Analysis of Longitudinal Data

Sonja Greven, Jona Cederbaum

Summer Term 2015

With thanks to Anne-Laure Boulesteix for slides from previous years

Language

- The lecture is held in English, with German summaries of the last lecture at the beginning of each lecture.
- There will be German and English versions of the Übungsblätter/work sheets and the exam.
- The language for the lab will be discussed in the first lab (you can send Jona Cederbaum an email if you cannot come and have a preference).
- You can always ask questions in German.
- Falls jemand große Schwierigkeiten mit Englisch hat, reden Sie bitte mit mir und wir finden einen Lösung.

Dates

- Lecture (Prof. Dr. Sonja Greven): Wednesday 10.10 - 11.50 (Leo 13, 1201) and ca. every second Monday, 12.10-13.50 (M 014)
- Lectures July 1 and 8 are cancelled - instead all other lectures will be 10 minutes longer.
- Lab session (Dipl.-Stat. Jona Cederbaum): ca. every second Monday, 12.15-13.45 (CIP-Pool 042)
- Up-to-date times and rooms will be on the course website `http://www.statistik.lmu.de/institut/ag/fda/ALD_2015/`.
- Sprechstunde / consultation times: by appointment

Exam

- 23 July 2015, 10:15-12:15
- Two hour long exam
- Lectures and lab sessions are both relevant for the exam.
- The official version will be German, there will also be an English version, which can be accepted for credit instead.
- You may bring two pages with notes (front and back) in addition to a calculator and a dictionary if necessary (**not** open book).

References

- Diggle, Heagerty, Liang, and Zeger (2002). Analysis of longitudinal data. Oxford University Press.

- Fitzmaurice, Laird, Ware (2004). Applied longitudinal analysis. Wiley.

- Molenberghs and Verbeke (2005). Models for Discrete Longitudinal Data. Springer.

- Verbeke and Molenberghs (2000). Linear Mixed Models for Longitudinal Data. Springer.

Additional papers and books are referenced in the slides. A bibliography of the most important ones will be on the website for (voluntary) further reading.

Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

1.2 Examples

1.3 Correlation and modeling approaches

What are longitudinal data?

Repeated Measures Data are data for which the variable of interest is measured repeatedly for the same subjects under different conditions.
$\longrightarrow$ Example: repeated blood pressure measurements for several subjects in different stress situations.

Longitudinal Data are a particular type of **Repeated Measures Data**, for which the variable of interest is measured for several subjects **repeatedly over time**.
$\longrightarrow$ Example: repeated blood pressure measurements for several subjects over 12 months.

[We will use the term "subject" for convenience, even if the unit of observation may also be an animal, crop field, country etc.]

Examples of longitudinal studies

- **Cohort studies** set up a cohort of people sharing some characteristic (e.g. born in the same year, free of a certain disease that is prospectively studied) and follow it over time. Often used in medicine/epidemiology, but also in other areas.

- **Panel studies** are similar to cohort studies, often collecting repeated measurements at specified time intervals, but the term is more common in the social and economic sciences. In some uses of the term, the panel is drawn to represent a cross-section of the population being studied and this sometimes involves replacement of panel members leaving the study.

- In **randomized** (clinical) **trials**, subjects are randomly assigned to treatment groups and in some trials followed up over time.

Notation and special cases

- Let n_i be the number of observations per subject for subjects $i = 1, \ldots, N$.
- Let $t_{i1}, \ldots, t_{in_i}$ be the time points where subject i is measured.
- **Balanced data** has the same number of observations $n_1 = \cdots = n_N$ and the same time points $t_{ij} \equiv t_j, j = 1, \ldots, n_i$, for all subjects i.
- If the observation times also have the same distance $d = t_{j+1} - t_j$ for all j, they are called **equally spaced**.

Some observations on longitudinal data

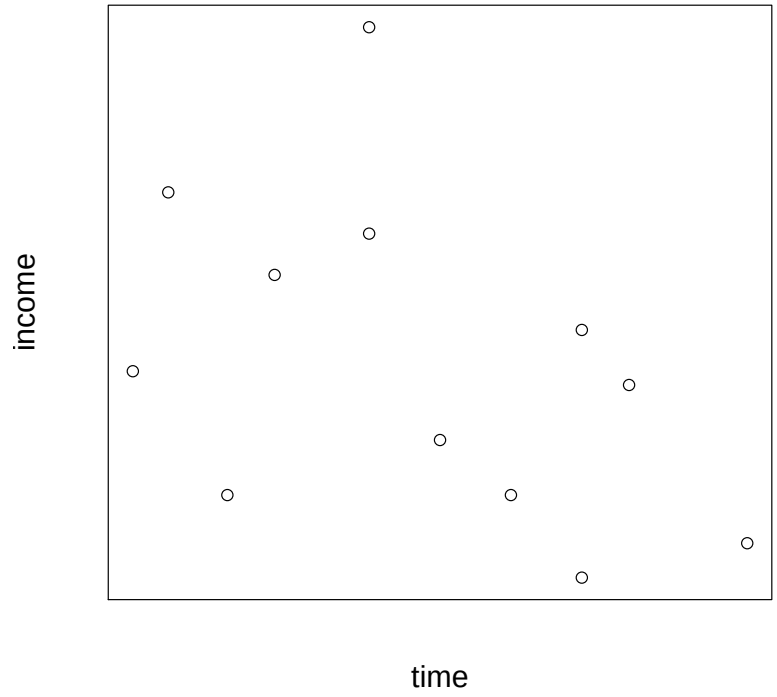
- Covariates can be **time-invariant** and only measured at baseline, e.g. gender. Or they can be **time-varying** and measured over time, e.g. physical activity.

- Sometimes longitudinal data is measured together with **survival / time-to-event data** (more in Chapter 12).

