Analysis of Longitudinal Data

Sonja Greven, Almond Stöcker

Summer Term 2017

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

Language

- The lecture and lab are held in English.
- There will be German summaries of the last lecture at the beginning of each lecture.
- As an additional service, we offer a voluntary Tutorial in German.
- There will be German and English versions of the Übungsblätter/work sheets and the exam.
- You can always ask questions in German.

Dates

- Lecture (Prof. Dr. Sonja Greven): Thursday 14.15 - 15.50 (DZ 003) and ca. every second Tuesday, 12.15-13.50 (W 101, Lehrturm)
- Lab session (Almond Stöcker):
  ca. every second Tuesday, 12.15-13.45 (W 101, Lehrturm)
- Tutorial (Johanna Völkl):
  ca. every second Tuesday, 8.15-9.45 (Cip-Pool 42, Ludwigstr. 33)
- Up-to-date times and rooms are on the course website (code 'REML')
  `https://moodle.lmu.de/course/view.php?id=1429`.
- Sprechstunde / consultation times: by appointment

Exam

- 90 minutes long. Our proposal: 8.8.2017.
- Lectures and lab sessions are both relevant for the exam. The Tutorial is additional voluntary opportunity for repetition and exercise.
- 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).
- Exercises will contain **old exam** questions.

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 will be on the website for (voluntary) further reading.

ARSnova

- ARS stands for 'Audience Response System' and will be used for live voting and live feedback.
- Browser-based, no installation or registration required.
- Link or QR code: `https://arsnova.eu/mobile/#id/56421973`

Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

1.2 Examples

1.3 Advantages of longitudinal data

1.4 Challenges of longitudinal data

1.5 Correlation and modeling approaches

What are longitudinal data?

For **repeated measures data**, the variable of interest is measured repeatedly for the same subjects under different conditions. Example: heart rate measurements for several subjects after different exercises.

Longitudinal data are a type of repeated measures data, for which the variable of interest is measured for several subjects **repeatedly over time**. Example: heart rate measurements for several subjects over 12 months.

[We will use the term "subject" for convenience, even if the independent **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**.

Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

1.2 Examples

1.3 Advantages of longitudinal data

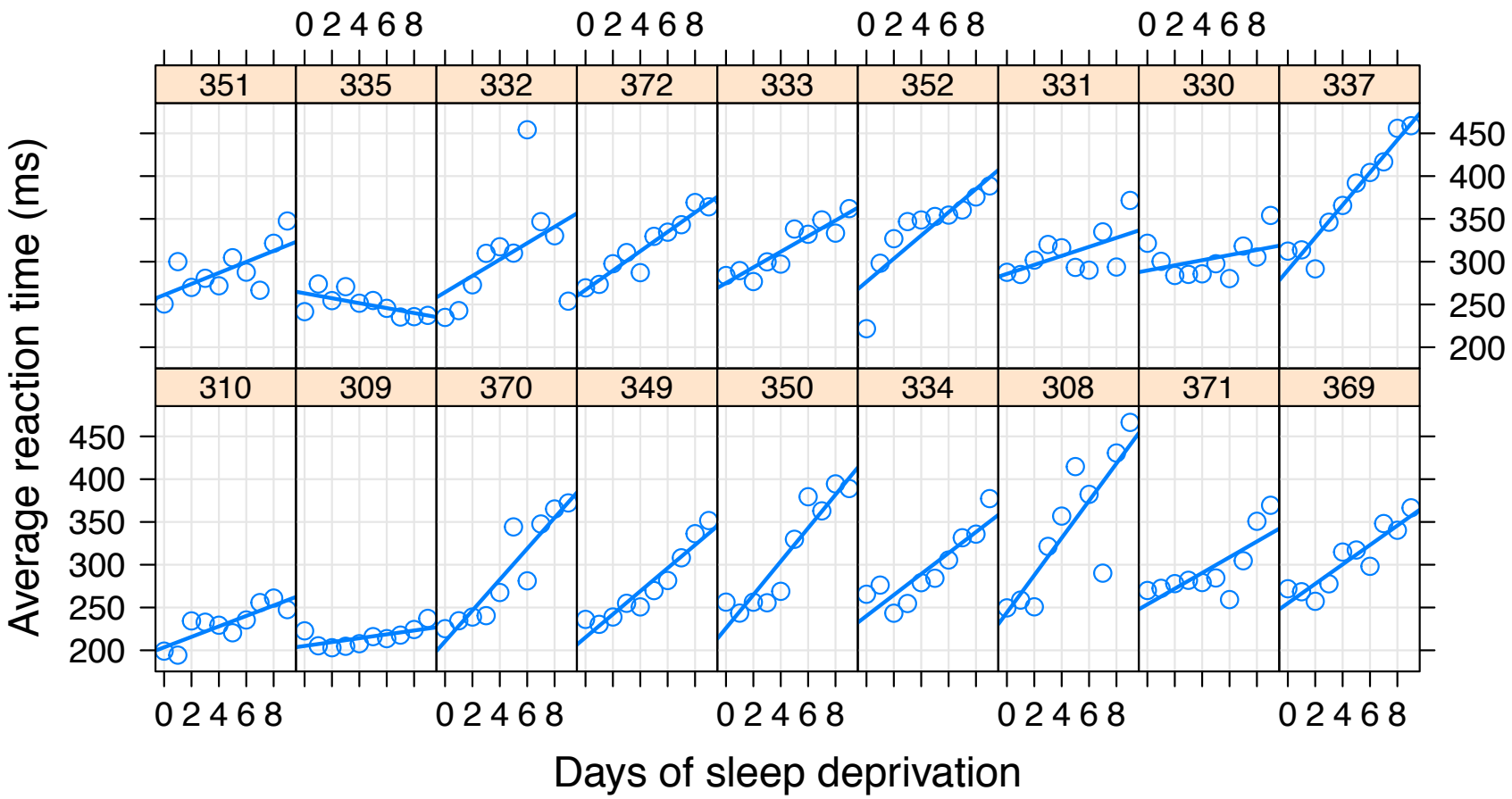
1.4 Challenges of longitudinal data

1.5 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, time is only covariate
- First analyzed in Belenkey et al (2003), re-analyzed in Bates et al (2014) and part of the R-package `lme4`.

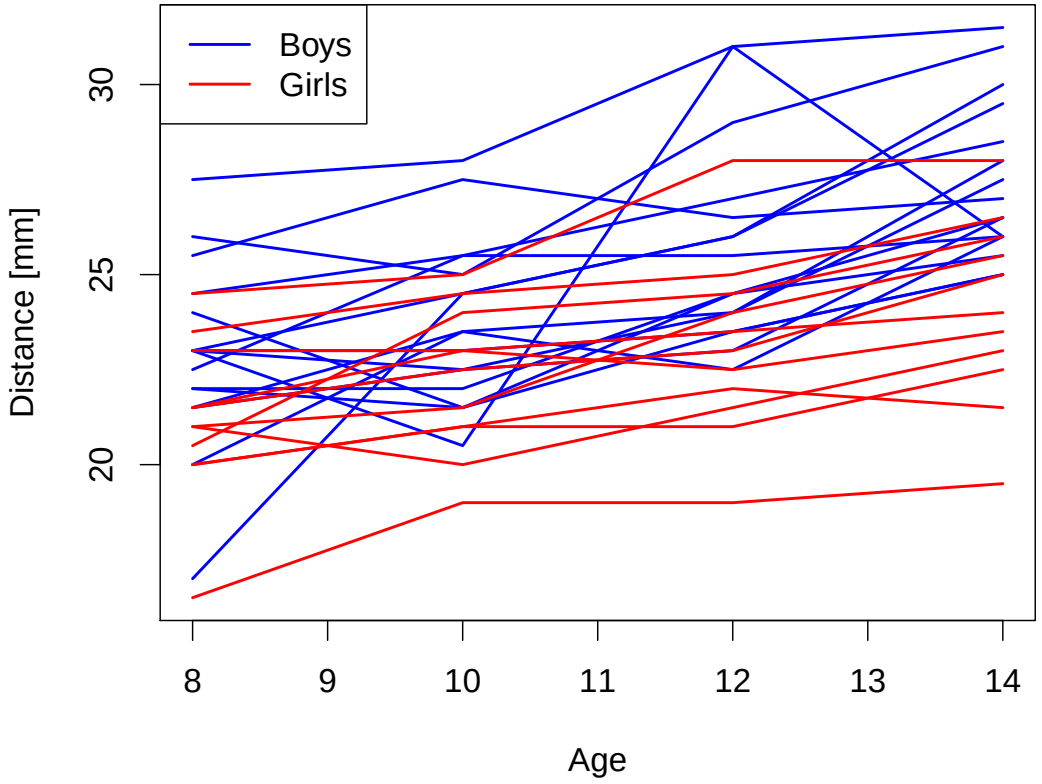
Example 1: Sleep deprivation study



Example 2: Growth in children (Orthodont data)

- Data from Potthoff and Roy (1964), re-analyzed in the book by Little and Rubin (1987) and part of the R-package `nlme`.
- $N = 27$, 11 girls, 16 boys
- Response: distance between two points in the face (in mm)
- 4 measurements at the ages 8, 10, 12, 14 (balanced data, equally spaced)
- **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)

Background: US children can be exposed to lead-based paint in deteriorating housing from before 1978 (when the paint was banned). High blood levels of lead result in risk of several adverse health effects.

