing

- **Features of cluster/grouped data**

  ▷ measurements in the same cluster are correlated

  ▷ distinction of between units and within units effects

  ▷ often some measurements are missing for various reasons

$$\Downarrow$$

Statistical techniques that ignore these features may produce spurious results

- **Two major modeling frameworks**

  ▷ marginal models

  ▷ mixed effects models

- **Continuous vs discrete data**

  ▷ continuous/normal data $\Rightarrow$ mixed models imply a specific marginal model

  ▷ discrete data $\Rightarrow$ more substantial differences between the two frameworks

7.1 Concluding Remarks (cont'd)

- **What we did not cover?**

  ▷ See *Advanced Topics* in the web app

     * grouped count data
     * grouped ordinal data
     * grouped bounded data
     * dynamic predictions for longitudinal data

  ▷ nonlinear models for grouped data

  ▷ transition models & alternating logistic regression

  ▷ weighted GEE & doubly robust methods

7.2 Statistical Analysis Section

- Example writing of 'Statistical Analysis' Section:

To assess changes in the biomarkers' levels over time while accounting for the correlation between the repeated measurements of each patient, we utilized the framework of linear mixed-effects models. In the fixed-effects part we allowed for a nonlinear effect of time using natural cubic splines with two internal knots placed at the corresponding percentiles of the follow-up times. In addition, we corrected for age and sex, and we also included the interaction of the nonlinear time effect with sex. In the random-effects structure we included random intercepts and random nonlinear splines using the same splines as in the fixed-effects part. The appropriate random-effects structure that best fitted the data was selected based on likelihood ratio tests. The appropriate fixed-effects structure was selected using F and likelihood ratio tests. Residual plots were used to validate the models' assumptions.

The End!

Practicals

Practical 1: Marginal Models Continuous

- We will use the PBC dataset; this is available as the object pbc2 in the R workspace available on GitHub

- To load this workspace and make the data and packages available execute the following steps:

  1. Open a new Rstudio session

  2. Create a new R script file (File → New File → R Script)

  3. Copy-paste and execute the following lines
     ```
     con <- url("https://raw.github.com/drizopoulos/Repeated_Measurements/master/Data.RData")
     load(con)
     close(con)

     library("lattice")
     library("nlme")
     library("splines")
     ```

- We will need the following variables:

  * `id`: patient id number
  * `prothrombin`: prothrombin time in sec (the response variable of interest)
  * `year`: follow-up times in years
  * `drug`: the randomized treatment
  * `sex`: the gender of the patients
  * `age`: the age of the patients

> **Aim:** To build an appropriate marginal model to investigate the relationships between the prothrombin time and the aforementioned variables

- Q1: We will start by producing some descriptive plots for the prothrombin time, similar to those we have seen in Chapter 1, i.e.,

  ▷ spaghetti plot per treatment group including the loess curve

  ▷ spaghetti plot per sex including the loess curve

  (hint: see code for Section 1.1)

  What observations can you make?

Practical 1: Marginal Models Continuous (cont'd)

- **Remove outliers:** From the plots you produced in Question 1 it was evident that we have some outlying observations

  ▷ for the rest of this practical we will exclude prothrombin times which were larger than 18 sec – to do that use the following piece of code:

```
pbc2 <- pbc2[pbc2$prothrombin < 18, ]
```

- We will continue by starting our model building exercise
  Remember

  ▷ we start with a full specification of the mean structure, and investigate the covariance structure

  ▷ based on our chosen covariance structure we can make inferences for the mean structure

- ☐Q2:☐ Start by fitting a marginal model with independent error terms using `gls()` and the following specification of the mean structure (hint: see code for Section 2.4)

  ▷ nonlinear time evolutions using natural cubic splines with 3 degrees of freedom

  ▷ correct for `sex`, `drug` and `age`

  ▷ interactions of the time effect with `sex` and `drug`

- Q2:

  ▷ interpret the results you obtained

  ▷ should we simplify the model by excluding the non-significant terms?