- Longitudinal data can be measured **prospectively** or **retrospectively** (e.g. via a survey or by searching through archives). Prospective studies are typically more reliable (e.g. recall bias, when for example patients who developed a disease better remember risk factors they deem important).

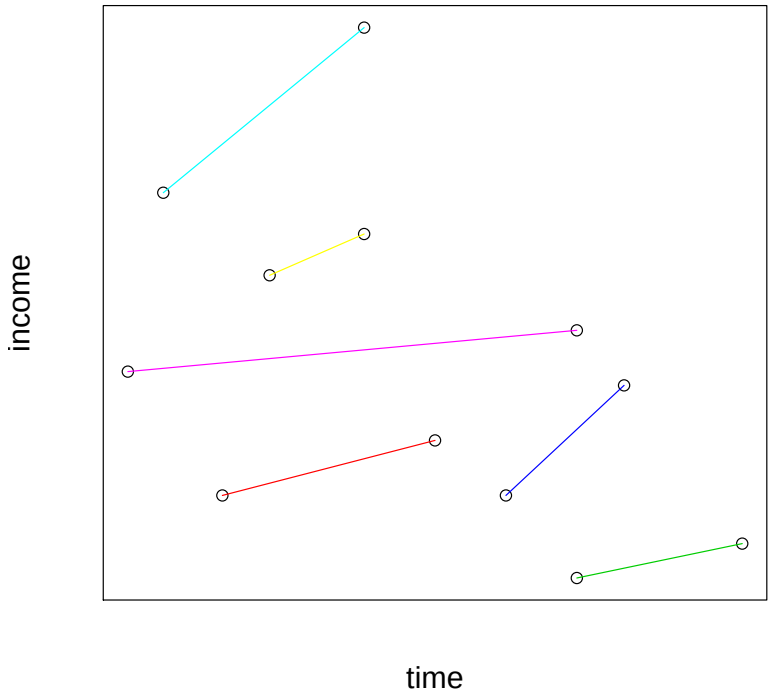
Advantages of longitudinal studies

We can distinguish longitudinal from cross-sectional effects.



Is income decreasing over time?

Advantages of longitudinal studies



Income is increasing over time for each person.

Starting salaries seem to be decreasing over time.

Advantages of longitudinal studies

Longitudinal studies can follow individual change over time and are thus more informative than cross-sectional studies.

We can distinguish **cross-sectional** (β_C) and **longitudinal** (β_L) **effects**, e.g.

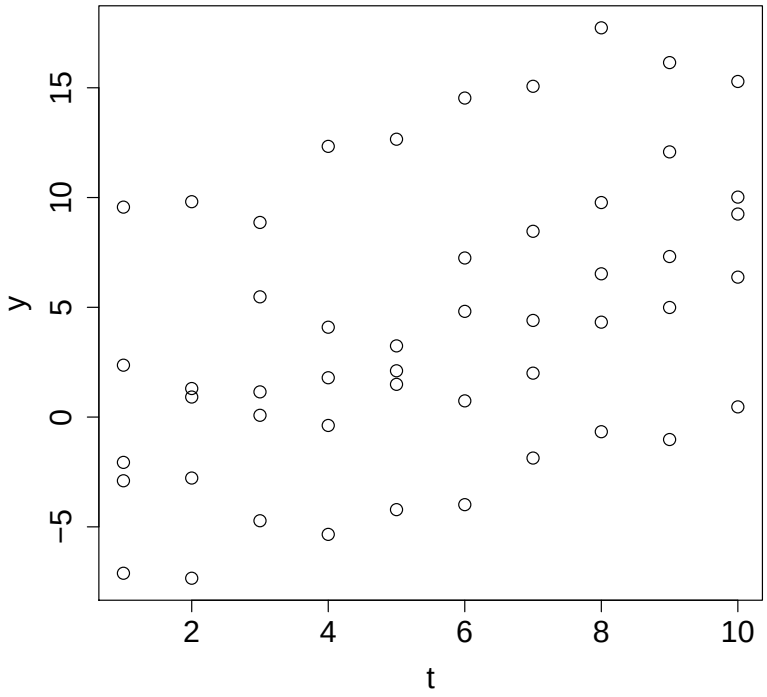
$$Y_{ij} = \beta_0 + \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \varepsilon_{ij}.$$

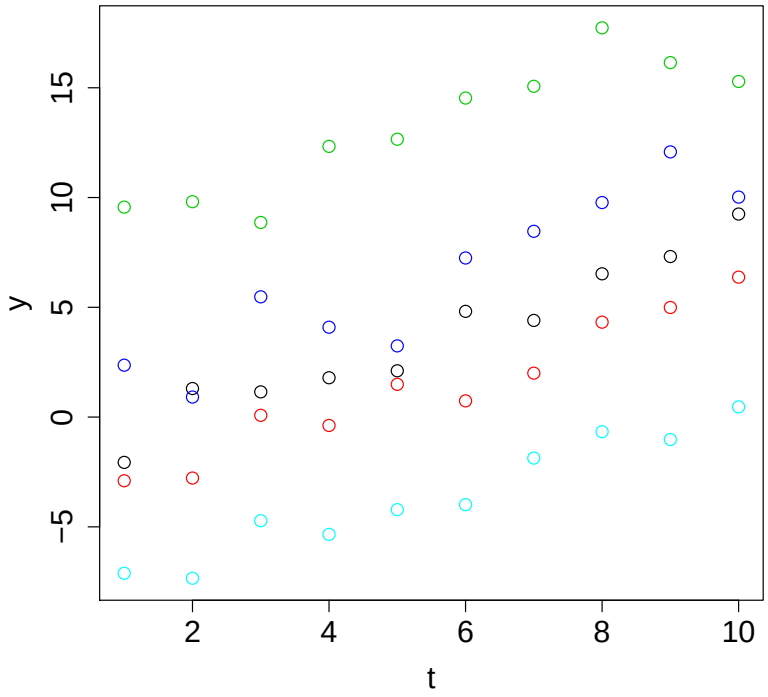
Without longitudinal information, we have to assume $\beta_C = \beta_L$. This is a strong assumption!

