Longitudinal Analysis Workshop

Anteneh Tesema and Yebelay Berehan

Ethiopian Public Health Institute (EPHI)

National Data Management Center for Health (NDMC)

2023-08-10

Topics covered

1. **Day 1: Introduction to Longitudinal Analysis**
    - xx
    - xx
    - xx
2. **Day 2: Modelling of longitudinal data**
    - Linear Mixed Effects Models
    - Marginal Models
    - Estimation of Marginal model
3. **Day 3: Methods for Discrete Data**
    - Generalized Estimating Equations (GEE)
    - Generalized Linear Mixed Models (GLMM)
4. **Day 4: Addressing Missing Data in Longitudinal Studies**
    - Types of Missing Data Mechanisms (MCR, MR, NMR)
    - Handling Missing Data: Multiple Imputation
    - Handling Missing Data: Weighted GEE

Goal of the Workshop

The goal of this course is to:

- Provide an overview of fundamental statistical models and methods for the analysis of **longitudinal data**, including key theoretical results presented.
- Focus on the practical implementation of these methods in **R** .
- Help **trainees** gain a comprehensive understanding of the properties and use of modern methods for **longitudinal data analysis**.
- Enable trainees to pose scientific questions within the context of appropriate statistical models and **carry out and interpret analyses effectively**.

Primary Objectives of the Workshop

- Understand the effect of non-independence in longitudinal data.
- Recognize limitations of classical analysis methods in longitudinal studies.
- Explore and analyze the marginal distribution of longitudinal data.
- Learn and apply methods for analyzing continuous outcomes using linear mixed effects models.
- Utilize methods for analyzing discrete data in longitudinal studies using GEE and GGLMM.
- Gain knowledge about missing data mechanisms and techniques for handling them.

Day 1: Introduction to Longitudinal Analysis

- Understanding Non-Independence in Clustered Data
- Limitations of Classical Analysis Methods
- Exploratory Analysis of Marginal Distribution (Average Evolution, Variance Structure, and Correlation Structure)

Repeated Measures

- Statistical techniques like ANOVA and regression have a basic assumption that the residual or error terms are iid.
- In applied sciences, often confronted with the collection of correlated data.
- The term embraces a multitude of data structures such as Multivariate observations, Clustered data, repeated measurements, Longitudinal data & spatially correlated data.
- The distinguishing feature of repeated data is that they are correlated.

Familiar examples of clustered data

- Families, towns, litters, etc. are familiar examples of clustered data.
- In each of these examples, a cluster is a collection of subunits on which observations are made.
- Another form of clustering arises when data are measured repeatedly on the same unit.
- When these repeated measurements are taken over time, it is called a longitudinal study.
- When the correlation occurs over space, it is called a Spatial study.

Longitudinal data

- Special forms of repeated measurements.
- Longitudinal Studies: Studies in which individuals are measured repeatedly through time.
- Longitudinal data (LD) sets differ from Time Series data sets.
- LD: usually consists of a large number of a short series of time points.
- TS: usually consists of a single, long series of time points.
- Examples of LD:
  - Monthly CD4 count of a patient over time.
  - Psychological change of a patient.
  - The effect of a treatment to cure a disease over time.

Features of Longitudinal Data

- Defining feature of longitudinal studies is that measurements of the same individuals are taken repeatedly through time.
- Longitudinal studies allow direct study of change over time.
- Objective: primary goal is to characterize the change in response over time and factors that influence change.
- With repeated measures on individuals, we can capture within-individual change.
- In longitudinal studies, the outcome variable can be:
    - Continuous (e.g., blood lead levels).
    - Binary (e.g., presence/absence).
    - Count (e.g., number of epileptic seizures).
- The data set can be incomplete (missing data/dropout).
- Subjects may be measured at different occasions.

- In this module we will master a set of statistical tools that can handle all of these cases.
- Emphasis on concepts, model building, software, and interpretation.

Advantages of modern longitudinal methods

- You have much more flexibility in research design.
    - Not everyone needs the same rigid data collection schedule.
    - Not everyone needs the same number of measurements—can use all cases, even those with just one measurement!
- You can identify temporal patterns in the data.
    - Does the outcome increase, decrease, or remain stable over time?
    - Is the general pattern linear or non-linear?
    - Are there abrupt shifts at substantively interesting moments?
- You can include time-varying predictors.
- Can provide information about individual change.

- You can include interactions with time (to test whether a predictor's effect varies over time).
    - Some effects dissipate—they wear off.
    - Some effects increase—they become more important.
    - Some effects are especially pronounced at particular times.
- Can provide more efficient estimators than cross-sectional designs with the same number and pattern of observations.
- Can separate aging effects (changes over time within individuals) from cohort effects (differences between subjects at baseline) ⇒ cross-sectional design can't do this.

Challenges of Longitudinal Data Analysis

- Observations are not, by definition, independent → must account for dependency in data.
- Analysis methods not as well developed, especially for more sophisticated models.
- Difficulty of using state-of-the-art software.
- Computationally intensive.
- Unbalanced designs, missing data, attrition.
- Carry-over effects (when the repeated factor is condition or treatment, not time).

Recap:

- Longitudinal studies:
  - Measurements of the same individuals are taken repeatedly through time.
  - Allow direct study of change over time.
  - We can capture within-individual change.
- Objective: to characterize the change in response over time and factors that influence change.

Motivating Examples

The Jimma Infant Survival Data

- A follow-up study of newborn infants in Southwest Ethiopia.
- Wide ranges of data were collected on the following characteristics:
    - Basic demographic information.
    - Feeding practice.
    - Anthropometric measurements.
- Infants were followed for 12 months.
- Measurements were taken at seven time points from each child.

```
library(readxl)
Infant <- read_xls("Data/Infant.xls")
```

```r
library(dplyr)
output <- table(Infant$age) %>%
  as.data.frame() %>%
  mutate(Time = as.numeric(as.character(Var1)),
         N = Freq,
         Percentage =round((Freq/Freq[1])*100,1)) %>%
  select(Time, N, Percentage)
print(output)
```

```
##   Time   N Percentage
## 1    0 971      100.0
## 2    2 949       97.7
## 3    4 894       92.1
## 4    6 857       88.3
## 5    8 833       85.8
## 6   10 811       83.5
## 7   12 784       80.7
```

- Infants were followed during 12 months.
- Measurements were taken at seven time points from each child, resulting in a maximum of seven measurements per subject.
- For our purpose, we will consider the variable weight.
- Due to a variety of reasons, 80.7% continues up to the end of the study.
- The profile plot is produced by using the following R code.

The Income Dynamics (PSID) Study

- The Panel Study of Income Dynamics (PSID) began in 1968 and is still continuing.
- It is the longest-running longitudinal household survey in the world.
- The PSID is a longitudinal study of a representative sample of U.S. individuals.
- The data that represents a small subset of this data (1661 observations) is available in R software under the library "faraway".
- Variables included in the dataset: Age, Education (years of education), Sex, and Annual Income.

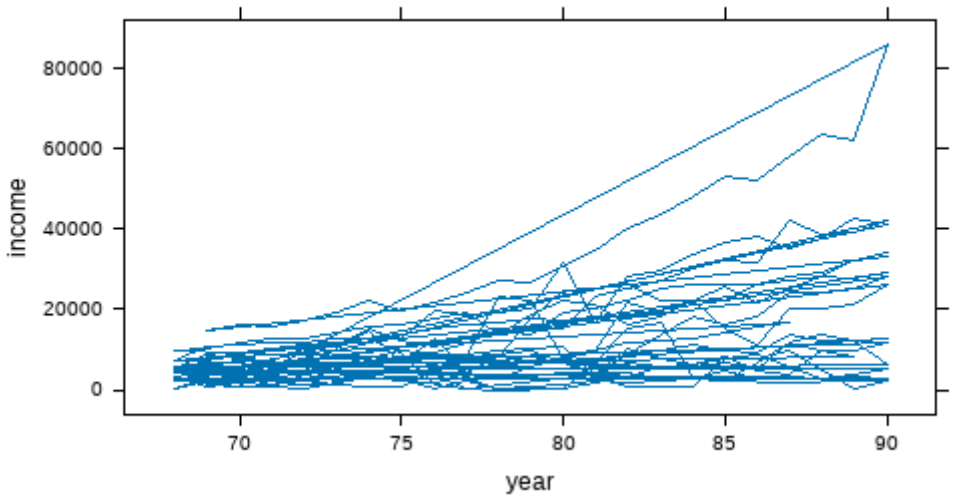
The Question of Interest

The PSID dataset raises the following questions:

- Is there a change in income over the years?
- Is there variation in income by sex?

PSID: Profile Plots for 20 Subjects

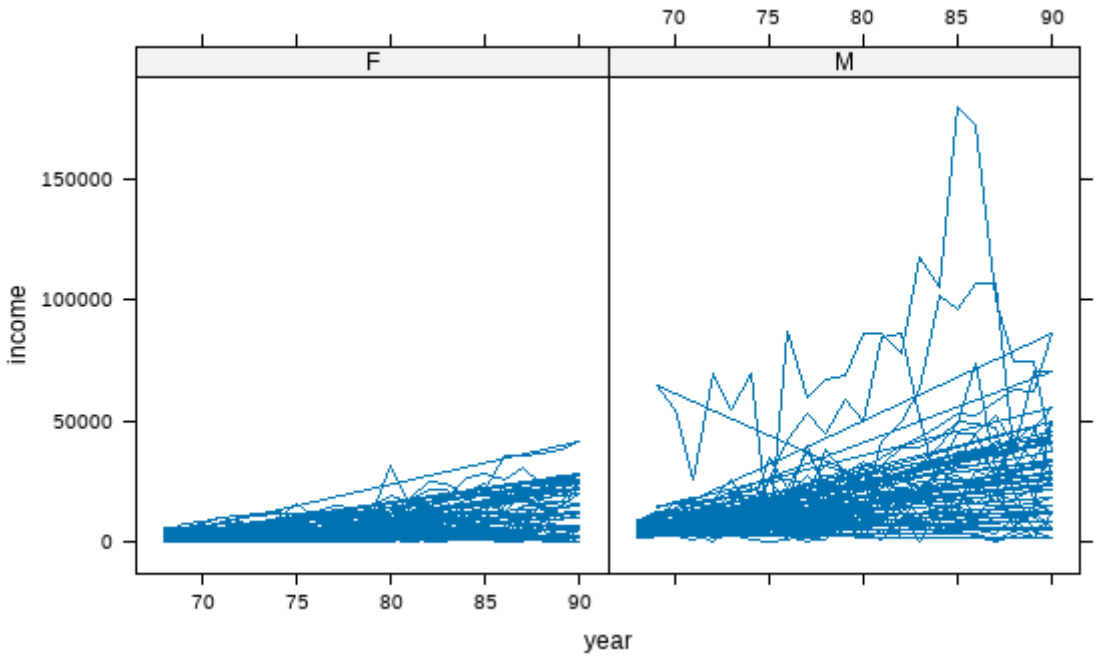
```
library(faraway);library(lattice)
data(psid)
mypsid<-subset (psid, (subset=(person <4)))
mypsid1<-subset(mypsid, (subset=(year < 75)))
xyplot(income ~ year , psid, type="l", subset=( person < 20),strip=TRUE)
```



- In the PSID dataset, some individuals have a slowly increasing income, while others have more erratic incomes.
- There is small variation at the beginning compared to the end.
- Income may possibly vary by sex, so we may need profile plots by sex.

PSID: Profile of Income by Sex

```
xyplot(income ~ year | sex, psid, type="l", subset=( person < 500),strip=TRUE)
```



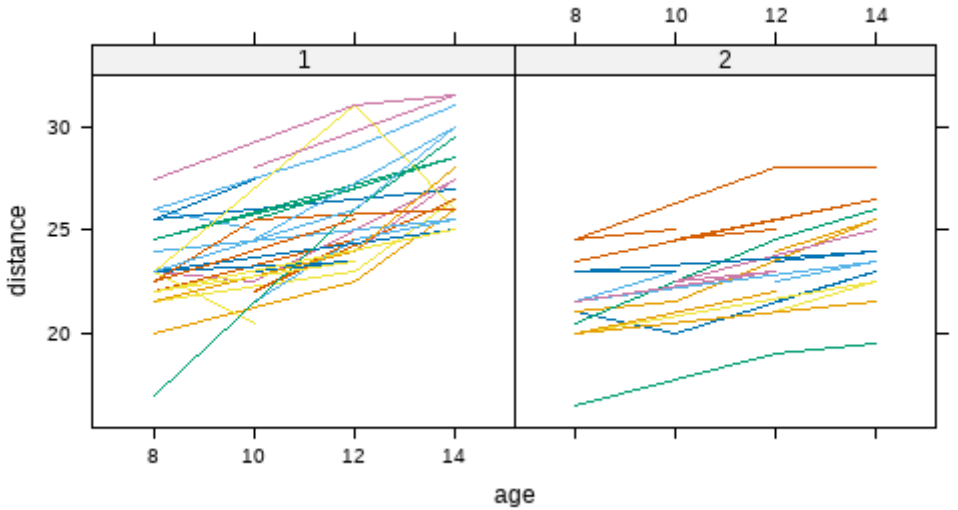
- The variation for males is higher than that of females.
- Income data for males is more erratic.
- Variation at the beginning is smaller than the variation at the end for both groups.