- Q3: Continue with the same mean structure and try different covariance structures

  ▷ first try different correlation structures, i.e., compound symmetry, continuous AR1, linear & Gaussian, and

  ▷ then extend the above structures by assuming heteroscedastic errors, i.e., that the variance increases (or decreases) over time

  (hint: see code for Section 2.9)

- Q4: Using appropriate tools (hypothesis tests, information criteria) decide which structure is the best

  ▷ which models are nested to which models?

- For the remainder we will use the covariance structure you have chosen in Q4

- Q5: Check if we can drop **all** the interaction terms

  ▷ with an F-test

  ▷ with a Likelihood Ratio Test

  (<u>hint:</u> see code for Section 2.9)

- Q6: Continue and check whether you can drop the nonlinear terms for the time effect

  ▷ to do that fit a model that assumes a linear time trend, and

  ▷ then do the likelihood ratio test to compare it to the model that includes the nonlinear terms

- Q7: Interpret the results of your final model

  ▷ regression coefficients

  ▷ covariance structure

- Q8: Use an Effect Plot to depict the model with the following settings

  ▷ year: in the range from 0 to 12 years of follow-up

  ▷ sex: both males and females

  ▷ drug: both treatment groups

  ▷ age: fixed at 49 years old

  (hint: see code for Section 2.4 – Effect Plot)

- Q9: Check the assumptions of the model using scatterplots of the standardized & normalized residuals versus the fitted values,

  ▷ overall

  ▷ separately per sex

  ▷ separately per treatment group

    (hint: see code for Section 2.11)

What are your conclusions?

Practical 2: Mixed Models Continuous

- We will use the PBC dataset; this is available as the object pbc2 in the R workspace available on GitHub

- To load this workspace and make the data and packages available execute the following steps:

  1. Open a new Rstudio session

  2. Create a new R script file (File → New File → R Script)

  3. Copy-paste and execute the following lines
     ```
     con <- url("https://raw.github.com/drizopoulos/Repeated_Measurements/master/Data.RData")
     load(con)
     close(con)

     library("lattice")
     library("nlme")
     library("splines")
     ```

- We will need the following variables:

  * `id`: patient id number

  * `prothrombin`: prothrombin time in sec (the response variable of interest)

  * `year`: follow-up times in years

  * `drug`: the randomized treatment

  * `sex`: the gender of the patients

  * `age`: the age of the patients

> **Aim:** To build an appropriate linear mixed effects model to investigate the relationships between the prothrombin time and the aforementioned variables

- Q1: Compute summary statistics for the number of repeated measurements per patient

  ▷ do we have enough information to model potential nonlinearities in the subject-specific trajectories?

- Q2: Examine graphically for samples of patients

  (hint: see code for Section 1.1)

  ▷ How do the individual longitudinal trajectories of the prothrombin time look like?

  ▷ What observations can you make?

- Q3: Start by fitting a linear mixed effects model using `lme()` with the following specification of the fixed and random effects

  (hint: see code for Section 3.2)

  ▷ *fixed effects:*
  - * linear & quadratic time evolutions, nonlinear effect of age using natural cubic splines with 3 degrees of freedom
  - * correct for `sex` and `drug`
  - * interactions of time with `sex` and `drug`, and `age` with `sex` and `drug`

  ▷ *random effects:* random intercepts

Note: As in Practical 1, in the analysis requested above, and for the remainder of this practical exclude the prothrombin times that were above 18 sec.

Practical 2: Mixed Models Continuous (cont'd)

- Q4: Keeping the mean structure (i.e., the fixed effects as is), start elaborating the random-effects structure that captures the within subject correlations, i.e., consider

  ▷ random intercepts & random slopes

  ▷ random intercepts, linear & quadratic random slopes

  ▷ random intercepts, linear, quadratic & cubic random slopes

  For each extra random effect that you add, perform the likelihood ratio test to see if it is required to add it

  ▷ which are the null and alternative hypotheses for each of these tests?