(E.g. $\beta_C =$ Increase in average starting salaries, $\beta_L =$ increase in salary after starting to work $\rightarrow$ opposite signs in our example. Or age vs. cohort effects.)

Advantages of longitudinal studies

Even if $\beta_C = \beta_L$, longitudinal studies are typically more powerful than cross-sectional studies to estimate β_L.





Advantages of longitudinal studies

Better protection against **confounding**: When looking at changes in the response, each subject can serve as its own control for time-constant variables such as age, gender, socio-economic background, education, genetics, disease history, ...

[**Confounder**: a variable that is associated with both the response and the covariate of interest and will lead to biased effect estimates if ignored.]

But even then, confounding is possible by time-varying variables (e.g. seasonality, long-term trends can be confounders for air pollution effects).

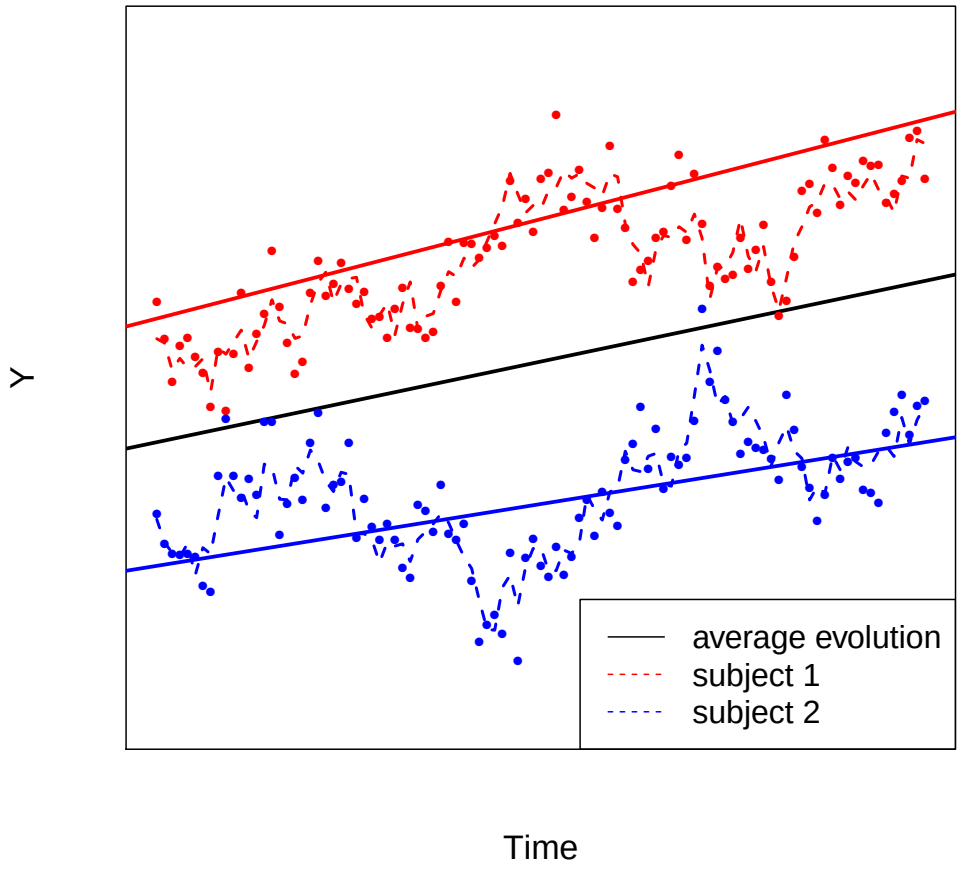
Advantages of longitudinal studies

We can distinguish different **sources of variation**:

- between subjects (**inter-subject variability**)
- within a subject over time (**intra-subject variability**)
- additional intra-individual variability due to e.g. **measurement error**

  (at least if we have repeated measurements at the same time)

Different sources of variability



Sources of variation in longitudinal data

Example: Hourly measurements of a bloodmarker:

- **Differences between people**: In average level and in average evolution over time. $\longrightarrow b_{0i}, b_{1i}$
- **Within a person over time**: Serial correlation due to e.g. long half-life of blood-marker, longer-term influences (alcohol, . . .) etc. $\longrightarrow \epsilon_{ij}^{(1)}$
- **Measurement Error**: On top of that, we will probably not have exact measurements of the bloodmarker at time t. $\longrightarrow \epsilon_{ij}^{(2)}$

Possible model, with $\epsilon_{ij} = \epsilon_{ij}^{(1)} + \epsilon_{ij}^{(2)}$, $\epsilon_{ij}^{(1)}$ auto-correlated, $\epsilon_{ij}^{(2)}$ i.i.d. error:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij}^{(1)} + \epsilon_{ij}^{(2)}.$$

Often, we cannot estimate all of these well (especially $\epsilon_{ij}^{(1)}$ and $\epsilon_{ij}^{(2)}$).

What is special about longitudinal data?

- Observations on the same subject tend to be more similar than observations on different subjects and are thus not independent. They are (marginally) **correlated**.

- Observations have an ordering in time. (In contrast, for example, to clustered data, e.g. measurements on litters of mice.)

- Often observations closer in time are more similar than observations taken further apart, i.e. the correlation is decreasing with the time difference. $\longrightarrow$ Difference to other repeated measures data.

- Missing data are common, e.g. because of drop-out, which is methodologically challenging.