Orthodontic Growth Data

- Taken from Potthoff and Roy, Biometrika (1964).
- The distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14, for 11 girls and 16 boys.
- Data were collected by orthodontists from x-rays of the children's skulls.
- 108 total records were grouped into 27 groups by Subject.
- This is an example of balanced repeated measures data, with a single level of grouping (Subject).
- Research question: Is dental growth related to gender?

The Orthodontic Growth Data: Individual Profiles by Sex

```
library(readr); library(dplyr)
growth <- read_csv("Data/growth2.csv")
  growth$age <- as.factor(growth$age); growth$Sex <- as.factor(growth$sex)
 growth %>%  xyplot(measure ~ age|Sex, data = ., groups = ind, type = "l",
         xlab ="age",ylab= "distance")
```



- Much variability between children.
- Considerable variability within children.
- Fixed number of measurements per subject.
- Measurements taken at fixed time points.

The Orthodontic Growth Data: Mean Distance Profile by Sex

```
mean1<-tapply(growth$measure, growth$age, mean)
age1<-as.numeric(unique(growth$age)) %>% sort()
plot(age1, mean1, type= "l",ylim=c(20,30), xlab="age",
     ylab=" The mean distance", lwd=3,
     main=" The mean profile of the growth data set")
```

- The relationship between age and distance appears to be linear.
- It appears that there is a linear growth pattern.
- The mean profile for males is higher than that of females.

Cross-sectional versus Longitudinal Data

Cross-sectional data and longitudinal data are two primary types of data used in statistical analysis:

- Cross-sectional data: Collected at a specific point in time and involves observations of different individuals at that particular time.
- Longitudinal data: Collected over an extended period, involving repeated observations of the same individuals over time.

t-test Example

Diastolic Blood Pressures from the Captopril Data

- Consider the diastolic blood pressures.
- It includes 15 patients with hypertension.
- The response of interest was the supine blood pressure before and after treatment with CAPTOPRIL.
- Research question: How does treatment affect blood pressure?

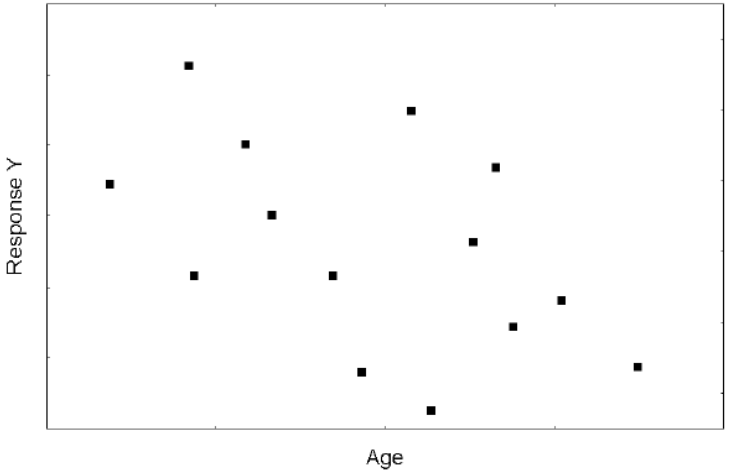
```r
library(tidyverse)
before <- c(130, 122, 124, 104, 112, 101, 121, 124, 115, 102, 98, 119, 106, 107, 100)
after <- c(125, 121, 121, 106, 101, 85, 98, 105, 103, 98, 90, 98, 110, 103, 82)
data <- data.frame(
  Group = rep(c("Before", "After"), each = 15),
  Diastolic_BP = c(before, after))
paired <- data %>% group_by(Group) %>%
  summarize(Mean = mean(Diastolic_BP),  N = n(),Std_Deviation = sd(Diastolic_BP),
            Std_Error_Mean = sd(Diastolic_BP)/sqrt(n()))
print(paired)
```

```
## # A tibble: 2 × 5
##   Group   Mean     N Std_Deviation Std_Error_Mean
##   <chr>  <dbl> <int>         <dbl>          <dbl>
## 1 After   103.    15          12.6           3.24
## 2 Before  112.    15          10.5           2.70
```

- Paired data analysis: Examines related variables within the same subjects.
- Diastolic blood pressure: Analyzed before and after treatment.
- Average decrease: More than 9 mmHg after treatment.
- Classical Analysis: Compares measurements within each subject.
- Focus: Changes from before to after treatment.
- Testing for Treatment Effect: Assesses if the average difference equals zero.

Cross-sectional vs Longitudinal Data

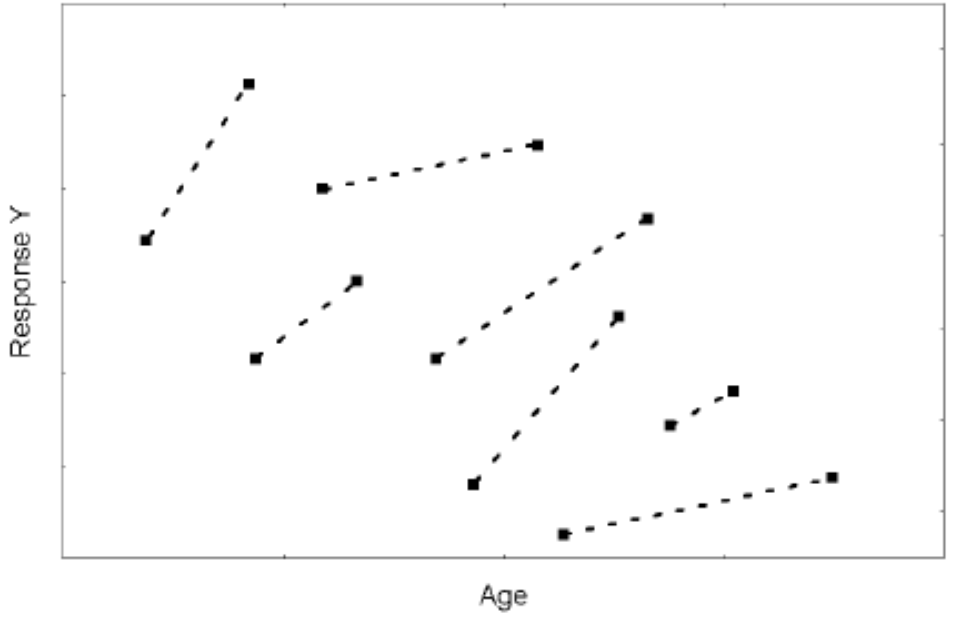
- Cross-sectional data refers to the data collected at a specific point in time.
- Observations from cross-sectional data are uncorrelated.
- Longitudinal data refers to measurements made repeatedly over time to study how the subjects evolve over time.
- That means the concern of longitudinal data analysis is change over time.
- In longitudinal study, the measurements made for subjects over a period of time are correlated.
- Suppose it is of interest to study the relation between some response Y and age.
- A cross-sectional study yields the following data:



(Graph suggesting a negative relation between Y and age)

Cross-sectional vs Longitudinal Data

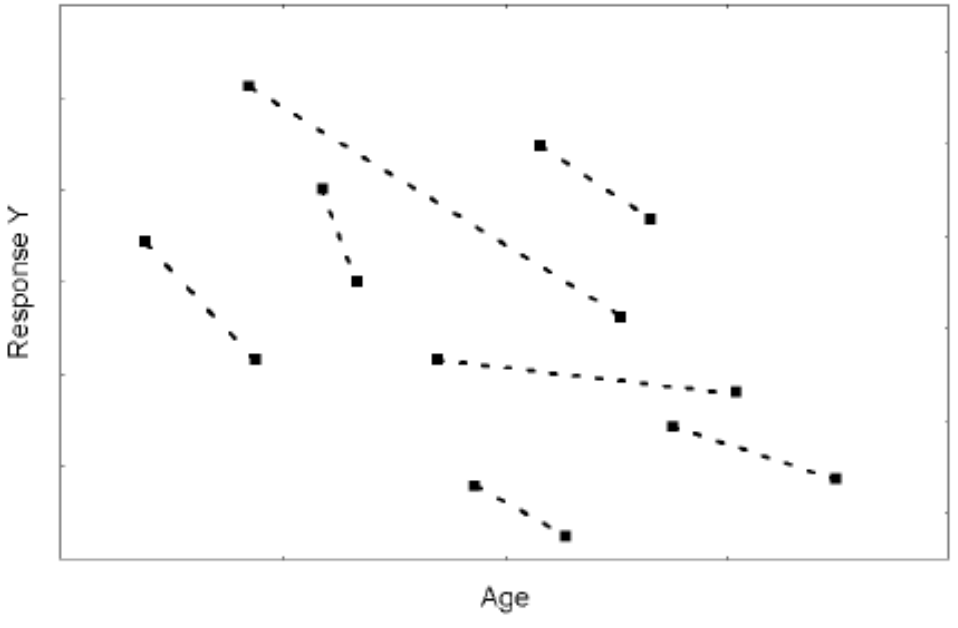
- Exactly the same observations could also have been obtained in a longitudinal study, with 2 measurements per subject.
- First case:



(Graph suggesting a negative cross-sectional relation but a positive longitudinal trend)

Cross-sectional vs Longitudinal Data

- Second case:



(Graph suggesting the cross-sectional as well as longitudinal trend to be negative)

- **Conclusion:** Longitudinal data allow distinguishing differences between subjects from changes within subjects.

Longitudinal data: broad form

- Wide format of data

Subject	Time 1 (y)	Time 2 (y)	Time 3 (y)	Time 1 (x)	Time 2 (x)	Time 3 (x)
1	10	6	6	4	4	4
2	7	5	3	2	2	2
3	12	9	8	6	6	6
4	11	14	16	8	8	8

- ```
  # Using pivot_wider
  wide_data <- long_data %>%
  pivot_wider(names_from = Time,
              values_from = c(y, x))
  ```