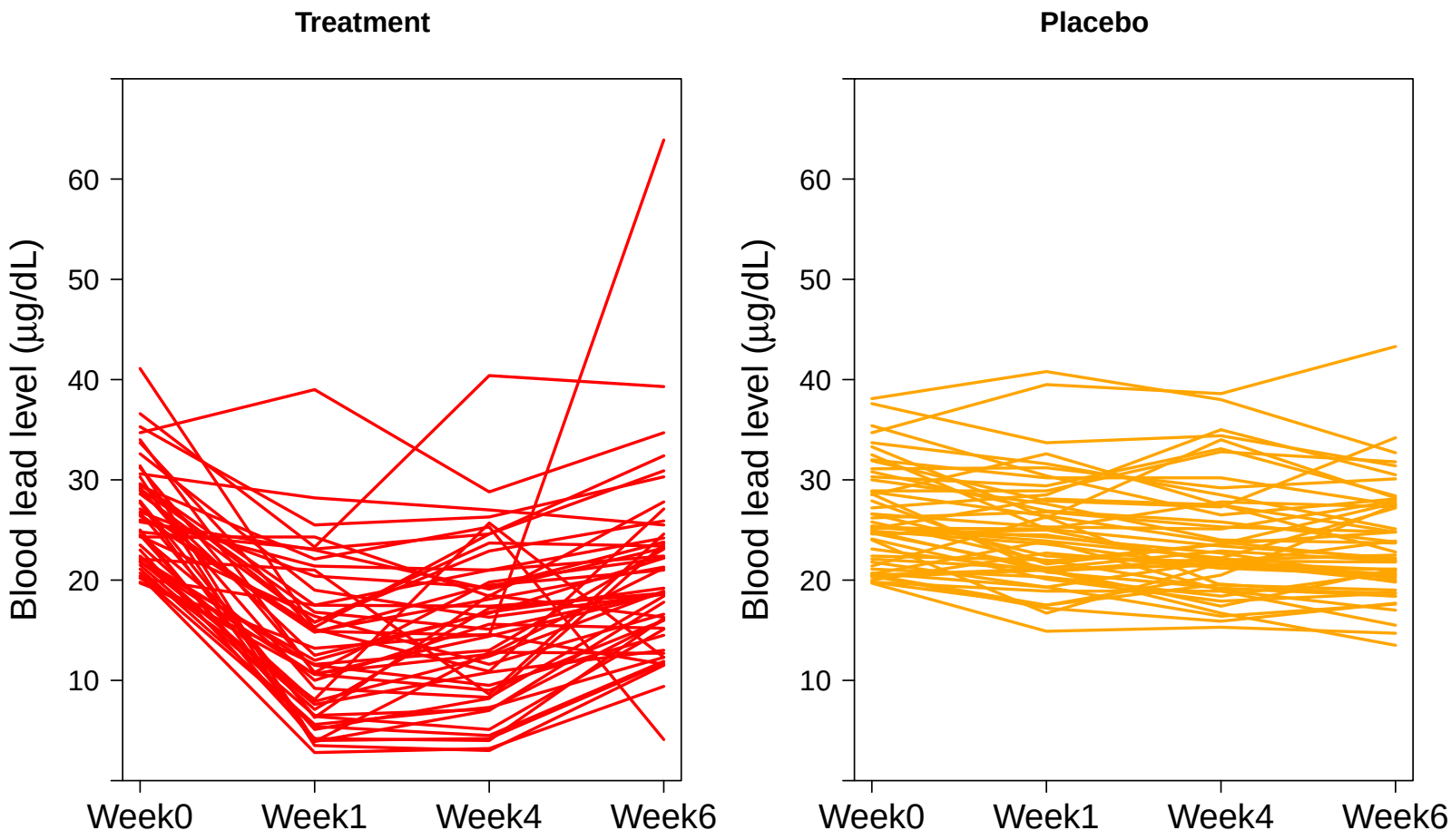
TLC trial (see Fitzmaurice et al, 2004): In this data set

- N=100 children 12-33 months old with high blood lead levels
- Response: Blood lead level (μg/dL)
- Treatment: placebo or succimer (enhances urinary excretion of lead)
- Measurements: baseline, week 1, week 4 and week 6 (balanced data).

Question: Is the treatment effective?