Some challenges in longitudinal data

- Appropriate modeling of correlation structure.
- There has been a lot of development in recent years, but flexibility and robustness of software can still be an issue.
- Missing values are a challenge and constistute a problem depending on the missing data mechanism and the model used (more in Chapter 11).

Some challenges in longitudinal data

Time-varying covariates are challenging:

- determining an appropriate **lag structure** of covariate effects. Examples:
  - does air pollution increase mortality immediately? After hours? Days? Cumulatively?
  - carry-over effects in cross-over trials

- **covariate endogeneity** when the response predicts the covariate values at later times. Examples:
  - the treatment is changed when the response values indicate that the patient is not responding
  - in a study on the effects of physical activity in reducing blood glucose levels in patients with type 2 diabetes: if patients with high blood glucose levels at one visit increase their physical activity subsequently (feeling guilty! - **feedback** mechanism).

More in Chapter 12.

Longitudinal and other data

- **Multivariate data**: (Balanced) longitudinal data can be viewed as a type of multivariate data. But with a special correlation structure!
- **Hierarchical / multi-level / clustered data**: Similar nested structure and approaches (random effects etc.), but without the temporal structure.
- **Spatial data**: 2-D / 3-D, no inherent ordering, usually no independent subunits. But many similar approaches to modeling correlation: Marginal models, Gaussian random effects / fields, Markov chains / random fields
- **Functional data**: There are approaches to model longitudinal data as functional data (in time). $\rightarrow$ more at the end, if time

Analysis of longitudinal data vs. time series analysis

- As in time series analysis, longitudinal data analysis tries to take into account **correlation** between observations on the same subject and to model the time course during the analysis.

- In contrast to classical time series analysis, the focus in longitudinal data analysis is usually on the estimation of **covariate effects**.

- Longitudinal data typically span shorter time periods than time series do. But: In longitudinal data, we have (typically) independent replications in the form of subjects, which allows us to borrow strength (can be more robust to model assumptions).

- Many concepts from time series analysis are useful in longitudinal data analysis.

Connections to further classes

- Linear models
- Generalized linear models
- Lifetime data analysis
- Stochastic processes
- Biostatistical methods

Typical questions with longitudinal data

- Are there changes over time?
- If so, which shape do they take? Is the shape linear? Are there break points?
- Do the changes over time depend on covariates? E.g. does the change differ between treatment groups or depend on gender or age?
- Are changes associated with the baseline value at $t = 0$?
- How large is the intra-individual variability compared to the inter-individual variability?

Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

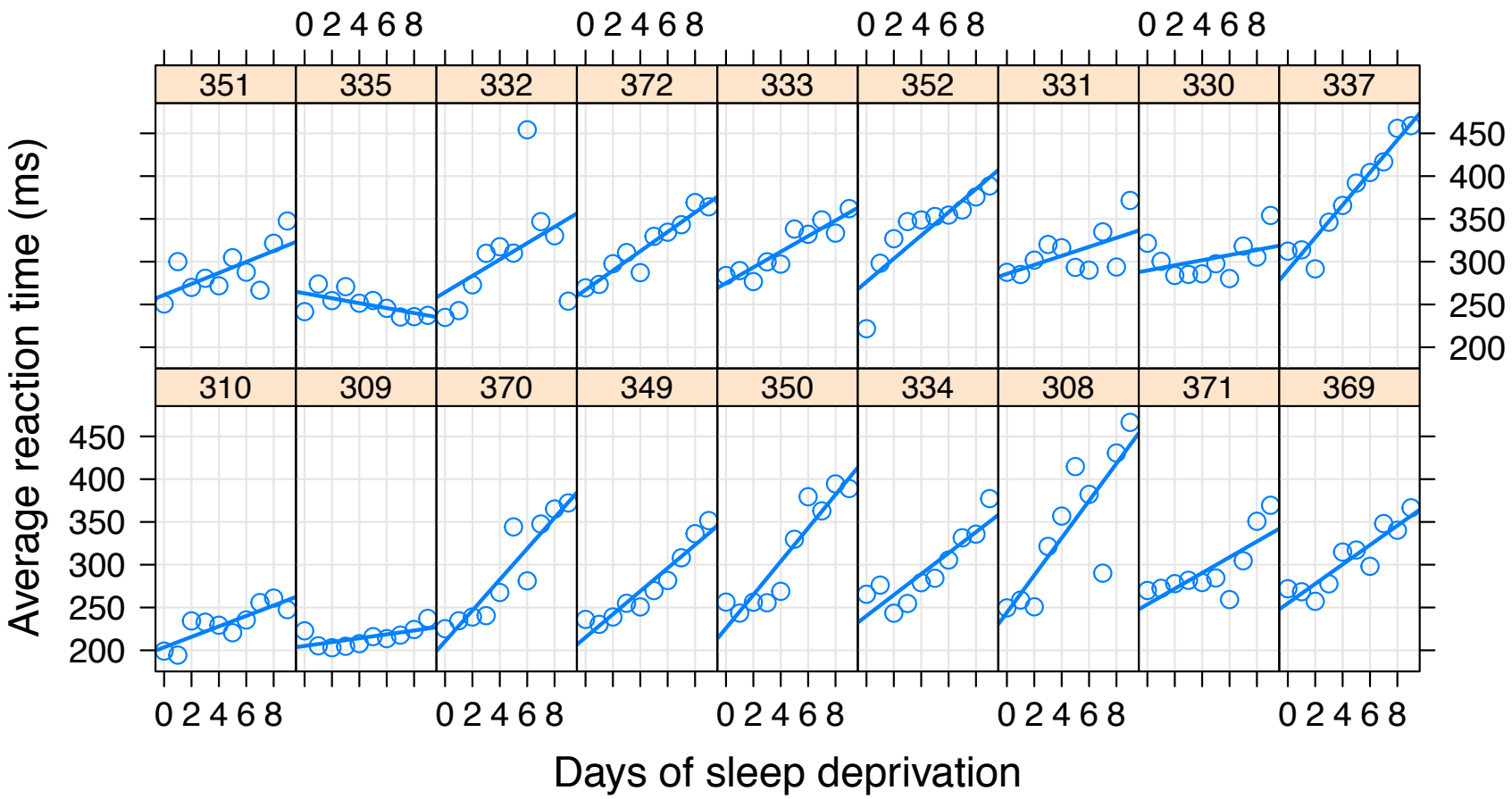
1.2 Examples

1.3 Correlation and modeling approaches

Example 1: Sleep deprivation study

- Sleep deprivation study with daily measurements from day 0 (normal sleep) to day 8 (3 hours sleep per night on subsequent nights) for $N = 18$ subjects.
- Response: average reaction time (in milliseconds, ms) on a series of tests
- No missings, balanced and equally spaced data
- First analyzed in Belenkey et al (2003), re-analyzed in Bates et al (2014) and part of the R-package `lme4`.

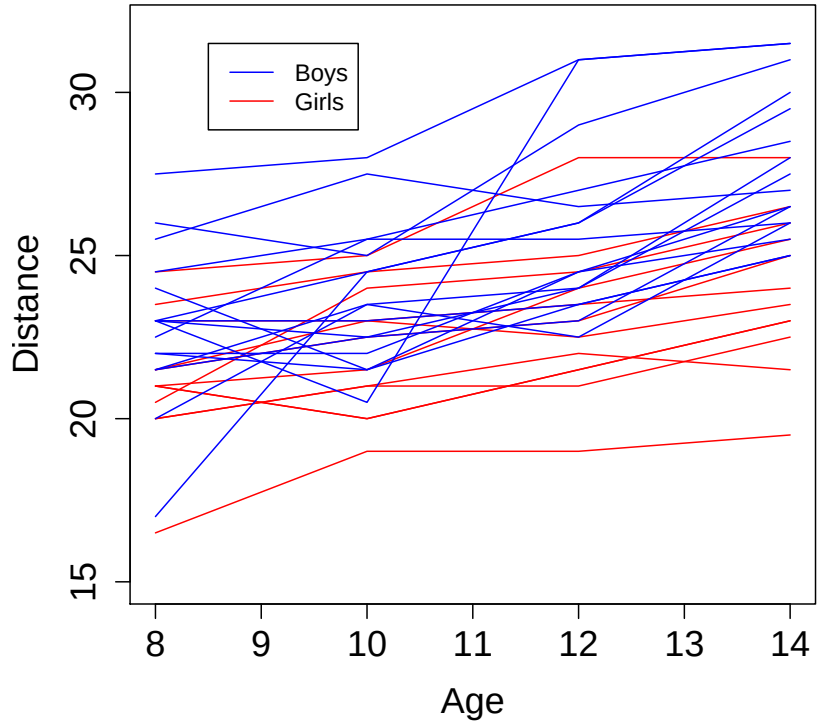
Example 1: Sleep deprivation study



Example 2: Growth in children (orthodont data)

- Data from Potthoff and Roy (1964), re-analyzed in the book by Little and Rubin (1987)
- 11 girls, 16 boys
- Response: distance between two points in the face
- 4 measurements at the ages 8, 10, 12, 14 (balanced data)
- **Questions of interest:** Comparison of intercept and slope between boys and girls. Heterogeneity between subjects?

Example 2: Growth in children (orthodont data)