- Long format of data

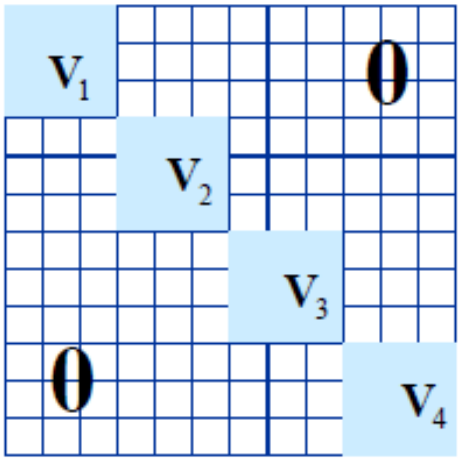
subject	time	y	x
1	1	10	4
1	2	6	4
1	3	6	4
2	1	7	2
2	2	5	2
2	3	3	2
3	1	12	6
3	2	9	6
3	3	8	6
4	1	11	8
4	2	14	8
4	3	16	8
  ```

Longitudinal data

- With LD: multiple measurements taken on each subject.
- You not only can examine the differences between subjects, but you can also examine the change within subjects across time.
- The number of observations is not the number of subjects but rather the number of measurements taken on all the subjects.
- There are three repeated measurements on each subject, you now have 12 observations.
- How does this change your variance-covariance matrix?

Variance-Covariance Matrix for Longitudinal Data

Subject	Time	X	Y
1	1	4	10
1	2	4	6
1	3	4	6
2	1	2	7
2	2	2	5
2	3	2	3
3	1	6	12
3	2	6	9
3	3	6	8
4	1	8	11
4	2	8	14
4	3	8	16



- 3 repeated measurements on each subject:
  - We now have 12 observations and a 12x12 variance-covariance matrix.
- For a simple longitudinal model, the matrix is a block-diagonal matrix.
- The matrix is a block-diagonal matrix:
  - Observations within each block are assumed to be correlated.
  - Observations outside of the blocks are assumed to be independent (subjects are still assumed to be independent of each other).

Introduction to Longitudinal Analysis

Longitudinal data: Data in the form of repeated measurements over time or other factors on each individual or unit in a sample from a population of interest.

Examples:

- Weekly measurements of growth on experimental plots with different fertilizers.
- Monthly measurements of viral load on HIV-infected patients with different treatment regimens.

Defining Characteristic

The same response or outcome is measured repeatedly on each unit.

Scientific Questions

- How mean response differs across treatments or other factors.
- How the change in mean response over time differs.
- Other features of the relationship between response/outcome and time.

Required Statistical Model

A statistical model that acknowledges this data structure in which the questions can be formalized and associated specialized methods of analysis based on the model.

- Longitudinal data studies have become increasingly common and widespread across various scientific disciplines.
- We will study both classical and more modern approaches to representing and interpreting these data.

Terminology

- Longitudinal data refers to data in the form of repeated measurements that might be over time but could also be over some other set of conditions.
- Time is most often the condition of measurement.
- "Response" and "outcome" are used interchangeably to denote the repeated measurement or outcome of interest.
- "Unit," "individual," and "subject" are used interchangeably to refer to the entity being measured.

Applications

Next, we consider several applications that exemplify longitudinal data situations and the range of ways data are collected and types of responses and questions of interest.

Day 2: Simple Methods

- Introduction: longitudinal studies
- Choosing outcome measures
- Simple methods
    - Overview of frequently used methods
- Summary statistics
- Practical using R

Simple Methods

- The reason why classical statistical techniques fail in the context of longitudinal data is that observations within participants 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 participant specific differences $\Delta_i = Y_{i1} - Y_{i2}$
    - This reduces the number of measurements to just one per participant, which implies that classical techniques can be applied again.

Overview of frequently used methods

- In the case of more than 2 measurements per participant, similar simple techniques are often applied to reduce the number of measurements for the i^{th} participant, from n_i to 1.

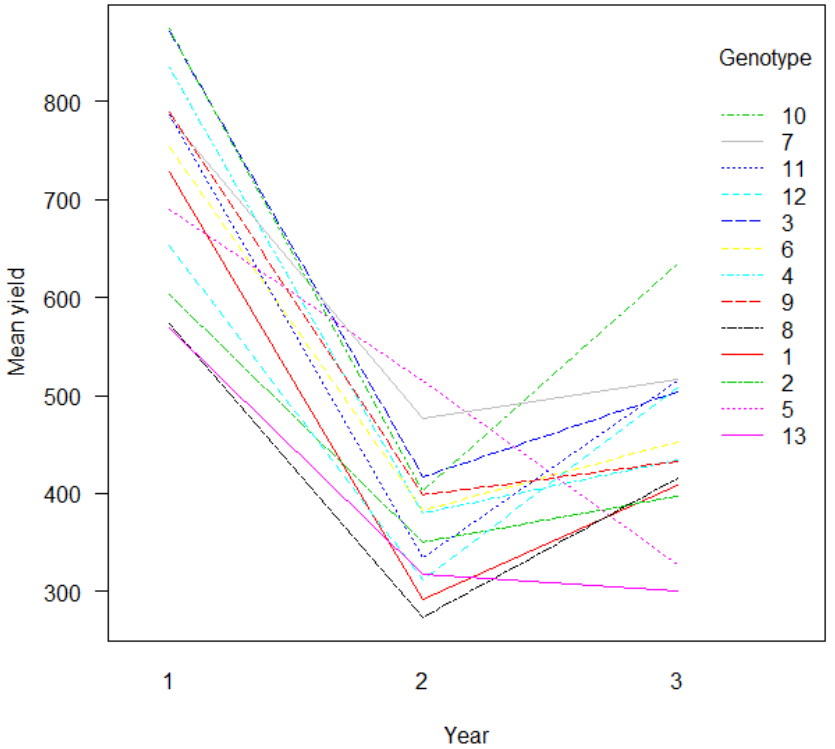
Some Examples:

- Analysis at each time point separately
- Analysis of Area Under the Curve (AUC)
- Analysis of endpoints
- Analysis of increments
- Analysis of covariance

The Sesame Data

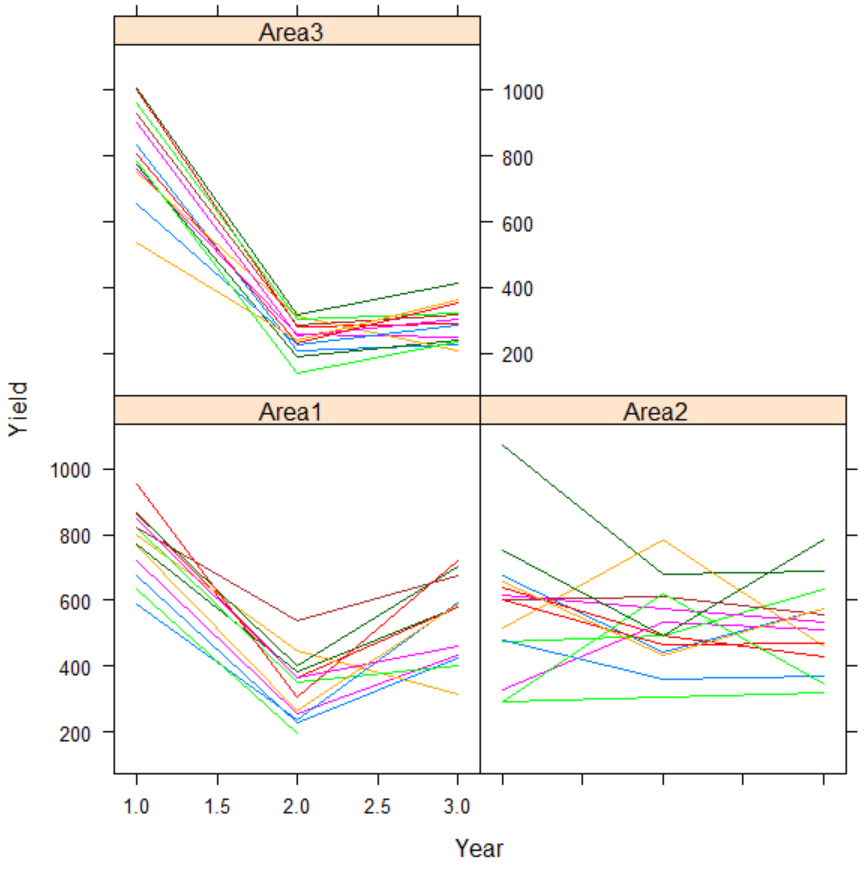
- Cross-year by locations trial of Sesame Genotypes
- Sesame is a short day plant & sensitive to photo-period, temperature, and moisture stress.
- The yield is reported to vary across years and locations.
- Variation in rainfall could lead to the change of yield across locations.
- The study area was characterized by uni-modal rainfall pattern
    - low in amount, short in duration, and poor in distribution within a short distance.
- To examine the effect of Genotype X Environment Interaction, 13 genotypes were tested across three locations over three years.
- Aim: to see the change in yield of different genotypes over years.

The Sesame Data



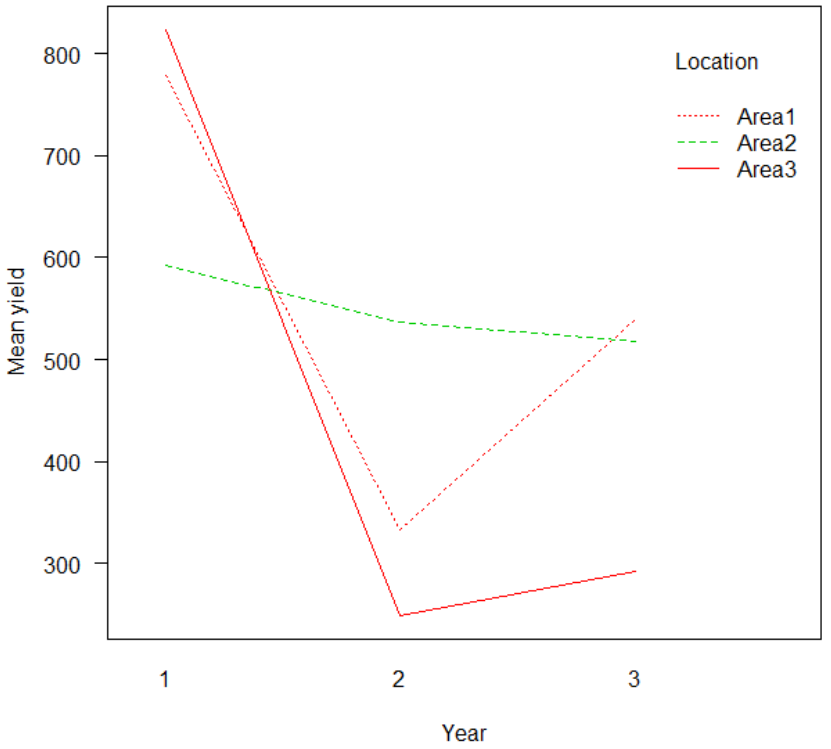
- Three years average yield over the three locations by thirteen genotypes
- Sesame Data: Profile plot by Year and Location

Sesame Data: Profile plot by Year and Location



- Fixed # of measurements per genotypes
- It seems that the yield on the first year is >> yield on the 2nd year for the majority of the observations.
- Some lines on the plot show an increasing trend
- Some genotypes have a larger yield than others
- Variability between genotypes
- Variability within genotypes
- May be it is good to see the mean profile plot by location.

Mean profile of Sesame yield by Location



- Area2: the mean yield is almost constant over time
- Area1 & Area3: the mean yield decrease from Yr 1 to Yr 2 and then increase to Yr 3.

Overview of Frequently Used Methods

Analysis at Each Time Point Separately

- The data are analyzed at each occasion separately.
- Example: Use the Sesame Data set to analyze the number of days to maturity using one-sample t-test for each location at each time point.

Summary of Days to Maturity

- A simple summary of days to maturity for each location at a given time point.
- Comparison of different locations for year 1.
- Performed two-sample t-test, despite having three locations.

Problem of Multiple Testing

- Multiple testing: Conducting multiple t-tests leads to inflation of Type I error.
- Experiment-wise Type I error rate: Probability of falsely rejecting at least one null hypothesis among multiple tests.

# of Comparisons (K)	Experiment-wise Type I Error
1	0.05
2	0.0975
3	0.1426
5	0.2262

- With 5 tests at 5% significance level, there's a 22.6% chance of observing at least one significant result even if all tests are not significant.

Experiment-wise Type I Error

- Increasing the number of comparisons also increases the overall significance.

Advantages of Analysis at Each Time Point

- Simple to interpret.
- Uses all available data.

Disadvantages of Analysis at Each Time Point

- Does not consider overall differences.
- Does not allow studying evolutionary differences.
- Problem of multiple testing.
- Possible issues with missing data.

Analysis of Area Under the Curve (AUC)

- For each participant, the area under its curve is calculated:

$$AUC_i = (t_{i2} - t_{i1}) * (y_{i2} - y_{i1})/2 + (t_{i3} - t_{i2}) * (y_{i3} - y_{i2})/2 + \dots$$

- Afterwards, these AUC_i are analyzed
- Ex: we use the days to CFU data to calculate the area under the curve.

Advantages

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

Disadvantage

- uses only partial information: AUC_i
- participants 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
- In randomised studies, there are no systematic differences at baseline.
- Hence, ‘treatment’ effects can be assessed by only comparing the measurements at the last occasion

Advantages

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

Disadvantages

- uses only partial information
- only valid for large data sets
- the last time point must be the same for all participants
- does not consider ‘overall’ differences
- possible problems with missing data

Analysis of increments

- A simple method to compare evolutions between participants, correcting for differences at baseline, is to analyze the participant-specific changes: $y_{in_i} - y_{i1}$

Advantages

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

Disadvantage

- uses only partial information
- the last time point must be the same for all participants
- possible problems with missing data

Analysis of covariance

- Another way to analyse endpoints, correcting for differences at baseline, is to use analysis of covariance techniques, where the first measurement is included as covariate in the model.

Advantages:

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

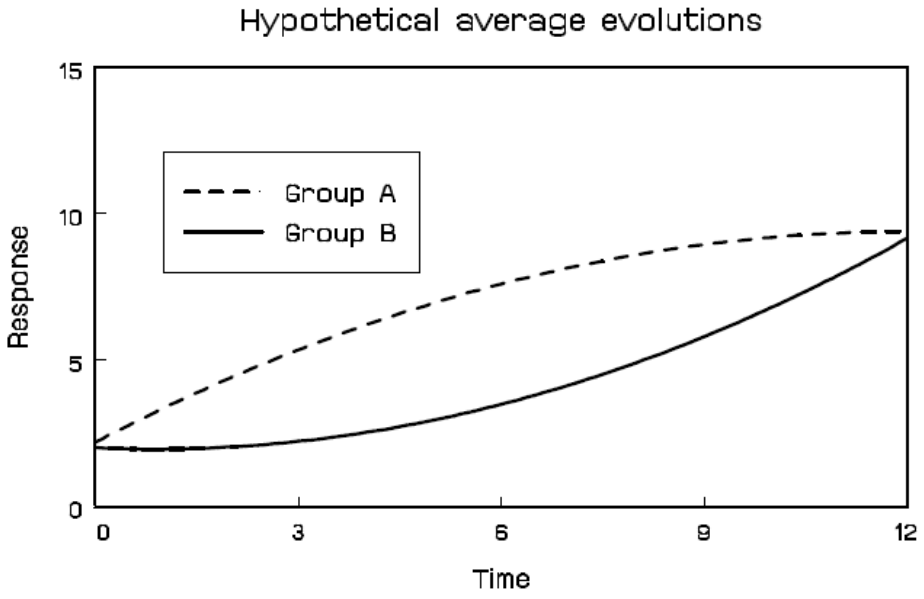
Disadvantages:

- uses only partial information: y_{i1} and y_{in_i}
- does not take into account the variability of y_{i1}

Summary Statistics

- The AUC, endpoints and increments are examples of summary statistics.
    - Such summary statistics summarise the vector of repeated measurements for each participant separately.
- This leads to the following general procedure:
    - Step 1: Summarize data of each participant into one statistic, a summary 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.
- These techniques are based on extensions of simple regression models for univariate data

- However, all these methods have the disadvantage that (lots of) information is lost
- Further, they often do not allow to draw conclusions about the way the endpoint has been reached



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

Exploratory Data Analysis

Introduction

Exploratory analysis comprises techniques to visualize patterns in the data.

Data analysis must begin by making displays that expose patterns relevant to the scientific question.

A linear mixed model makes assumptions about:

- mean structure: (non-)linear, covariates, ...
- variance function: constant, quadratic, ...
- correlation structure: constant, serial, ...
- subject-specific profiles: linear, quadratic, ...

In practice, linear mixed models are often obtained from a two-stage model formulation.

However, this may or may not imply a valid marginal model.

Exploratory Data Analysis

Longitudinal data analysis, like other statistical methods, has two components which operate side by side:

- exploratory and
- confirmatory analysis.

Exploratory analysis comprises techniques to visualize patterns in the data. Confirmatory analysis is judicial work, weighing evidence in data for, or against hypotheses.

Data analysis must begin by making displays that expose patterns relevant to the scientific question.

The best methods are capable of uncovering patterns which are unexpected.

In this regard graphical displays are so important. At this stage, the following guidelines are very useful.

Jimma Infant Data

Follow-up study of newborn infants in Southwest Ethiopia.

Wide ranges of data were collected on the following characteristics:

- basic demographic information,
- feeding practice,
- anthropometric measurements, ...

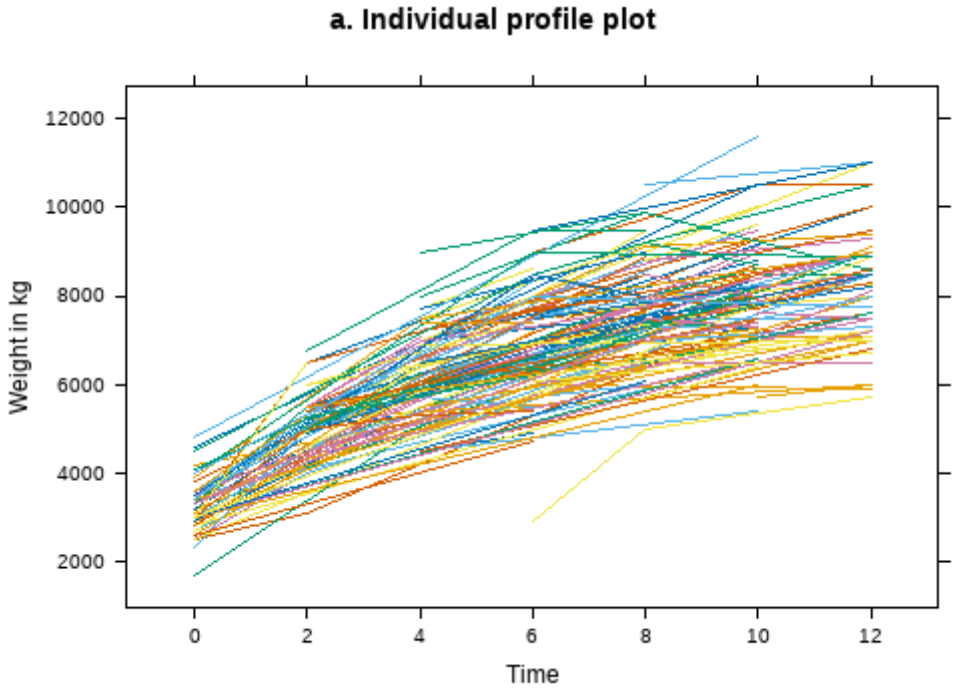
Infants were followed during 12 months. Measurements were taken at seven time points every two months from each child. Weight was one of the variables recorded at each visit.

Research question: How does weight change over time?

Jimma Infant Data

The individual profiles support a random-intercepts model