Data source: `http://www.hsph.harvard.edu/fitzmaur/ala/tlc.txt`

Example 3: TLC trial

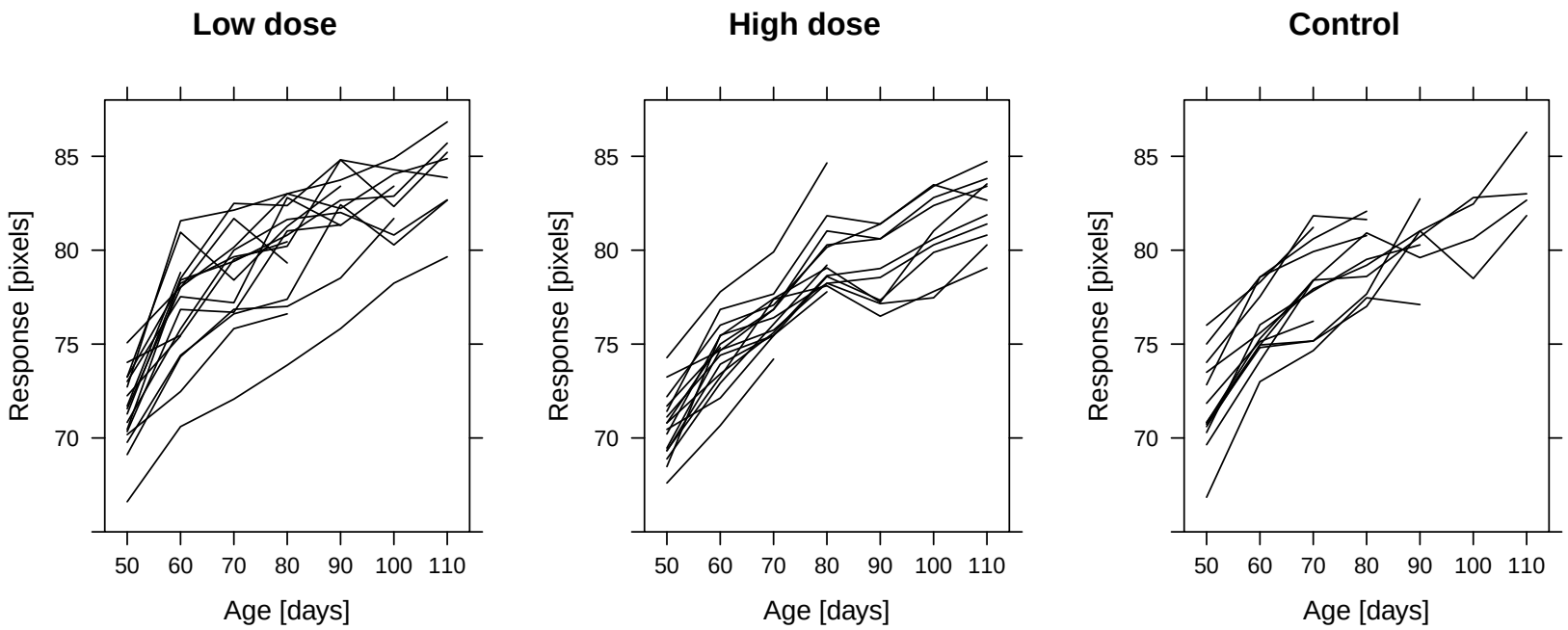


Example 4: Rats

Question: Effect of an inhibitor for testosterone production in rats on their craniofacial growth (see Verbeke and Molenberghs, 2000, `https://perswww.kuleuven.be/~u0018341/documents/rats.sas`).

- $N = 50$ male rats
- randomized into three groups: control, low dose, high dose
- **Response:** Distance between two well-defined points on X-ray pictures of the skull, characterizing the height of the skull (in pixels)
- Same measurement times t_j for all rats, but **dropout** due to rats not surviving the anesthesia (unequal n_i).