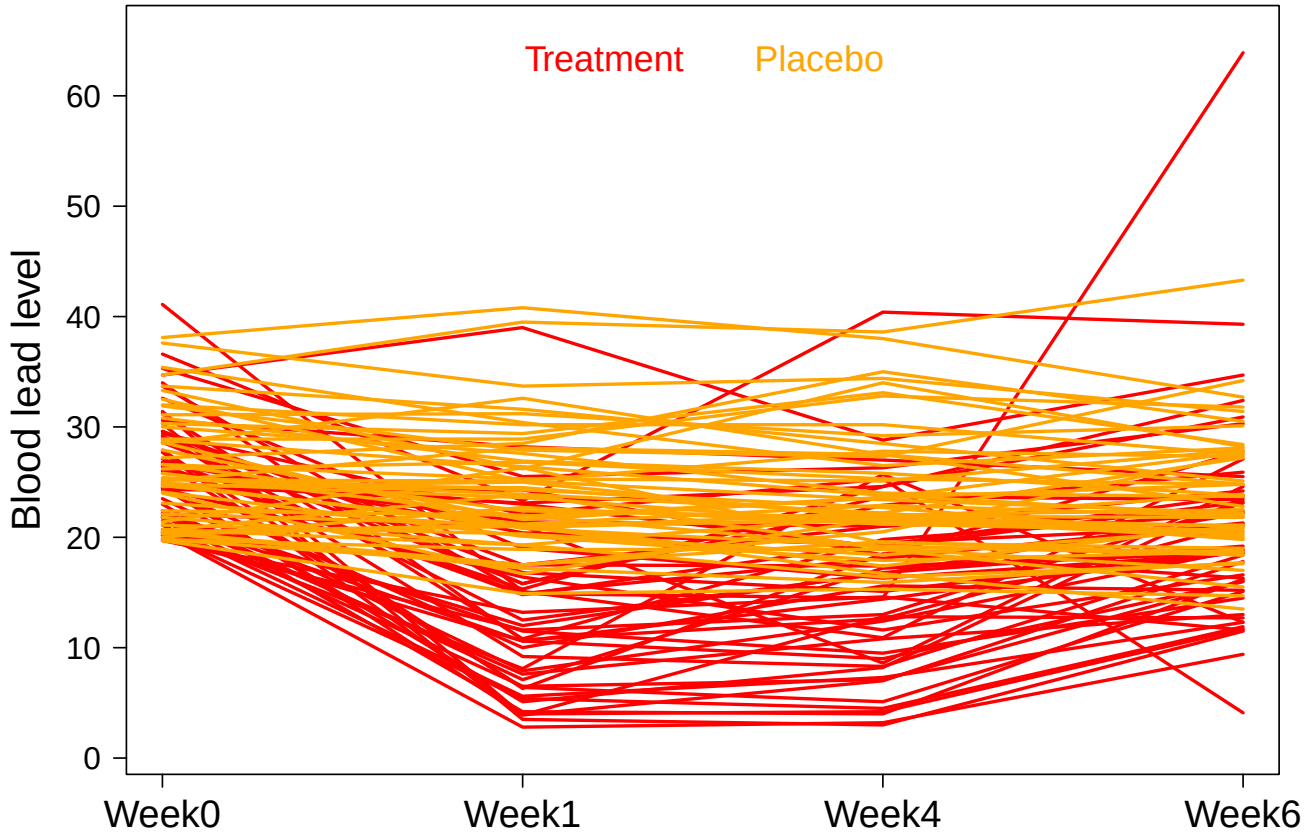


Example 3: Treatment of lead-exposed children (TLC) trial

Background: US children can be exposed to lead by exposure to lead paint dust and by eating lead paint chips, both of which can occur in deteriorating housing with lead-based paint from before 1978 (when the US government banned it). High blood levels of lead results in risk of several adverse health effects.

The **TLC trial** enrolled children 12-33 months old with high blood lead levels. These received either a placebo or succimer, which enhances urinary excretion of lead. Measurements in this data set are for baseline, week 1, week 4 and week 6 for 100 children (see Fitzmaurice et al, 2004).

Example 3: TLC trial



Example 3: TLC trial - wide format

```
> leadwide <- read.table("lead.txt",
 col.names = c("id", "group", "week0", "week1", "week4", "week6"))
> head(leadwide)
  id group week0 week1 week4 week6
1  1     P  30.8  26.9  25.8  23.8
2  2     A  26.5  14.8  19.5  21.0
3  3     A  25.8  23.0  19.1  23.2
4  4     P  24.7  24.5  22.0  22.5
5  5     A  20.4   2.8   3.2   9.4
6  6     A  20.4   5.4   4.5  11.9
```

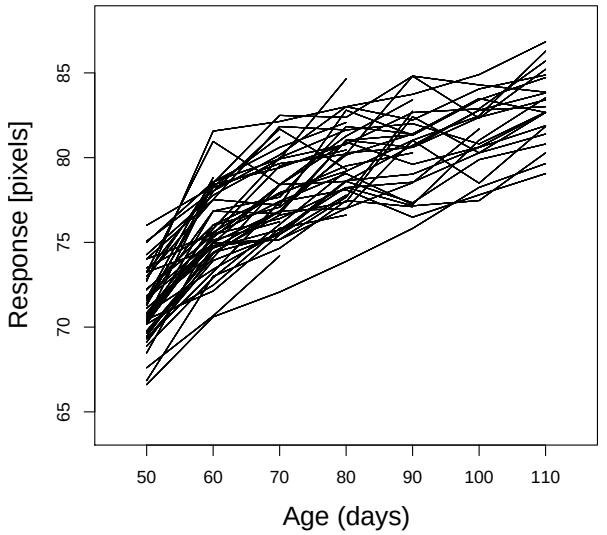
Example 3: TLC trial - long format

```
> lead <- reshape(leadwide, times = c(0,1,4,6),
        direction = "long", v.names = "lead",
        varying = c("week0", "week1", "week4", "week6"))
> lead <- lead[order(lead$id),]
> lead[1:9,]
    id group time lead
1.0  1     P    0 30.8
1.1  1     P    1 26.9
1.4  1     P    4 25.8
1.6  1     P    6 23.8
2.0  2     A    0 26.5
2.1  2     A    1 14.8
2.4  2     A    4 19.5
2.6  2     A    6 21.0
3.0  3     A    0 25.8
```

Example 4: Rats

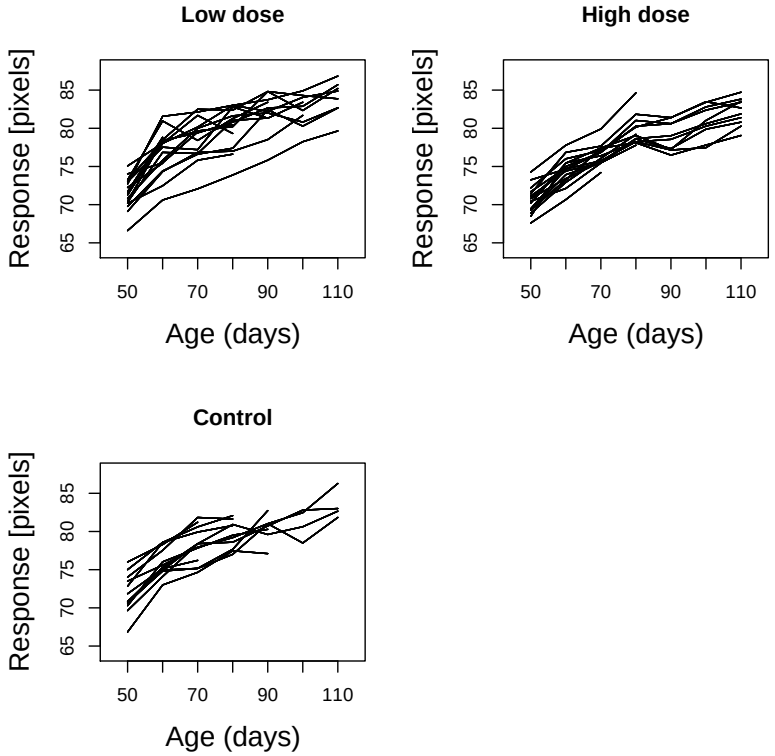
Question: Effect of an inhibitor for testosterone production in rats on their craniofacial growth (see Verbeke and Molenberghs, 2000).