- Q5: Based on the model you selected Question 4, test whether you can drop all the *interaction terms* in order to simplify the model

  ▷ first perform the omnibus test for all the interaction terms

  ▷ if it is (highly) non-significant, you can drop them

  ▷ if it is significant, find which group(s) are the significant ones

- Q6: In the same spirit as in Question 5, test whether you can drop all the *nonlinear terms* to simplify the model

  ▷ first perform the omnibus test for all the nonlinear terms

  ▷ if it is (highly) non-significant, you can drop them

  ▷ if it is significant, find which group(s) are the significant ones

- Q7: Interpret the results of your final model

  ▷ regression coefficients

  ▷ covariance structure

- Q8: Compare the marginal and subject-specific predictions from your final model, i.e.,

  ▷ add in you data frame the marginal and subject-specific fitted values from the final model (remember to use the dataset that excludes the outliers)

  ▷ select the following patients from the data set: 133, 36, 180, 11, 168, 116, 70, 58, 82, 104, 43, 21, 101, 210, 176, 157

  ▷ create the plot that compares the predictions

  (hint: see code for Section 3.4)

- Q9: Use an Effect Plot to depict the model with the following settings

  ▷ year: in the range from 0 to 12 years of follow-up

  ▷ sex: both males and females

  ▷ drug: both treatment groups

  ▷ age: the median age from the original data for the respective four groups of patients (i.e., the median age of male in placebo, females in placebo, males in active treatment & females in active treatment)

(hint: see code for Section 3.2 – Effect Plot)

- Q10: Check the assumptions of the model using scatterplots of the standardized subject-specific & standardized marginals residuals versus the fitted values,

  ▷ overall

  ▷ separately per sex

  ▷ separately per treatment group

    (hint: see code for Section 3.11)

What are your conclusions?

Practical 3: Marginal Models Discrete

- We will use the PBC dataset; this is available as the object pbc2 in the R workspace available on GitHub

- To load this workspace and make the data and packages available execute the following steps:

  1. Open a new Rstudio session

  2. Create a new R script file (File → New File → R Script)

  3. Copy-paste and execute the following lines

     ```
     con <- url("https://raw.github.com/drizopoulos/Repeated_Measurements/master/Data.RData")
     load(con)
     close(con)

     library("lattice")
     library("geepack")
     library("splines")
     ```

Practical 3: Marginal Models Discrete (cont'd)

- We will need the following variables:

  * `id`: patient id number

  * `prothrombin`: prothrombin time in sec (the response variable of interest)

  * `year`: follow-up times in years

  * `drug`: the randomized treatment

  * `sex`: the gender of the patients

  * `age`: the age of the patients

> **Aim:** To build an appropriate GEE model to investigate the relationships between a dichotomized version of the prothrombin time and the aforementioned variables

- Q1: A normal prothrombin time is between 11 and 13 sec

  ▷ create a dichotomous variable, with '0' denoting a normal prothrombin time, and '1' an abnormal one

- Q2: Examine graphically the probability of abnormal prothrombin time
  (hint: see code for Section 1.1)

  ▷ separately per treatment including the loess curve

  ▷ separately per sex including the loess curve

  ▷ separately for each age category $[25, 43]$, $[43, 50]$, $[50, 55]$ and $[55, 80]$ including the loess curve

  ▷ what observations can you make?

- Q3: The researchers in this study made the following conjectures

  ▷ the log odds of abnormal prothrombin time may evolve nonlinearly during follow-up;

  ▷ in addition, it is plausible that the log odds evolutions over time are different between males and females, and between placebo and treated patients;

  ▷ furthermore, age is an important risk factor, and the effect of age may be modified by sex

  Translate the above conjectures into a suitable GEE model for the log odds of abnormal prothrombin time

  ▷ use the exchangeable working correlation matrix, and

  ▷ for the nonlinear terms use natural cubic splines with 2 degrees of freedom

  (hint: see code for Section 4.3)

- $\boxed{\text{Q4:}}$ Re-fit the model you fitted in Question 3 by assuming

  ▷ an independence working correlation matrix, and

  ▷ an AR1 working correlation matrix

  ▷ Compare the estimated coefficients and the corresponding naive and sandwich standard errors using a coefficients' plot

  (<u>hint:</u> see code for Section 4.5)

  ▷ Which working correlation matrix do you choose and why?