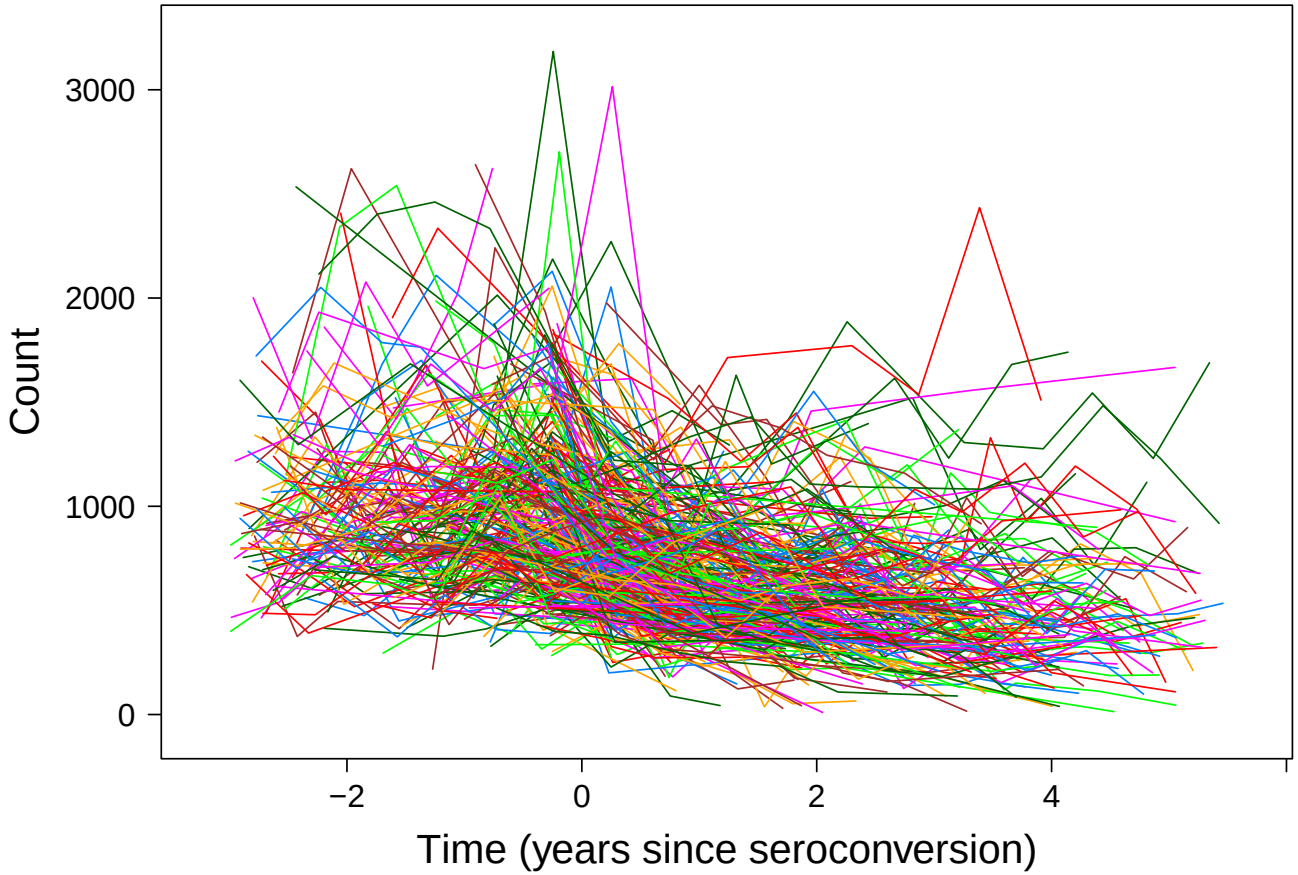
Example 4: Rats



Example 5: CD4

- **Background:** The human immunodeficiency virus (HIV) destroys CD4 cells, which are important in the body's immunoresponse. The CD4 cell count decreases after seroconversion (when anti-HIV antibodies develop, "HIV positive") and is a good indicator for the disease development.
- 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
  - time courses for individual men
  - heterogeneity between men
  - factors influencing the CD4 cell count change (info on age, cigarette and drug use, number of sexual partners, psychological health)

Example 5: CD4



Some observations on longitudinal data

- Covariates can be
  - **time-invariant** and only measured at baseline, e.g. gender or treatment for Examples 2-4.
  - **time-varying** and measured over time, e.g. cigarette and drug use in the CD4 data.

- Sometimes longitudinal data is measured together with **survival / time-to-event data** (more in Chapter 12). It can also be relevant to consider **dropout**, when subjects leave the study, as the event, as this might be related to the longitudinal outcome (e.g. in the CD4 data).

- 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, such as when subjects who developed a disease better remember risk factors than healthy controls). All considered examples are prospective studies.

Typical questions with longitudinal data

- Are there changes over time (e.g. sleep study)?
- If so, which shape do they take? Linear (e.g. growth in children)? Are there break points (e.g. TLC trial)?
- Do changes depend on covariates, e.g. on treatment group or gender (e.g. growth in children, rat data, TLC trial)?
- Are changes associated with the baseline value at $t = 0$ (e.g. CD4 data)?
- How large is the intra-individual variability compared to the inter-individual variability (e.g. CD4 data)?

Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

1.2 Examples

1.3 Advantages of longitudinal data

1.4 Challenges of longitudinal data

1.5 Correlation and modeling approaches

Advantages of longitudinal studies: Sources of variation

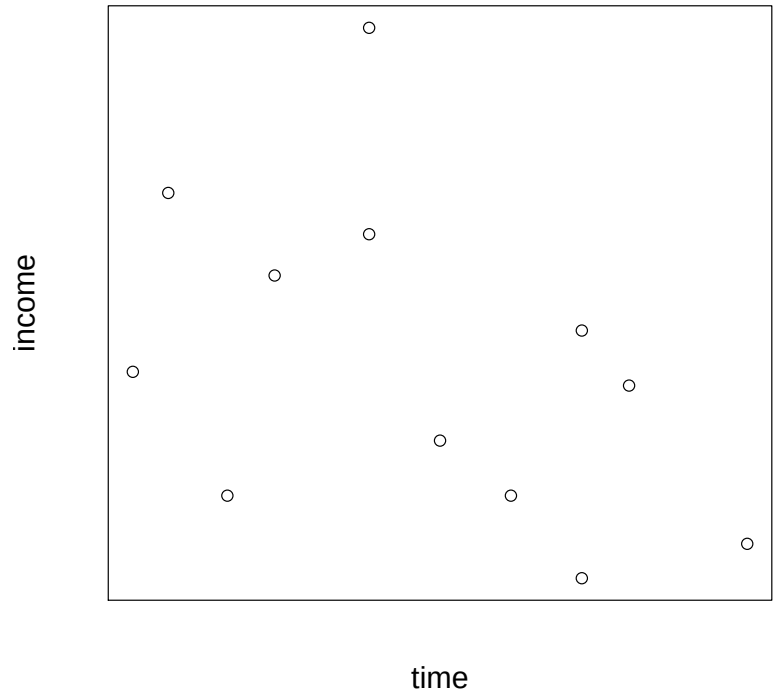
We can distinguish different **sources of variation**:

- differences between subjects (**inter-subject variability**)
- changes/variability within a subject over time (**intra-subject variability**).

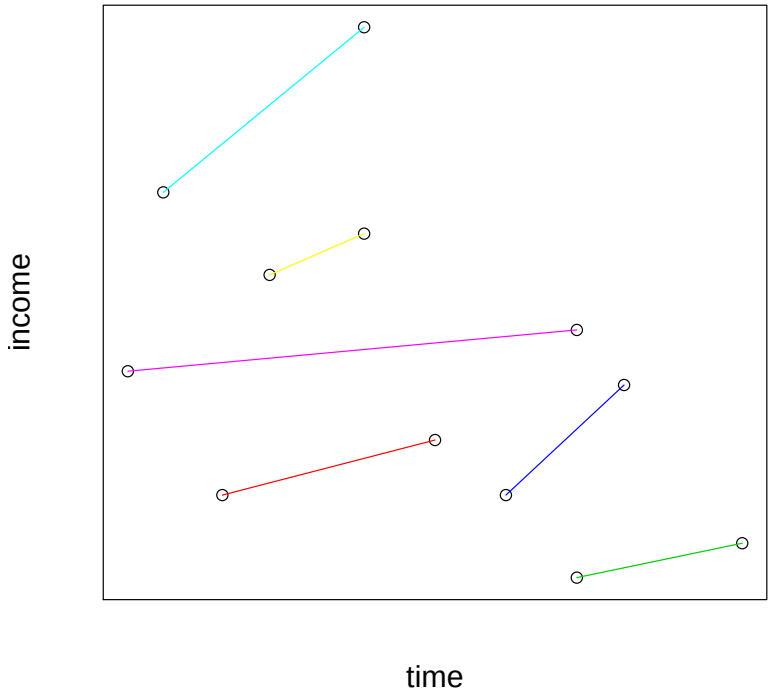
Looking at heterogeneity between subjects is directly of interest e.g. in the CD4 data.

Advantages of longitudinal studies: Distinguishing effects

We can distinguish changes over time within an individual (ageing effects) from differences in baseline levels between people (cohort effects).



Advantages of longitudinal studies: Distinguishing effects



Income is increasing over time for each person ("ageing" effect).

Starting salaries seem to be decreasing over time (cohort effect).

Advantages of longitudinal studies: Distinguishing effects

Longitudinal studies can follow individual change over time and are thus more informative than cross-sectional studies ($n_i = 1$).