- 50 male rats were randomized into three groups:
  - control
  - low dose
  - high dose
- **Response:** Measurement between two well-defined points on X-ray pictures of the skull, characterizing the height of the skull (in pixels)



Example 4: Rats

As the focus is on the effect of the treatment, it makes sense to look at the data by group:



Example 4: Rats

- The measurement time points t_j are the same for all rats.
- In contrast to the sleep study, growth and TLC data, the number of measurements is not equal for all subjects, as not all rats were observed up to 110 days $\longrightarrow$ dropout.
- This is due to not all rats surviving the anesthesia.

Example 5: CD4

- **Background:** The human immunodeficiency virus (HIV) destroys the CD4 cell, which is important in the body's immunoresponse. The CD4 cell count is thus a good indicator for the development of the disease. After seroconversion the number of CD4 cells decreases.

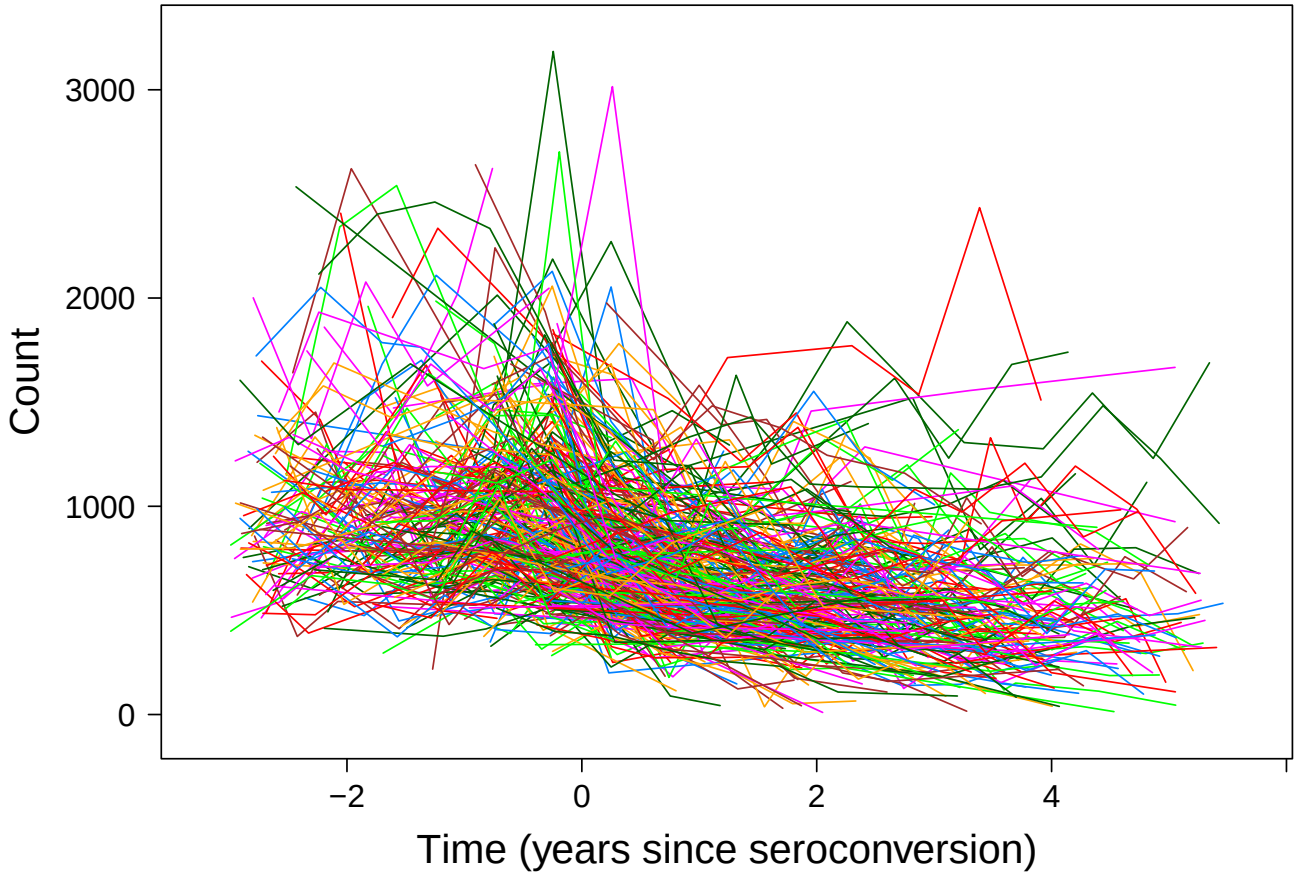
- The data with 2376 observations on 369 men infected with HIV is highly unbalanced (see Diggle et al, 2002).

- **Questions of interest:**
  - the average time course for the CD4 cell depletion
  - estimation of the time courses for individual men
  - heterogeneity between men
  - factors influencing the change in CD4 cell counts

Example 5: CD4 - variable description

- `ID`: subject ID
- `time`: time since seroconversion in years
- `count`: **CD4 cell count**
- `age`: age
- `cigarettes`: number of packs of cigarettes per day
- `drug`: illegal drugs (yes/no)
- `sexual`: number of sexual partners
- `mental`: psychological health score

Example 5: CD4



Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

1.2 Examples

1.3 Correlation and modeling approaches

Why are simple methods not adequate? Example orthodont data

Cross-sectional analysis (Question: Difference between genders?)