```
attach(Infant); library(lattice)
mydata1<-as.data.frame(Infant)
xyplot(weight ~ factor(age), group = ind, lty=1 ,
       data =mydata1[sample(nrow(mydata1), 800),], main="a. Individual profile plot",
       xlab = "Time", ylab = "Weight in kg", type = "a", lines = TRUE)
```



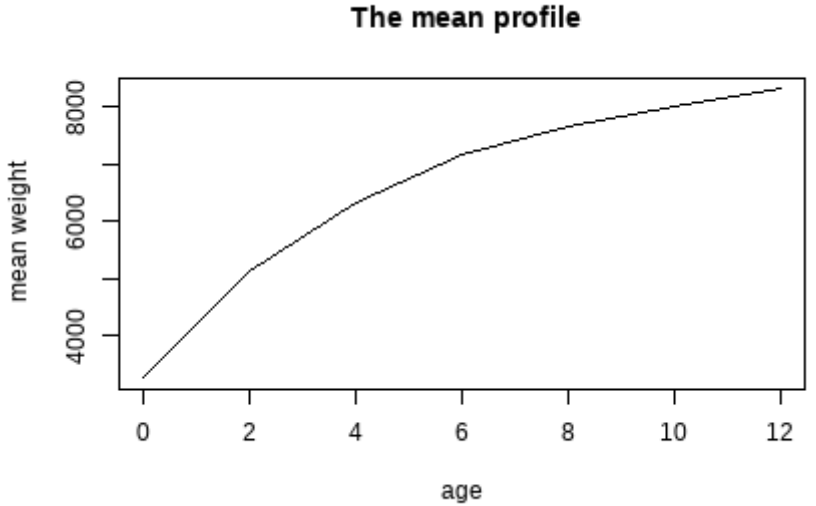
Conclusions From the profile

- Much variability between children
- Considerable variability within subjects
- Fixed number of measurements per subject
- Measurements taken at fixed time points

Mean Profile

The mean profile can be plotted using the following R code:

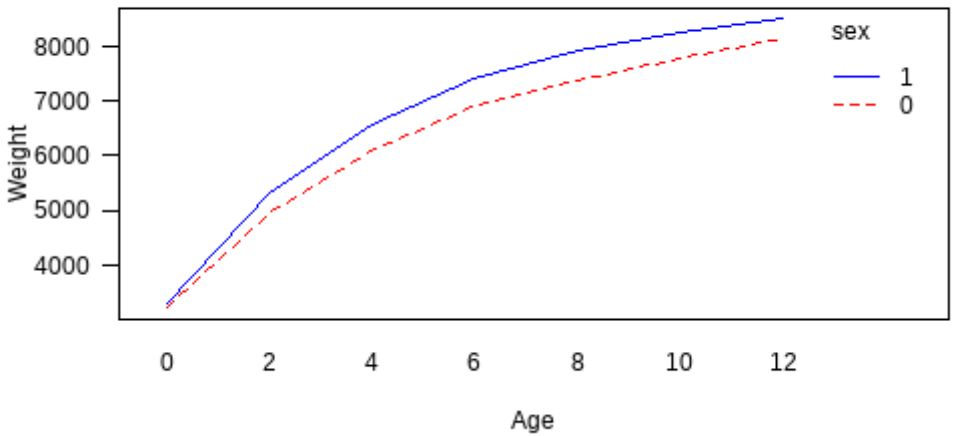
```
mean1<-tapply(Infant$weight, Infant$age, mean, na.rm=T)
age1<-as.numeric(unique(Infant$age))
plot(age1, mean1, type= "l", xlab="age",
     ylab=" mean weight", lwd=1, main=" The mean profile")
```



Mean Profile by Sex

- The mean profiles by sex:

```
interaction.plot(Infant$age, sex, Infant$weight, fun = mean,
                 col = c("red", "blue"), xlab = "Age", ylab = "Weight", las = 1)
```



Exploring the Random Effects

- The mean structure for linear mixed effect model can be determined based on the random effects.
- Choosing which parameters in the model should have a random-effect component included to account for between-group variation.
- The lmList function and the methods associated with it are useful for this.
- Continuing with the analysis of the Jimma infants data, we see from the individual profiles of these data that a simple linear regression model of weight as a function of age may be suitable.

Jimma Infant Survival

-The data was fitted this for each subject as follows:

```
library(nlme)
fit <- lmList(weight ~ age | ind, Infant)
fit
```

```
## Call:
##   Model: weight ~ age | ind
##    Data: Infant
##
## Coefficients:
##      (Intercept)           age
## 1       4200.000  2.714286e+02
## 3       5435.714  1.821429e+02
## 4       4435.714  2.392857e+02
## 5       4139.286  3.196429e+02
## 6       4485.714  4.571429e+02
## 7       4400.000  3.428571e+02
## 8       4550.000  3.250000e+02
## 9       3792.857  3.250000e+02
## 10      4635.714  3.821429e+02
## 11      3417.143  5.592857e+02
## 12      3860.714  5.089286e+02
## 14      3407.143  2.607143e+02
```

Exploring the Random Effects

- The main purpose of this preliminary analysis is to give an indication of what random effects structure to use in the model.
- We must decide which random effects to include in a model for the data, and what covariance structure these random effects should have.
- Objects returned by lmList are of class lmList, for which several display and plot methods are available.
- The pairs method provides one view of the random effects covariance structure.
- To identify outliers-points outside the estimated probability contour at level 1-id/2 will be marked in the plot, we use the R function.
- We see that subject 29 has high slope.

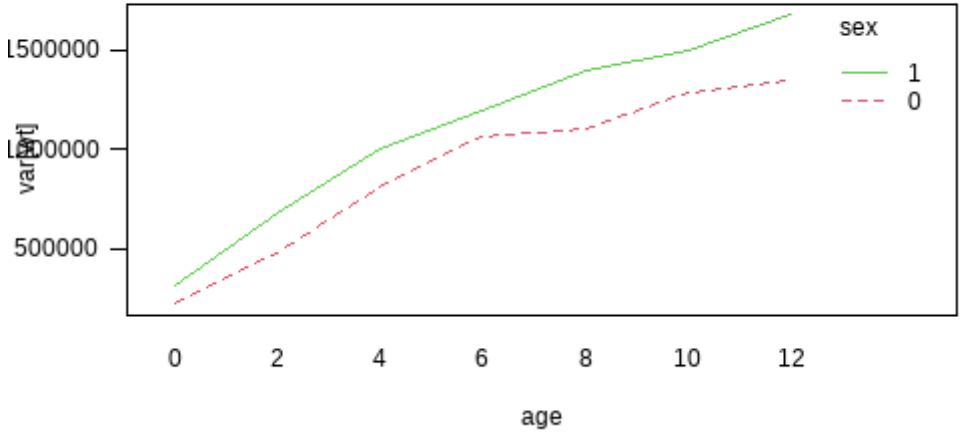
Exploring the Correlation Structure

- In longitudinal data analysis we model two key components of the data:
  - Mean structure
  - Correlation structure (after removing the mean structure)
- Modelling the correlation is important to be able to obtain correct inferences on regression coefficients.
- Correlation can be formulated in terms of:
  - **Random effects**
  - **Autocorrelation** or serial dependence
  - **Noise, measurement error**
- After we explore the mean function in the regression, we need to explore the correlation structure for the residuals, taking away the mean trend effect.

Observed Variance for Jimma dataset

- Having an appropriate model for studying the evolution of the variance is a very important step in the modeling approach.
- The observed variance shows an increase in variability over time.
- Hence, a heterogeneous variance structure may be a good starting point.
- Moreover, the variability for males and females seems to be more or less the same.
- Hence, the same variance structure may be assumed for both groups.

```
interaction.plot(Infant$age, sex, Infant$weight, fun=var, col=2:3, xlab="age",
                 ylab="var[wt]", las=1)
```



Growth Data

- The distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14, for 11 girls and 16 boys.

Research Question: Is dental growth related to **gender?**

- The individual profiles support a random-intercepts model.

```
xyplot(measure ~ factor(age) | sex, group = ind, data = growth,
       main="Individual profile plot by sex",xlab = "Age", ylab = "measure",
       type = "a", lines = TRUE)
```

From the exploratory analysis:

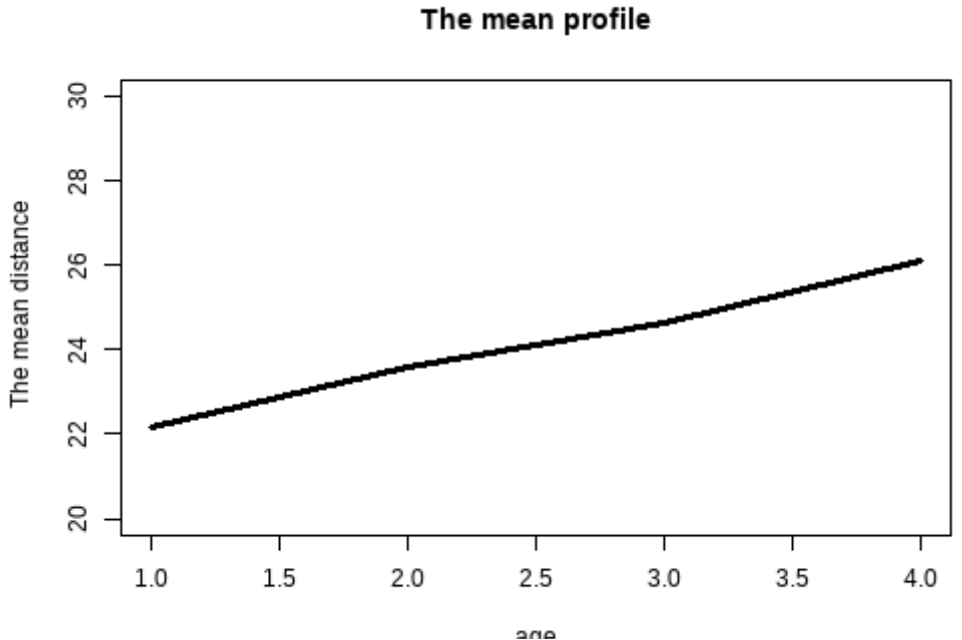
- Mean structure seems linear over time.
- Variability between subjects at baseline.
- Variability between subjects in the way they evolve.

Hence, a linear mean with random intercept and slope is a good idea...

Exploring the Mean Structure of Growth data

For balanced data, averages can be calculated for each occasion separately, and standard errors for the means can be added.