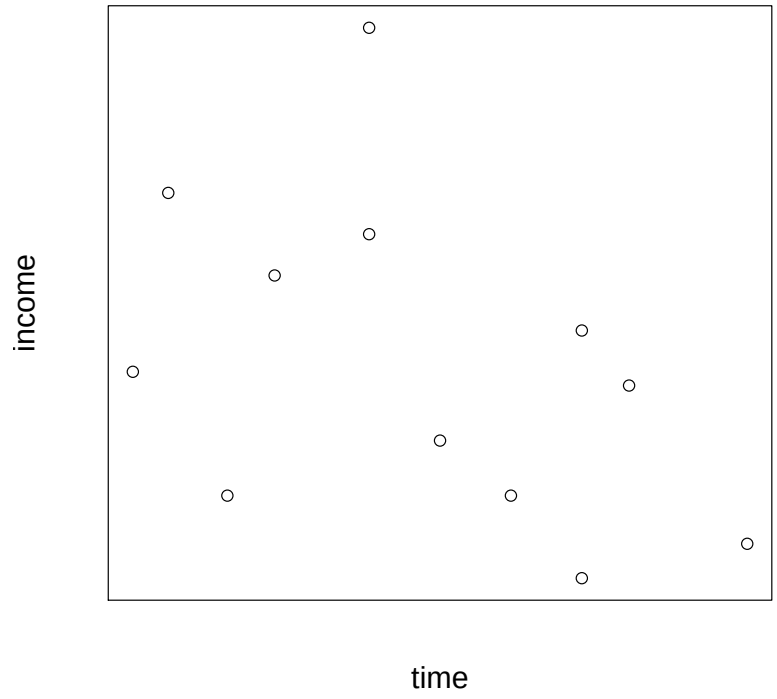
We can distinguish **ageing effects** and **cohort effects**, e.g.

$$\mathsf{E}[Y_{ij}] = \beta_0 + \beta_C t_{i1} + \beta_L (t_{ij} - t_{i1}),$$

- $\beta_C =$ increase in average starting salaries per year
- $\beta_L =$ average increase in salary per year after starting to work.

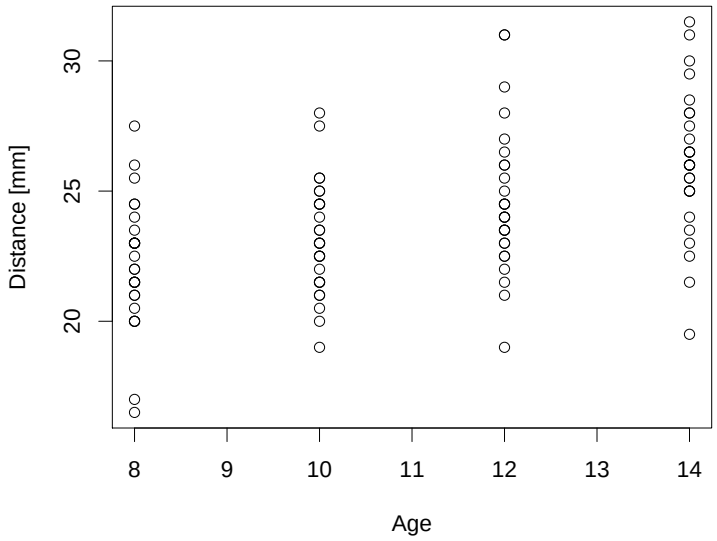
Note that $\mathsf{E}[Y_{ij} - Y_{ik}] = \beta_L (t_{ij} - t_{ik})$, i.e. **changes** in Y for subject i when t changes contribute to the estimation of β_L.

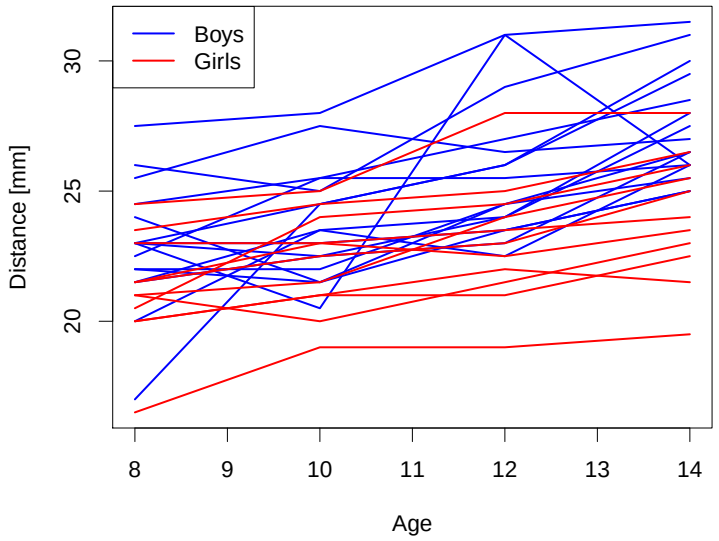
In a cross-sectional study, without longitudinal information, we have to assume $\beta_C = \beta_L$. This is a strong assumption! In our example, β_C and β_L have opposite signs.



Advantages of longitudinal studies: Power

Even if $\beta_C = \beta_L$, longitudinal studies are typically more powerful than cross-sectional studies to estimate β_L. E.g. in the Orthodont data, β_L can be more precisely estimated as there is less variability in the slopes per subject than there is overall in the data.