- Compare boys and girls at each age with a t-test for independent samples or with a Mann-Whitney-Test.

- **Problems:**
  - Handling of multiple testing issue?
  - At each single time point there is less data and thus less power.
  - Comparison between different ages?
  - Information about changes over time is not considered.
  - (Only works if time points are the same for each child.)

Why are simple methods not adequate?

Analysis stratified by gender (Question: Change over time?)

- Compare the mean at age 8 with the mean at age 10, 12, 14 with a t-test for paired samples.
- **Problems:**
  - Handling of multiple testing issue?
  - For each comparison there is less data and thus less power.
  - Characterization of the change?
  - Comparison between genders?
  - (Only works if time points are the same for each child.)

Why are simple methods not adequate?

Linear regression 1 (Question: Change over time + difference between genders?)

- Estimate for each subject a linear regression model with covariate age. Compare the subject-specific regression coefficients between boys and girls.

- **Problems:**
  - (There may not be enough observations per subject to estimate a regression model.)
  - The (varying) uncertainty in the estimated regression coefficients is not considered in the second step.

  $\longrightarrow$ One would like to do everything at once.

Why are simple methods not adequate?

Linear regression 2 (Question: Change over time + difference between genders?)

- For all data, estimate a linear regression model with covariates gender, age and their interaction.

- **Problem:** One important assumption in linear regression is that observations are independent. Here, observations on the same subject are correlated. Estimated standard errors (and thus confidence intervals, p-values) will be incorrect when ignoring this. We also lose efficiency in estimating the mean parameters.

Why are simple methods not adequate?

In addition to incorrect inference when ignored, the correlation and/or decomposition of the variance is sometimes actually of scientific interest.

Example: New bloodmarker - how large is the

- measurement error $\longrightarrow$ how precisely can the lab measure?
- intra-subject variability $\longrightarrow$ is one measurement representative of the subject's average marker level?
- inter-subject variability $\longrightarrow$ is a general reference value useful in detecting individual unusual results?

Different viewpoints of correlation

- **Marginal models**: Marginally, observations are correlated. We can model this and/or account for it with robust standard errors (GEE).
- **Mixed models**: Observations are correlated, because they are from the same subject and share the same underlying processes. Conditional on these, observations are independent.
  (Or residual serial correlation is left, then additionally model that.)
- **Transition/Markov models**: Observations are correlated, because the past influences the presence.
  (Typical here: Past = last q observations $\longrightarrow$ Markov property.)

We can model correlation or try to explain it, depending on our objectives.
Shift: Covariance structure $\longrightarrow$ mean structure.

The three approaches for the linear model

Consider a simple linear regression model (e.g. for child growth, $Y =$ height)

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \epsilon_{ij}.$$

Marginal model: Specifies a model for mean (population average), variance and correlation (between measurements on the same subject), e.g.

$$\begin{aligned} \mathsf{E}(Y_{ij}) = \mu_{ij} &= \beta_0 + \beta_1 t_{ij} \\ \mathsf{Var}(Y_{ij}) &= \sigma^2 \\ \mathsf{Corr}(Y_{ij}, Y_{ik}) &= \rho(\boldsymbol{\alpha}) \end{aligned}$$

The three approaches for the linear model

Mixed model: Models curves with subject-specific means, e.g.

$$Y_{ij} = (\beta_0^* + b_{i0}) + (\beta_1^* + b_{i1})t_{ij} + \epsilon_{ij}$$

$$(b_{i0}, b_{1i})^T \overset{iid}{\sim} \left(\mathbf{0}, \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}\right) \text{ ind. of } \epsilon_{ij} \overset{iid}{\sim} (0, \sigma^2)$$

Transition model: Models the present in terms of the past, e.g. $(q = 1)$

$$\begin{aligned} Y_{ij} &= \beta_0^{**} + \beta_1^{**} t_{ij} + \epsilon_{ij} \\ \epsilon_{ij} &= \alpha\epsilon_{ij-1} + \xi_{ij}, \quad \xi_{ij} \overset{iid}{\sim} (0, \tau^2), \epsilon_{i1} \sim (0, \sigma^2) \end{aligned}$$

The three approaches for the linear model

Note that in the linear model, the β parameters in all three approaches have a marginal interpretation, i.e. in all three models do they measure the unit increase in the mean of Y for a unit increase in the corresponding covariate. In our example:

$$\begin{aligned} \textbf{Marginal model}: \quad & \mathsf{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} \\ \textbf{Mixed model}: \quad & \mathsf{E}(Y_{ij}) = \beta_0^* + \beta_1^* t_{ij} \\ \textbf{Transition model}: \quad & \mathsf{E}(Y_{ij}) = \beta_0^{**} + \beta_1^{**} t_{ij} \end{aligned}$$

This is no longer the case in the generalized setting, see Chapter 8.

The three approaches for the linear model

Also, the transition and the linear mixed model imply certain marginal models with particular correlation structures, see Chapters 3.5 and 6.1.

- For the linear case, we will focus on the linear mixed model, see Chapters 3-7. The generalized case is discussed in Chapter 9.

- Marginal models will be discussed for the generalized case in Chapter 10, but we will discuss choice of the correlation structure and robust standard errors for linear mixed models as well.

- If there is time, we may briefly discuss transition models in Chapter 12.

Outlook

1. Introduction
2. Exploring and displaying longitudinal data
3. The longitudinal linear mixed model
4. Estimation in the LLMM
5. Inference in the LLMM
6. Flexible extensions of the LMMM
7. Model building and model choice
8. Non-normal data
9. The generalized linear mixed model
10. Marginal models
11. Missing data
12. Selected topics