```
attach(growth)
mean1<-tapply(measure, age, mean)
age1<-sort(as.numeric(unique(age)))
plot(age1, mean1, type= "l",ylim=c(20,30), xlab="age", ylab=
" The mean distance", lwd=3, main="The mean profile")
```

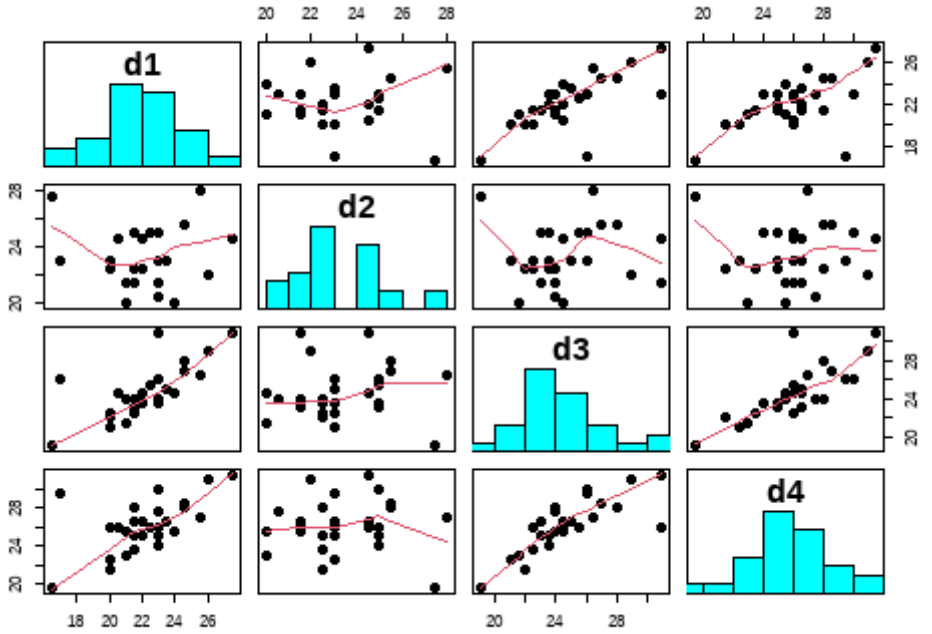


```r
##### R-code for Correlation matrix ####
d1<-measure[age==8]
d2<-measure[age==10]
d3<-measure[age==12]
d4<-measure[age==14]
response1<-cbind(d1, d2, d3, d4)
cor_matrix <- cor(response1)
cor_matrix
```

```
##             d1         d2         d3         d4
## d1 1.00000000 0.04600947 0.71080794 0.59983380
## d2 0.04600947 1.00000000 0.06913238 0.01175495
## d3 0.71080794 0.06913238 1.00000000 0.79499798
## d4 0.59983380 0.01175495 0.79499798 1.00000000
```

Scatter plot matrix for growth data

```r
# Scatter plot matrix
panel.hist <- function(x, ...)
{usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) );  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)}
pairs(response1, panel=panel.smooth, cex = 1.5, pch = 16,  bg="light green",
      diag.panel=panel.hist, cex.labels = 2, font.labels=2)
```



Exploring the Variability of the Observed Data

- The individual profile plots of the growth data set exhibit substantial variability within and between subjects.
- This intricate variability can be further elucidated by considering the variance-covariance matrix of the observed data, as indicated below:
- By examining the variance-covariance matrix, we gain deeper insights into the extensive variability present within and between subjects in the growth data set.

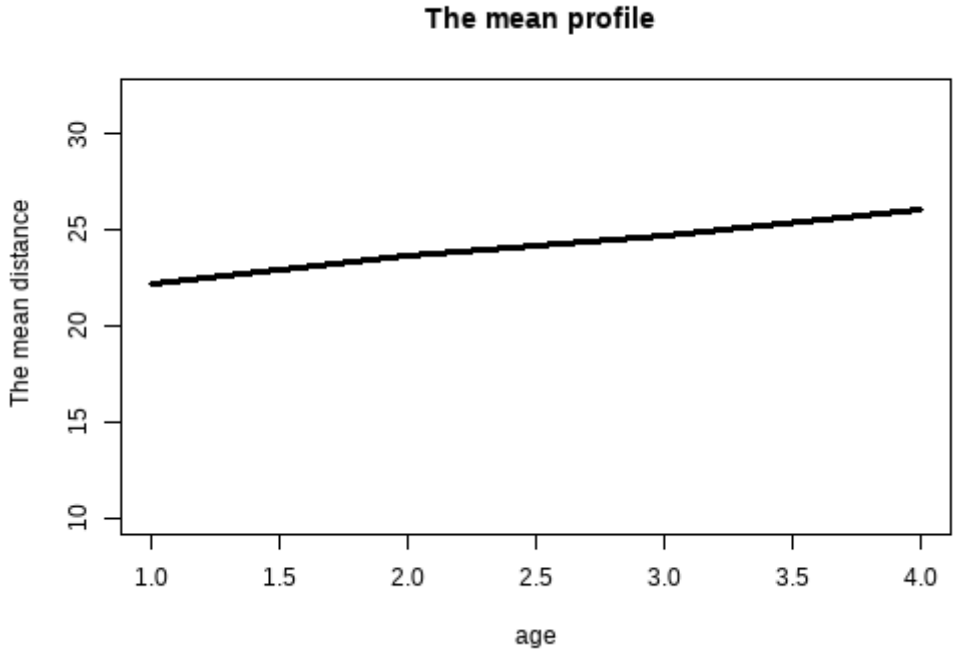
Covariance Matrix for Growth Data:

```
cov(response1)
```

```
##           d1         d2        d3         d4
## d1 5.9259259 0.22934473 4.8753561 4.03988604
## d2 0.2293447 4.19301994 0.3988604 0.06659544
## d3 4.8753561 0.39886040 7.9387464 6.19729345
## d4 4.0398860 0.06659544 6.1972934 7.65455840
```

Exploring Overall Variability

```
## Mean evolution profile ##
mean1<-tapply(measure, age, mean)
age1<-sort(as.numeric(unique(age)))
plot(sort(age1), mean1, type= "l",ylim=c(10,32), xlab="age",
     ylab=" The mean distance", lwd=3, main=" The mean profile")
```



Variability by Group