Advantages of longitudinal studies: Confounding

Consider a time-varying covariate x_{ij} of interest and a time-constant confounder z_i and say we have

$$\mathsf{E}[Y_{ij}] = \beta_0 + \beta_1 x_{ij} + \beta_2 z_i \quad \Rightarrow \quad \mathsf{E}[Y_{ij} - Y_{ik}] = \beta_1 (x_{ij} - x_{ik}).$$

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

Note that z_i does not appear in the mean of the **change** in Y. Longitudinal studies thus offer better **protection against confounding**: For changes in the response, each subject serves as its own control for time-constant variables such as age, gender, socio-economic background, education, genetics, disease history, ...

Similarly, if we include an intercept per subject in our model, this offers protection against any time-constant confounders z_i: Using

$$\mathsf{E}[Y_{ij}] = \beta_i + \beta_1 x_{ij} \quad \Rightarrow \quad \mathsf{E}[Y_{ij} - Y_{ik}] = \beta_1 (x_{ij} - x_{ik})$$

β_i now captures the effect of $\beta_0 + \beta_2 z_i$.

Note however, that confounding is still possible by time-varying variables.

Confounding: Example air pollution study

Consider a study comparing cities with respect to their mortality counts and PM_{10} levels (particulate matter $< 10\mu m$) to determine whether higher PM_{10} levels increase mortality.

- Cross-sectional study comparing average PM_{10} levels and mortality:
  - Confounding by time-constant variables per city, e.g. different industrialization, poverty levels, climates, . . .
- Longitudinal study comparing daily PM_{10} levels and mortality counts:
  - No confounding by time-constant variables if each city is allowed their own average mortality level in the model.
  - Confounding by time-varying variables possible, e.g. seasonality (PM_{10} and influenza higher in the winter) and long-term trends.

Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

1.2 Examples

1.3 Advantages of longitudinal data

1.4 Challenges of longitudinal data

1.5 Correlation and modeling approaches

What is special about longitudinal data?

- Observations on the same subject are more similar than observations on different subjects. They are not independent, but **correlated**.
- Observations have an ordering in time.
- Often, observations are more similar the closer they are in time, i.e. the correlation is **decreasing with the time difference**. (In contrast to clustered data, e.g. on families.)
- Missing data are common, e.g. because of **drop-out**, when subjects leave the study.

Some challenges in longitudinal data

- Appropriate modeling of correlation structure (more in Chapters 3-10).
- There has been a lot of development in recent years, but flexibility and robustness of software can still be an issue.
- Missing values are methodologically challenging and constitute a problem depending on the missing data mechanism and the method used (more in Chapter 11).

Challenges in longitudinal data: Time-varying covariates

- 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 (**feedback** mechanisms). Examples:
  - the treatment is changed when the response values indicate that the patient is not responding
  - patients in a study on the effects of physical activity on blood glucose levels increase their physical activity after high glucose measurements.

More in Chapter 12.

Overview Chapter 1 - Introduction

1.1 Introduction to longitudinal data

1.2 Examples

1.3 Advantages of longitudinal data

1.4 Challenges of longitudinal data

1.5 Correlation and modeling approaches

Why are simple methods not adequate?

Example Orthodont data. Question: Difference between genders in change over time? Possible naive approaches:

- Linear regression model with covariates gender, age and their interaction.
  Problems:

- Cross-sectional analysis, comparison of boys and girls at each age
  Problems:

- Linear regression model with covariate age for each subject. Comparison of subject-specific regression coefficients between boys and girls.
  Problems:

Different viewpoints of correlation

- **Marginal models**: Model marginal correlation 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.
- **Transition/Markov models**: Observations are correlated, because the past influences the presence.
  (Typical here: Past $=$ last q observations $\rightarrow$ Markov property.)

The three approaches for the linear model

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

$$\mathsf{E}[Y_{ij}] = \beta_0 + \beta_1 t_{ij}. \tag{1}$$

- **Marginal model**: In addition to (1), specify a model for variance $\mathsf{Var}(Y_{ij})$ and correlation $\mathsf{Corr}(Y_{ij}, Y_{ik})$.

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

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

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

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

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

- In the linear model: Transition and linear mixed model imply marginal models with particular correlation structures (cf. Ch. 3.5 and 6.1). The β parameters in all three approaches have the same marginal interpretation. This is no longer the case in the generalized setting, see Ch. 8 ff.

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

- Marginal models are discussed for the generalized case in Chapter 10.

Outlook

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

Longitudinal and other 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
- Longitudinal data can be viewed as realizations of **stochastic processes**.

Analysis of longitudinal data (ALD) vs. time series analysis

- Both model time courses and try to take into account **temporal correlation** between observations.
- In contrast to time series analysis, ALD usually focuses on the estimation of **covariate effects**.
- Longitudinal data typically span shorter time periods than time series, but they contain **independent replications** in the form of subjects. This allows us to borrow strength (can be more robust to model assumptions).
- Many concepts from time series analysis are useful in ALD.