Practical 3: Marginal Models Discrete (cont'd)

- Q5: The researchers in the study want to see if the model can be simplified by dropping the *interaction terms*

  ▷ first perform the omnibus test for all the interaction terms

  ▷ if it is (highly) non-significant, you can drop them

  ▷ if it is significant, find which group(s) are the significant ones

- Q6: Do the same for the *nonlinear terms*, i.e.,

  ▷ first perform the omnibus test for all the nonlinear terms

  ▷ if it is (highly) non-significant, you can drop them

  ▷ if it is significant, find which group(s) are the significant ones

- Q7: Interpret the results of your final model

- Q8: Use an Effect Plot to depict the model with the following settings

  ▷ year: in the range from 0 to 12 years of follow-up

  ▷ sex: both males and females

  ▷ drug: both treatment groups

  ▷ age: 49 years old

    (hint: see code for Section 4.3 – Effect Plot)

Do the plot in both the log odds and probability scales

- Q9: From the effect plot we observe that the trajectories of the log odds for males and females in the D-penicillamine group are nonlinear (more so for the females)

  ▷ test in males and females separately

  ▷ with age 49 years old

  ▷ whether there are differences in the log odds of abnormal prothrombin time

  ▷ at the follow-up years 2, 6, 8 and 10

  ▷ in other words, perform all the pairwise comparisons for the aforementioned follow-up times

  ▷ should you adjust for multiple comparisons?

  (hint: see code for Section 4.6 – complex effects)

Practical 4: Mixed Models Discrete

- We will use the PBC dataset; this is available as the object pbc2 in the R workspace available on GitHub

- To load this workspace and make the data and packages available execute the following steps:

1. Open a new Rstudio session

2. Create a new R script file (File → New File → R Script)

3. Copy-paste and execute the following lines
```
con <- url("https://raw.github.com/drizopoulos/Repeated_Measurements/master/Data.RData")
load(con)
close(con)

library("lattice"); library("splines")
library("lme4"); library("GLMMadaptive"); library("MASS")
```

- We will need the following variables:

  * `id`: patient id number

  * `prothrombin`: prothrombin time in sec (the response variable of interest)

  * `year`: follow-up times in years

  * `drug`: the randomized treatment

  * `sex`: the gender of the patients

  * `age`: the age of the patients

Aim: To build an appropriate GLMM to investigate the relationships between a dichotomized version of the prothrombin time and the aforementioned variables

---

- Q1: A normal prothrombin time is between 11 and 13 sec

  ▷ create a dichotomous variable, with '0' denoting a normal prothrombin time, and '1' an abnormal one

- Q2: Examine graphically the probability of abnormal prothrombin time for each patient

  (hint: see code for Section 1.1)

  ▷ create the subject-specific smooth trajectories of abnormal prothrombin time *for patients who had more than five measurements*

  ▷ use as a smoother the `"splines"` option in the `'type'` argument of `xyplot()`

  ▷ what observations can you make?

- Q3: The researchers in this study made the following conjectures

  ▷ the subject-specific log odds of abnormal prothrombin time evolve linearly during follow-up;

  ▷ in addition, it is plausible that the subject-specific log odds evolutions over time are different between males and females;

  ▷ furthermore, drug is expected to affect prothrombin time, and its effect may be modified by sex

Translate the above conjectures into a suitable GLMM for the log odds of abnormal prothrombin time using random intercepts, and 15 quadrature points for the adaptive Gauss-Hermite rule

(hint: see code for Section 5.2)

- Q4: Test wether it is required to also include a random slopes component using a likelihood ratio test

  ▷ depending on the result keep the model that best fits the data

- Q5: Continue by testing whether you can drop all interaction terms from the model using a likelihood ratio test

- Q6: Interpret the parameters in your final selected model

- Q7: Use an Effect Plot to depict the conditional log odds ratios for the following settings

  ▷ year: in the range from 0 to 12 years of follow-up

  ▷ sex: both males and females

  ▷ drug: both treatment groups

    (hint: see code for Section 5.2 – Effect Plot)

- Q8: Create a second effect plot with the same settings as in Question 7 but for the marginal probabilities

  ▷ also include the probabilities of the median subject

No slides available for the Quizzes

No slides available for the Advanced Topics