```
interaction.plot(age, sex, measure, lty=c(1, 2), fun=var,
                 ylab="Distance from Pituitary to Pterygomaxillary Fissure (mm)",
                 xlab="Age", trace.label="Group")
title(main="The Variance of the Growth Data Set by Sex")
```

Day 3

A Model for Longitudinal Data

- Linear Mixed Models (LMM)
- Hierarchical versus Marginal Model

Marginal Model: Estimation and Inference

Inference for the Random Effects

Linear Mixed Models

- **Linear mixed models (LMM)** are models that handle data where observations are not independent.
- That is, LMM correctly models correlated errors, whereas procedures in the general linear model family (GLM) usually do not.
    - (GLM includes: t-tests, ANOVA, correlation, regression, and factor analysis, to name a few.)
- LMM can be considered as a further generalization of GLM to better support the analysis of a continuous response.
- Mixed models contain both fixed and random effects.
- These models are useful in a wide variety of disciplines in the physical, biological, and economic sciences.
- They are particularly useful in settings where repeated measurements are made on the same statistical units or where measurements are made on clusters of related statistical units.
- Let us see some of the terms associated with mixed models.

Types of Effects in Linear Mixed Models

Fixed Effects

- **Factors for which the only levels under consideration are contained in the coding of those effects.**
  - Example: **Sex** where both male and female genders are included in the factor, is considered a fixed effect.
  - Example: **Agegroup** with levels "Minor" and "Adult" included in the factor is also considered a fixed effect.

Random Effects

- **Factors for which the levels contained in the coding of those factors are a random sample of the total number of levels in the population for that factor.**
  - Example: **Subject** can be considered a random effect if it represents a random sample of the target population.
- Random effects models allow researchers to make inferences over a wider population in Linear Mixed Models (LMM) than would be possible with Generalized Linear Models (GLM).

Hierarchical Effects

- **Hierarchical designs have nested effects.**
  - Nested effects are those with subjects within groups. For instance, in a medical study, "Patients" may be nested within "Doctors," and "Doctors" may, in turn, be nested within "Hospitals."
  - We can have a hierarchical effect when the predictor variables are measured at more than one level (ex., reading achievement scores at the student level and teacher-student ratios at the school level).

- Considering hierarchical effects in Linear Mixed Models allows researchers to account for the nested structure of the data and make more accurate inferences about the relationships between variables at different levels of the hierarchy.

In Practice: Handling Unbalanced Data

- Often, data is unbalanced:
    - Unequal number of measurements per subject
    - Measurements not taken at fixed time points
- As a result, traditional multivariate regression techniques may not be applicable.

Subject-Specific Profiles

- Subject-specific longitudinal profiles can be well approximated by linear regression functions.
- A 2-stage model formulation is common:
    1. **Stage 1**: Linear regression model for each subject separately.
    2. **Stage 2**: Explaining variability in subject-specific regression coefficients using known covariates.

Stage 1 Model

$$Y_i = Z_i \beta_i + \varepsilon_i$$

- $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$
- Z_i is a $n_i \times q$ matrix of known covariates.
- β_i is a q dimensional vector of subject-specific regression coefficients.
- $\varepsilon_i \sim N(0, \Sigma_i)$.
- Often, $\Sigma_i = \sigma^2 I_{N_i}$.
- This model describes the observed variability within subjects.

A 2-stage Model

- Between-subject variability can now be studied by relating the β_i to known covariates.
- **Stage 2 model**:

$$\beta_i = K_i\beta + b_i$$

  - K_i is a $q \times p$ matrix of known covariates.
  - β is a p-dimensional vector of unknown parameters.
  - $b_i \sim N(0, D)$.

The General Linear Mixed-effects Model

- A 2-stage approach can be performed explicitly in the analysis.
- However, this is just another example of the use of summary statistics.
  - Y_i is summarized by $\hat{\beta}_i$.
  - Summary statistics $\hat{\beta}_i$ are analyzed in the second stage.
- The associated drawbacks can be avoided by combining the two stages into one model:

$$\begin{cases} Y_i = Z_i\beta_i + \varepsilon_i \\ \beta_i = K_i\beta + b_i \end{cases}$$

$$\Rightarrow Y_i = Z_iK_i\beta + Z_ib_i + \varepsilon_i$$

$$= X_i\beta + Z_ib_i + \varepsilon_i$$

The model is given by:

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

Where:

$$\{\, b_i \sim N(0, D)\ \varepsilon_i \sim N(0, \Sigma_i)\ b_1, b_2, \ldots, b_N, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N\ are\ ,independent$$

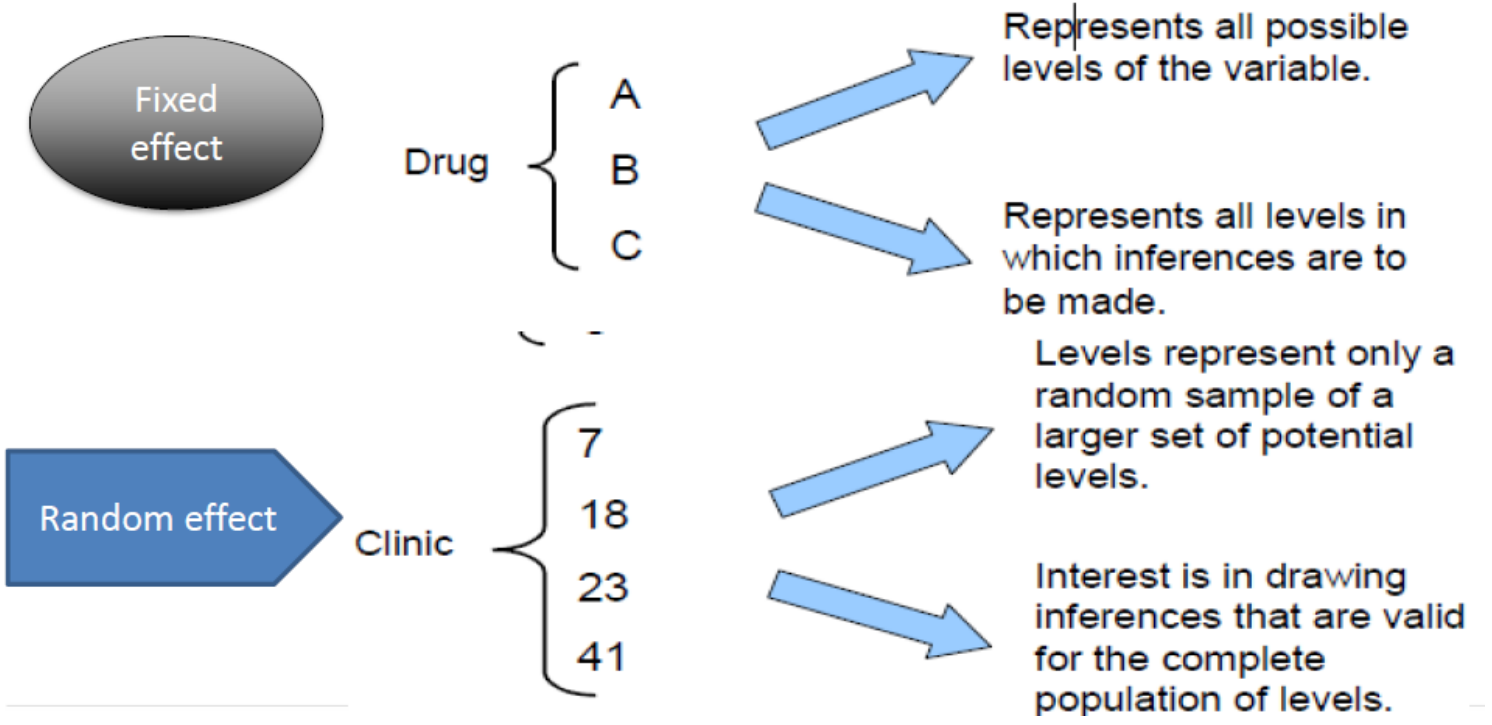
Terminology:

- Fixed effects: β
- Random effects: b_i
- Variance components: D and Σ_i

- For Gaussian data, **GLMM** extends the General Linear Model (GLM) by the addition of random effect parameters and by allowing a more flexible specification of the covariance matrix of the random errors.
- **GLM**: $Y_i = X_i\beta + \epsilon_i$
  - **GLM** includes t-tests, analysis of variance (ANOVA), correlation, regression, and factor analysis, etc.
- **GLMM**: $Y_i = X_i\beta + Z_i b_i + \epsilon_i$
- Difference?
- ϵ_i
  - **GLM**: vector of random errors
  - **GLMM**: is no longer required to be independent and homogenous
- **Mixed Effects Models**
  - Applicable to all types of outcomes (continuous, discrete)
  - Can handle both time-variant and time-invariant covariates
  - Robust to missing data (irregularly spaced observations)

- Contains both fixed and random effects

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$



Summary

- LMM extends the GLM by the addition of random effect parameters and by allowing a more flexible specification of the covariance matrix of the random errors
- LMM can easily fitted to longitudinal data
- Estimation is more difficult in mixed models than GLM
- Longitudinal models have three sources of variation
    - between subject variability/represented by random effect
    - Within subject variability/represented by serial correlation
    - Measurement error

Source of random variation

- $Y_i = X_i\beta + Z_i b_i + \epsilon_i$
- $Y_i = X_i\beta + Z_i b_i + \epsilon_{(1)i} + \epsilon_{(2)i}$

3 stochastic components:

- b_i: between-participant variability
- $\epsilon_{(1)i}$: measurement error
- $\epsilon_{(2)i}$: serial correlation component

Random effects (variation between participants)

- Characteristics of individual participants
- For example, intrinsically high or low responders

Serial correlation (variation over time within participants)

- Measurements taken close together in time are strongly correlated than those taken further apart in time
- On a sufficiently small scale, this kind of structure is almost inevitable

Measurement error

- Extra component of measurement error reflecting variation added due to the measurement process.

Model Families

- Marginal (population average) models
- Subject specific models
- Conditional models

Each can be

- Random intercept model
- Random slope model
- Random higher order model

Model families

Marginal (population average) models

- Responses are marginalized over all other responses
- Parameters characterize the marginal expectation

Subject specific models

- If the aim is to study how subjects change overtime & what characteristics influence such changes
- Subject specific models differ from marginal models by the inclusion of parameters specific to the subject

Conditional models

- Any response within the sequence of repeated measures is modeled conditional upon the other outcomes
- Parameters describe a feature of outcomes, given values for the other outcomes (Cox 1972) i.e. log linear models

1. Random intercept model

$$\begin{cases} Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + \epsilon_{ij} \\ b_{0i} \sim N(0, \sigma_b^2), \varepsilon_i \sim N(0, \sigma_\epsilon^2) \\ b_{0i}, \varepsilon_{ij} \, are \; independent \end{cases}$$

- Each subject has his/her own intercept: $\beta_0 + b_{0i}$
- Interparticipant variability at **baseline**
- Slope remains the same: β_1
- Fixed effects can be added to the model

Conditional distribution:

- Take $E(Y/b)$ and $Var(Y/b)$
  - $E(Y_{ij}/b_{0i}) = \beta_0 + \beta_1 t_{ij} + b_{0i}$
  - $Var(Y_{ij}/b_{0i}) = \sigma_\epsilon^2$
  - $Y_{ij}/b_{0i} \sim N(\beta_0 + \beta_1 t_{ij} + b_{0i}, \sigma_\epsilon^2)$

1. Random intercept model

$$\begin{cases} Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + \epsilon_{ij} \\ b_{0i} \sim N(0, \sigma_b^2), \varepsilon_i \sim N(0, \sigma_\epsilon^2) \\ b_{0i}, \varepsilon_{ij} \, are \; independent \end{cases}$$

Marginal distribution (marginal over the random intercepts):

- Take $E(Y_{ij})$ and $Var(Y_{ij})$
  - $E(Y_{ij}) = \beta_0 + \beta_1 t_{ij}$
  - $Var(Y_{ij}) = \sigma_b^2 + \sigma_\epsilon^2$
  - $Y_{ij} \sim N(\beta_0 + \beta_1 t_{ij}, \sigma_b^2 + \sigma_\epsilon^2)$

The random intercept model: ICC

- Measurements from the same participant share a random effect:
  - Means that marginally, there is a correlation structure
- Let's consider two measurements from the same participant:
  - Y_{ij} and Y_{ik}, $j \neq k$
  - $Cov(Y_{ij}, Y_{ik}) = \sigma_b^2$
- Correlation between two measurements from the same participant:

$$Cov(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}$$

2. Random Slope Model

$$\begin{cases} Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij} \\ b_{0i}, b_{1i} \sim N(0, D), \varepsilon_i \sim N(0, \sigma_\epsilon^2) \\ \epsilon_{ij} \, independent \, of \, b_{0i}, b_{1i} \end{cases} \qquad D = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix}$$

- Each subject has his/her own intercept: $\beta_0 + b_{0i}$
- Each subject has his/her own slope: $\beta_1 + b_{1i}$
- Allows the profiles to cross each other

- Fixed effects can be added to the model.

- This model has two random effects: b_{0i} and b_{1i}

- Their covariance $\sigma_{10} = \sigma_{01}$:
  - If positive: subjects higher at baseline also have a higher evolution
  - If negative: subjects higher at baseline have a slower evolution

2. Random Slope Model

- If the covariance $\sigma_{10} = \sigma_{01}$ is not restricted:
  - It is called unstructured.

Conditional distribution:

- $E(Y_{ij}|b_{0i}, b_{1i}) = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij}$
- $Var(Y_{ij}|b_{0i}, b_{1i}) = \sigma_\epsilon^2$

Marginal distribution:

- $E(Y_{ij}) = \beta_0 + \beta_1 t_{ij}$
- $Var(Y_{ij}) = \sigma_1^2 t_{ij} + 2\sigma_{01} t_{ij} + \sigma_0^2 + \sigma_\epsilon^2$
- Note: marginal variance is a function of time

The Random Slope Model: ICC

- Let's consider two measurements from the same subject:
  - Y_{ij} and Y_{ik}, $j \neq k$
  - $Cov(Y_{ij}, Y_{ik}) = \sigma_1^2 t_{ij} t_{ik} + \sigma_{01}(t_{ij} + t_{ik}) + \sigma_0^2$
- The ICC is now a function of time:

$$Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_1^2 t_{ij} t_{ik} + \sigma_{01}(t_{ij} + t_{ik}) + \sigma_0^2}{\sqrt{\sigma_1^2 t_{ij} + 2\sigma_{01} t_{ij} + \sigma_0^2 + \sigma_\epsilon^2}\sqrt{\sigma_1^2 t_{ik} + 2\sigma_{01} t_{ik} + \sigma_0^2 + \sigma_\epsilon^2}}$$

Marginal models Vs Subject specific Models

1. The General Linear Mixed-effects model:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

$$b_i \sim N(0, D), \epsilon_i \sim N(0, \Sigma_i)$$

; $b_1, b_2, \ldots, b_N, \epsilon_1, \epsilon_2, \ldots, \epsilon_N$ independent

- It can be written as:

$$Y_i/b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i)$$

$$b_i \sim N(0, D)$$

Marginal models Vs Subject specific Models

It is also called a hierarchical model:

- A model for Y_i given b_i
- A model for b_i

Marginally, we have that Y_i distributed as:

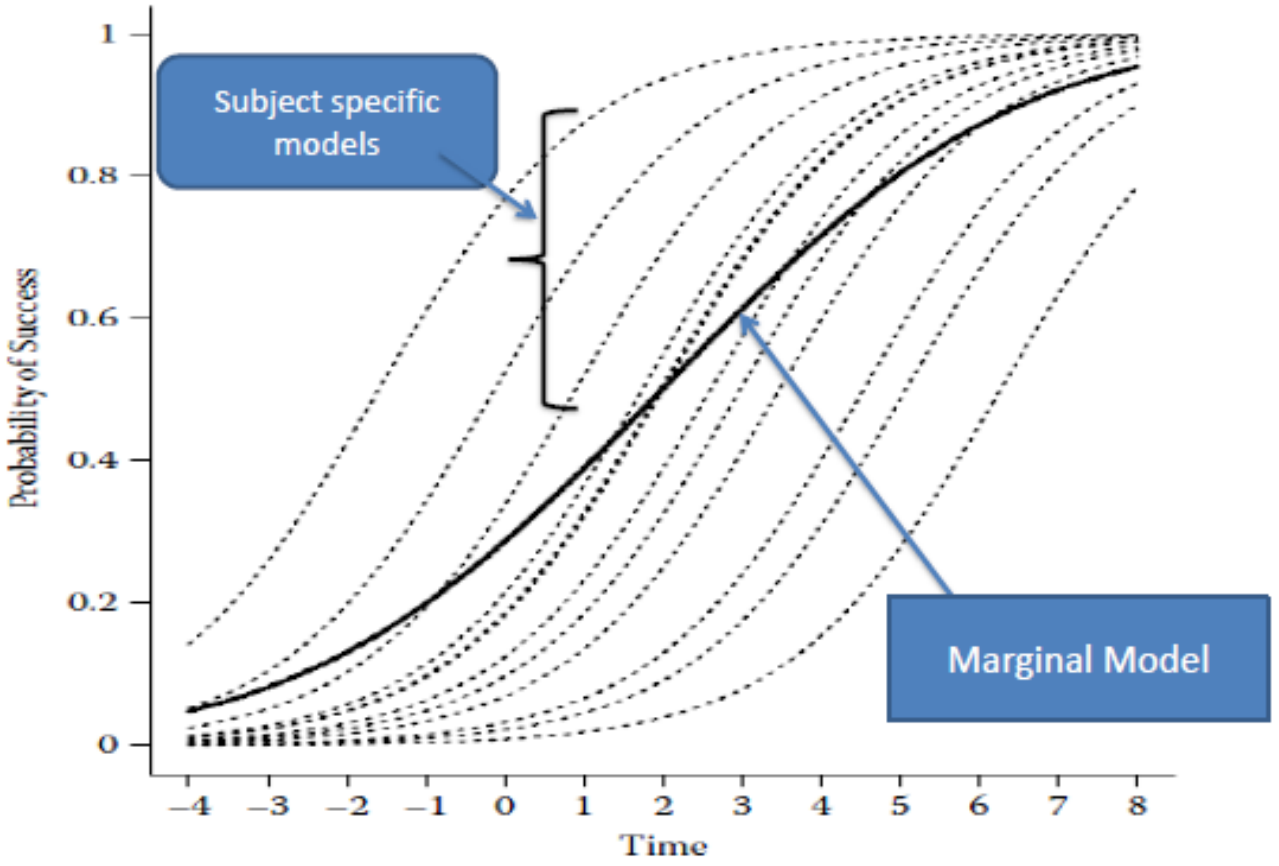
$$Y_i \sim N(X_i\beta, Z_iDZ_i' + \Sigma_i)$$

Hence, very specific assumptions are made about the dependence of mean and covariance on the covariates X_i and Z_i:

- Implied mean: $X_i\beta$
- Implied covariance: $V_i = Z_iDZ_i' + \Sigma_i$

Note that the hierarchical model implies the marginal, NOT vice versa.

Marginal models vs subject specific models … (3)



Marginal model: estimation and inference

- Estimation of the marginal model
  - Introduction
  - Maximum likelihood estimation (MLE)
  - Restricted maximum likelihood estimation (RMLE)
- Practical: Fitting LMM in R

Estimation of the marginal model

- Recall: GLMM

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

where $b_i \sim N(0, D)$, $\epsilon_i \sim N(0, \Sigma_i)$

$b_1, b_2, \ldots, b_N, \epsilon_1, \epsilon_2, \ldots, \epsilon_N$ independent

- The implied marginal model equals: $Y_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i)$
- Let:
  - Residual error covariance matrix: $\Sigma_i = \sigma_i^2 I_{n_i}$
  - The marginal covariance matrix: $V_i = Z_i D Z_i' + \Sigma_i$
- Inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects

Notation:

- β: vector of fixed effects (as before)
- α: Vector of all variance components in D and Σ_i
- $\theta = (\beta', \alpha')$ vector of all parameters in marginal model

- Marginal likelihood function is given by:

$$L_{LM}(\theta) = \prod_{i=1}^{N} \left\{ (2\pi)^{-n_i/2} |V_i(\alpha)|^{-1/2} \exp\left(-\frac{1}{2}(Y_i - X\beta)^T V_i(\alpha)^{-1}(Y_i - X\beta) \right) \right\}$$

If α were known, the maximum likelihood estimate (MLE) of β would be:

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} (X_i^T W_i X_i)^{-1} \sum_{i=1}^{N} X_i^T W_i y_i \right)$$

where $W_i = V^{-1}$.

- In most cases, α were unknown, and needs to be replaced by an estimate $\hat{\alpha}$
- Two frequently used estimation methods for α
    - Maximum likelihood (ML)
    - Restricted maximum likelihood (REML)

Maximum Likelihood Estimation

- **ML estimation of V_i:**
  - Does not take into account that β estimated from data.
  - Does not account for degrees of freedom lost.
  - Generally results in biased estimation of V_i.

Restricted Maximum Likelihood Estimation (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

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

  - 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.

- **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

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

  - 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.

Features of REML estimation:

- It corrects for the downward bias in the ML parameters.
- It handles strong correlations among the responses more effectively.
- Available in all software that fit marginal and mixed effects models.
- The default estimation method in most software (R, SAS).
- It works is by applying a transformation in the longitudinal outcome $\boldsymbol{Y}$ based on the chosen structure of the design matrix $\boldsymbol{X}$(i.e., which predictors you have included in the model).
- Models with different mean structures not comparable.
- Since different observations involved.

Restricted Maximum Likelihood Estimation (REML)

ML estimation of V_i:

- Does not take into account that β estimated from data
- Does not account for degrees of freedom lost
- Generally results in biased estimation of V_i

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 $Var(x) = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x})^2}{n-1}$ to obtain an unbiased estimate we need to divide by $n-1$ because we estimate the mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- 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

Features of REML estimation

- It corrects for the downward bias in the ML parameters
- It handles strong correlations among the responses more effectively
- Available in all software that fit marginal and mixed effects models
- The default estimation method in mos'"t software (R, SAS)
- It works 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)
- Models with different mean structures not comparable.
    - since different observations involved.

ML versus REML

- Both are based on the likelihood principle, which has the properties of consistency, asymptotic normality, and efficiency.
- The differences between ML and REML estimation increase as the number of fixed effects in the model increases.
- Difference between ML and REML is less marked if $n > p$
- We cannot compare the likelihoods of models fitted with REML and have different $X\beta$ part!

Components of the Linear Mixed Effects Model (LMM)

- $Y_i = X_i\beta + Z_i b_i + \epsilon_i$
- The implied marginal model equals: $Y_i \sim N(X_i\beta, Z_i D Z_i' + \Sigma_i)$
- The mean structure: $X_i\beta$
- The covariance structure:
  - When we estimate the covariance matrix without making any particular assumption about the covariance structure, we say that we are using an unrestricted or unstructured covariance matrix (UN).
  - As we shall see later, it is sometimes advantageous to model the covariance structure more parsimoniously.

Fitting Marginal Models in R

- **R>** The following code fits a marginal model for the growth data with a compound symmetry correlation structure:

```
gls.Symm <- gls(distance ~ Sex * I(age - 11), data = growth,
                correlation = corSymm(form = ~1|Subject),
                weights = varIdent(form = ~1|age))
```

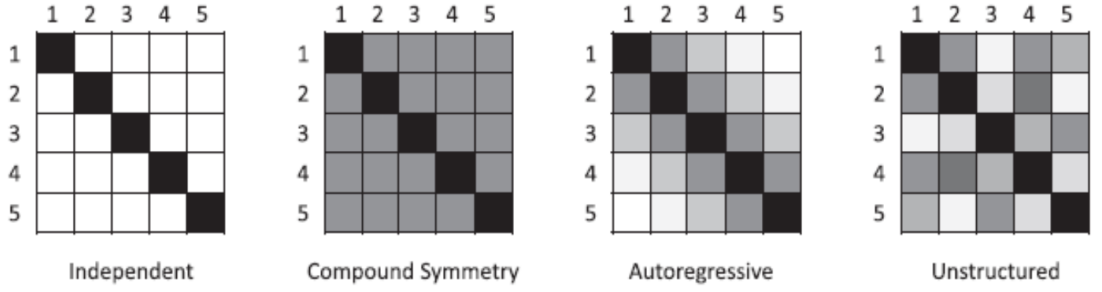
- The Restricted Maximum Likelihood (REML) method is commonly used as the default estimation approach for the Generalized Least Squares (gls).

Covariance Matrix

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)
- We need an appropriate choice for the marginal covariance matrix: $V_i = Z_i D Z_i' + \Sigma_i$ in order to appropriately describe the correlations between the repeated measurements
- Mostly used appropriate for the research question under study– independent, compound symmetry or exchangeable, autoregressive, unstructured.

Covariance Matrix



1. Independent Covariance Structure

- assumes that the measurements are uncorrelated. i.e., no linear relationship between them.

1. Compound Symmetry Covariance Structure

- assumes that all correlations between measurements are equal.

1. Autoregressive Covariance Structure

- assumes thatcorrelations decrease as the time interval between measurements increases.
  - As the time interval increases, the correlation decreases exponentially.

1. Unstructured Covariance Structure

- makes no assumptions about the correlations between measurements.

Example Code

```
library(nlme)
# Independent covariance structure
fit_i <- lme(response ~ time, random = ~ 1 | subject,
             correlation = corSymm(form = ~ 1))

# Compound symmetry covariance structure
fit_c <- lme(response ~ time, random = ~ 1 | subject,
             correlation = corCompSymm(form = ~ 1))

# Autoregressive covariance structure
fit_ar <- lme(response ~ time, random = ~ 1 | subject,
              correlation = corAR1(form = ~ time))

# Unstructured covariance structure
fit_un <- lme(response ~ time, random = ~ 1 | subject,
              correlation = corSymm(form = ~ 1))
```

Model Building

- We have seen that marginal models consist of two parts:
    - Mean part $X\beta$ **:** that describes how covariates we have put in the model explain the average of the repeated measurements.
    - Covariance part V_i **:** assumed covariance structure between the repeated measurements.
- In the majority of the cases, scientific interest focuses on the mean part.
- However, to obtain valid and efficient inferences for the mean part, the covariance part needs to be adequately specified.

General guidelines for model building

- Exploratory data analysis Descriptive statistics, individual group profiles, plots
- Begin with simple models and build towards more complex mean structure
    - 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
- Then select covariance structure covariance matrix V_i that adequately describes the correlations in the repeated measurements
- Finally, reduce the mean structure
    - return to the mean part and exclude non significant covariates
    - start by testing the interaction terms, and then the nonlinear terms
- Model diagnostics

Fitting Linear Mixed Models in R

- There are two packages in R for fitting multilevel models
- The older and more comprehensive package is **`nlme`**, an acronym for nonlinear mixed effects models
- Its limitation is that it only fits normal-based models and was not designed to fit mixed models to non-hierarchical data
- The newer package is **`lme4`**
- It can handle generalized linear mixed effect regression models such as logistic and Poisson regression
- It currently lacks the nonlinear features of **`nlme`**
- Since we are going to focus on examples based on normal theory, our focus will be on the **`nlme`** package

Model: Jimma infant data

$$W_{ij} = \beta_0 + b_{0i} + \beta_1 S_i + (\beta_2 + b_{1i}) A_{ij} + \beta_3 {A_{ij}}^2 + \beta_4 S_i A_{ij} + \beta_5 {A_{ij}}^2 S_i + \varepsilon_{ij}$$

- W_{ij}: weight (Kg) of the i^{th} infant at the j^{th} visit.
- A_{ij}: Age of the i^{th} infant at the j^{th} visit.
- S_i: Sex of the i^{th} infant (Female=0, Male=1)
- b_{0i}: random intercept; b_{1i}: random slope

Basic components from R

- The function `lme` under the library `nlme` in R fits:
  - Linear mixed-effects model
  - Multilevel linear mixed effects model
- It uses maximum likelihood or restricted maximum likelihood
- The command `lme` in R is as follows:

```
lme(fixed, data, random, correlation, weights, subset,
    method, na.action, control, contrasts = NULL, keep.data = TRUE)
```

- `fixed` is an argument to define the fixed effects portion.
- `random` is an argument to define the random effects portion.
- `data` is an optional data frame containing the variables named `correlation` describing the within-group correlation structure.
- `method` is an argument to `lme` that changes the estimation method.

- REML: the model is fit by maximizing the restricted log-likelihood.
- If ML, the log-likelihood is maximized. The Default is REML.

Fixed and Random Parts:

- The fixed part is `fixed = distance ~ Sex + Sex * age`.
- The random part is `random =~ 1|Subject`.

- If the random part is specified as above, it means we will fit a model with a random intercept.

- Here, the response is specified only on the fixed part.

- In the random part, the model statement begins with just a ~.

- If the random formula is omitted, its default value is taken as the right-hand side of the fixed formula.

- The vertical bar separates the model specification from the structural specification.

Model:

$$D_{ij} = \beta_0 + \beta_1 S_i + \beta_2 A_{0ij} + \beta_4 S_i A_{0ij} + b_{0i} + b_{1i} A_{0ij} + \varepsilon_{ij}$$

- D_{ij}: Orthodontic distance of the ith child at the jth visit.
- A_{ij}: Age of the ith child at the jth visit, $A_{0ij} = A_{ij} - 8$
- S_i: Sex of the ith child (boys = 1, girls = 2)
- b_{0i}: Random intercept
- b_{1i}: Random slope

Growth Data

We want to fit a random intercept model on growth data, and the following code can be used:

```
library(nlme)
growth.fit1 <- lme(fixed = measure ~ sex + sex * age,
                   data = growth, random = ~ 1 | ind)
```

Code Explanation:

- Fixed effect: `fixed = measure ~ sex + sex * age`
- Name for the data: `data = mydata22`
- Random intercept: `random = 1 | ind`

For the `growth.fit1` object, `print(growth.fit1)` gives...

```
print(growth.fit1)
```

```
## Linear mixed-effects model fit by REML
##   Data: growth
##   Log-restricted-likelihood: -196.9026
##   Fixed: measure ~ sex + sex * age
## (Intercept)         sex       age10       age12       age14   sex:age10
##  24.5681818  -1.6931818   0.5017488   3.7784091   6.2784091   0.1443845
##   sex:age12   sex:age14
##  -0.9346591  -1.6846591
##
## Random effects:
##  Formula: ~1 | ind
##         (Intercept) Residual
## StdDev:    1.874363 1.442797
##
## Number of Observations: 99
## Number of Groups: 27
```

Main `lme` Methods:

- `ACF`: Empirical autocorrelation function of within-group residuals
- `anova`: Likelihood ratio or conditional tests
- `augPred`: Predictions augmented with observed values
- `coef`: Estimated coefficients for different levels of grouping
- `fitted`: Fitted values for different levels of grouping
- `fixef`: Fixed-effects estimates
- `intervals`: Confidence intervals on model parameters
- `logLik`: Log-likelihood at convergence
- `pairs`: Scatter-plot matrix of coefficients or random effects
- `plot`: Diagnostic Trellis plots
- `predict`: Predictions for different levels of grouping
- `print`: Brief information about the fit
- `qqnorm`: Normal probability plots
- `ranef`: Random-effects estimates
- `resid`: Residuals for different levels of grouping
- `summary`: More detailed information about the fit
- `update`: Update the `lme` fit
- `Variogram`: Semivariogram of within-group residuals

The command `coef(growth.fit1)` in R produced;

```
coef(growth.fit1)
```

```
##    (Intercept)       sex     age10    age12    age14 sex:age10  sex:age12
## 1     23.51517 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 2     24.93052 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 3     25.30203 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 4     26.56361 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 5     24.60390 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 6     23.21436 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 7     24.93052 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 8     25.25713 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 9     23.21436 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 10    20.84833 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 11    27.87008 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 12    27.05404 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 13    23.15900 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 14    24.00561 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 15    26.07419 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 16    22.46311 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 17    25.85644 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 18    23.57012 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 19    23.67899 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 20    24.76772 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 21    28.57826 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
## 22    23.46124 -1.693182 0.5017488 3.778409 6.278409 0.1443845 -0.9346591
```

Command fixef (Ortho.fit1), the following output is produced

```
fixef(growth.fit1)
```

```
## (Intercept)         sex       age10       age12       age14   sex:age10
##  24.5681818  -1.6931818   0.5017488   3.7784091   6.2784091   0.1443845
##   sex:age12   sex:age14
##  -0.9346591  -1.6846591
```

- The parameters are average

- Producing Maximum Likelihood Estimates Using lme
- In all of the above outputs, we produced the Restricted Maximum
- Likelihood Estimates as REML is the default method in the lme.
- The argument method=ML requests that estimates be obtained using full maximum likelihood.

```
growth.fit2 <-lme(fixed = measure ~ sex+sex*age, method= "ML",
data = growth, random = ~ 1|ind)
```

- The output that follows is based on the maximum likelihood estimation.
- The intervals growth.fit2 command in R will produce the following confidence interval for the parameters of our model.

```
intervals(growth.fit2)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                 lower       est.     upper
## (Intercept) 21.919283 24.5681818 27.217080
## sex         -3.526164 -1.6931818  0.139800
## age10       -2.106542  0.5010569  3.108655
## age12        1.500635  3.7784091  6.056183
## age14        4.000635  6.2784091  8.556183
## sex:age10   -1.621355  0.1440557  1.909467
## sex:age12   -2.462636 -0.9346591  0.593318
## sex:age14   -3.212636 -1.6846591 -0.156682
##
##  Random Effects:
##   Level: ind
##                     lower     est.    upper
## sd((Intercept)) 1.318468 1.803986 2.468293
##
##  Within-group standard error:
##    lower     est.    upper
## 1.172963 1.381634 1.627428
```

- The summary(growth.fit2) command in R will produce the following output for the parameters of our model

```
summary(growth.fit2)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: growth
##        AIC      BIC    logLik
##   418.2713 444.2225 -199.1357
##
## Random effects:
##  Formula: ~1 | ind
##         (Intercept) Residual
## StdDev:    1.803986 1.381634
##
## Fixed effects:  measure ~ sex + sex * age
##                 Value Std.Error DF   t-value p-value
## (Intercept) 24.568182 1.3838177 66 17.753915  0.0000
## sex         -1.693182 0.9282931 25 -1.823973  0.0801
## age10        0.501057 1.3622421 66  0.367818  0.7142
## age12        3.778409 1.1899375 66  3.175300  0.0023
## age14        6.278409 1.1899375 66  5.276251  0.0000
## sex:age10    0.144056 0.9222729 66  0.156196  0.8764
## sex:age12   -0.934659 0.7982344 66 -1.170908  0.2458
## sex:age14   -1.684659 0.7982344 66 -2.110482  0.0386
##  Correlation:
##             (Intr) sex    age10  age12  age14  sx:g10 sx:g12
```

- The maximum likelihood is the estimation method that was used.
- The AIC and log likelihood can be used to make comparisons between models with different fixed effects (or random effects).
- The next estimates for the random effects part of the model.
- In the line where the numerical estimates appear, the label is StdDev, indicating that standard deviations are displayed.
- The estimates displayed are the standard deviations of between variability $(\sigma_b = 1.74)$ and the standard deviations of within variability $(\sigma_w = 1.369)$.
- In the Fixed Effects sections, we have the reported value of the intercept, its estimated standard error, and Wald test for whether its value is significantly different from zero or not.

Random Slope Model:

- Random intercept: `random = 1|ind`
- Random intercept and slope: `random =~ age|subject`

```
growth.fit2 <- lme(fixed = measure ~ sex + sex * age,
                   data = growth, random = ~ age | ind)
```

VarCorr(growth.fit2), variance components can be extracted from the model

```
VarCorr(growth.fit2)
```

```
## ind = pdLogChol(age)
##             Variance  StdDev    Corr
## (Intercept) 4.9698458 2.2293151 (Intr) age10  age12
## age10       1.2232671 1.1060140  0.122
## age12       3.1594142 1.7774741 -0.267 -0.287
## age14       4.0894536 2.0222397 -0.501  0.132  0.690
## Residual    0.4456557 0.6675745
```

Inference for the Marginal Model

- Having fitted a marginal model using maximum likelihood, we can use standard inferential tools for performing hypothesis testing:
    - Wald test
    - t-test / F-test
    - Score test
    - Likelihood ratio test (LRT)
    - Robust Inference
- 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
- 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):

The **likelihood ratio test (LRT)** is calculated as follows:

$$LRT = -2\left(\ell(\theta_0) - \ell(\theta_a)\right)$$

Where:

- $\ell(\theta_0)$ is the value of the log-likelihood function under the null hypothesis, i.e., the special case model.
- $\ell(\theta_a)$ is the value of the log-likelihood function under the alternative hypothesis, i.e., the general model.
- p denotes the number of parameters being tested.

Note: Provided that the mean structure in the two models is the same, we can either compare the REML or ML likelihoods of the models (preferably REML).

- We can rewrite the two hypotheses as:

$$LRT = -2\left(\ell(\theta_0) - \ell(\theta_a)\right) \sim \chi^2_p$$

$$H_0 : \begin{cases} \sigma_{12} = \sigma_{22} = \sigma_{32} = \sigma_{42} = \sigma^2 \\ \sigma_{12} = \sigma_{13}, = \cdots = \sigma_{34} = \tilde{\sigma} \end{cases}$$

$$H_0 : At\ least\ one\ variance\ or\ covariance\ is\ not\ equal\ to\ others.$$

Inference for the Marginal Model:

- When we have non-nested models, we cannot use standard tests anymore.
- When we compare two non-nested models, we choose the model that has the lowest **AIC/BIC** value.
- 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.

Information Criteria:

- LR tests can only be used to compare nested models.
- How to compare non-nested models?
- The general idea behind the **LR test** for comparing model A to a more extensive model B is to select model A if the increase in likelihood under model B is small compared to the increase in complexity.
- A similar argument can be used to compare non-nested models A and B.
- One then selects the model with the largest (log-)likelihood provided it is not (too) complex.
- Criterion: **Akaike (AIC), Schwarz (SBC)**.

AIC:

- For each model compute **AIC** = $-2\log(L) + kp$.
- p is the number of parameters in the model.
- L is the likelihood.
- k is a constant (often 2). k can be seen as a penalty for additional parameters. k between 2 and 6. The recommendation is to use a larger k with a small sample.
- Has theoretical basis in the prediction of future data.
- In practice, it sometimes overfits and chooses models that are sometimes too large.

Notes on IC:

- **Information criteria are not formal testing procedures!**
- The AIC and BIC do not always select the same model – when they disagree:
    - AIC typically selects the more elaborate model, whereas BIC selects the more parsimonious model.
- For the comparison of models with different mean structures, **IC should be based on ML rather than REML**, as otherwise the likelihood values would be based on different sets of error contrasts, and therefore would no longer be comparable.

Hypothesis Testing for Regression Coefficients:

- Hypothesis testing on β **:** We assume that 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:

- Tests for individual coefficients are based on the Wald-type statistic but assume the t-distribution for calculating p-values.
- The hypothesis is:

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

- We use the t test statistic:

$$\frac{\hat{se}(\beta)}{\hat{\beta}} \sim t_{df}$$

- df specified according to the number of subjects and the number of repeated measurements per subject.

Tests for Group Coefficients:

- Tests for group coefficients are based on the F-test.
- The hypothesis is:

$$H_0 : L\beta = 0 \; vs \; H_a : L\beta \neq 0$$

  where L is the contrasts matrix.

- The numerator df are always equal to the rank of the contrast matrix L.
- Denominator df need to be estimated from the data using methods such as the 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.

Notes on Hypothesis Testing for Regression Coefficients:

- Hypothesis testing for the regression coefficients β:
  - The **likelihood ratio test**, and the classical univariate and multivariate **Wald tests** (using the χ^2 distribution instead of the t or F distributions), are 'liberal'.
  - They give **smaller p-values** than they should give, especially in **small samples**.
- The LRT for comparing models with different β parts is only valid when the models have been fitted using maximum likelihood and not REML.

For Studies with Small Samples:

- **P-values of Wald test may be too small.**
- **Confidence intervals may be too narrow.**
- **t and F-tests** may be used to remedy this.
- Difficulty is with determining the df.

Inference for the Variance Components:

- Inference for the mean structure is usually of primary interest.
- However, inferences for the covariance structure are of interest as well:
  - Interpretation of the **random variation** in the data.
  - **Over-parameterized covariance structures** lead to inefficient inferences for the **mean**.
  - Too restrictive models invalidate inferences for the **mean structure**.
- The reported p-values often do not test meaningful hypotheses
- The reported p-values are often wrong.
- The sample size requirements for these tests are excessive & often not met (approximately 400 or more subjects).

Likelihood Ratio Test

- After a candidate model is selected, a LRT can be computed by comparing the candidate model with the reduced model.
- The mean structure of the model remains the same across both models, but the number of random effects is reduced by one in the reduced model.
- Note: as long as models are compared with the same mean structure, a valid LR test can be obtained under REML as well.
- Note: if H_0 is a **boundary value**, the classical χ^2 approximation may not be valid.
- For some very specific null-hypotheses on the boundary, the correct asymptotic null-distribution has been derived.
- Example: for the infant survival data, testing whether the **variance components** associated with the **random time effect** are equal to zero is equivalent to testing

$$H_0 : d_{12} = d_{22} = 0$$

Case 1: No Random Effects vs One Random Effect

- **Hypothesis of Interest:**

$H_0 : D = 0$ vs $H_a : D = d_{11}$ for some scalar d_{11}

- Asymptotic null distribution equals $-2 \ln \lambda_N \to \chi^2_{0:1}$, the mixture of χ^2_0 and χ^2_1 with equal weight 0.5.

Case 2: One vs Two Random Effects

- **Hypothesis of Interest:**

$$D = \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

for $d_{11} > 0$, versus H_a that D is a 2 by 2 positive semi definite matrix.

- Asymptotic null distribution: $-2 \ln \lambda_N \to \chi^2_{1:12}$, the mixture of χ^2_1 and χ^2_2 with equal weight 0.5.

Models for Non-Gaussian Longitudinal Data

- GEE
- GLMM

Recap: Marginal (Population Average) Models:

- Responses are marginalized over all other responses.
- Parameters characterize the **marginal expectation**.
- Inferences based on the marginal model do not explicitly assume the presence of **random effects** representing the natural heterogeneity between subjects.
- The implied marginal model equals:

$$\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_i\beta, \mathbf{Z}_i D\mathbf{Z}_i' + \Sigma_i.$$

Subject-Specific Models:

- If the aim is to study how subjects change overtime and what characteristics influence such changes.
- Subject-specific models differ from marginal models by the inclusion of parameters specific to the subject.

$$\mathbf{Y}_i^T\mathbf{b}_i \sim \mathcal{N}(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i, \Sigma_i),$$

where $\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D}_i)$.

Recap: Generalized Linear Models

- LMM have the assumption that the conditional responses are normally distributed.
- Normality assumption may not always be **reasonable**, i.e., **non-Gaussian responses**.
- Different methodology should be used when responses are **discrete**.
- Suppose we have a dichotomous outcome, Y, measured cross-sectionally.
- We are interested in making statistical inferences for this outcome, e.g.:
  - Is there any difference between placebo and treatment corrected for the age and sex of the patients?
  - Which factors best predict the outcome?

Generalized Linear Models

- Suppose we have a dichotomous outcome, Y, measured cross-sectionally.

$$\log\left(\frac{\text{odds of success}}{1-\text{odds of success}}\right) = \log(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- Odds of success = $\pi_i/(1-\pi_i) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$
- A unit change in X_1 from x to $x+1$ (while all other covariates are held fixed) corresponds to $\exp(\beta_1)$.

Example Toenail Data

- **Toenail Dermatophyte Onychomycosis:** Common toenail infection, difficult to treat, affecting more than **2%** of the population.
- Classical treatments with **antifungal compounds** need to be administered until the whole nail has grown out healthy.
- New compounds have been developed which reduce treatment to **3 months**.
- Randomized, double-blind, parallel group, multicenter study.

Toenail Data

- **Research question: style="color:red">Severity relative to treatment of TDO, coded as** .
- The question of interest was whether the **percentage of severe infection decreased** over time, and whether that evolution was different for the **two treatment groups**:
  - 2 × 189 patients randomized, 36 centers.
  - 48 weeks of total follow-up (12 months).
  - 12 weeks of treatment (3 months).
  - Measurements at months 0, 1, 2, 3, 6, 9, 12.

Toenail Data

```
library(readr)
toenail <- read_csv("Data/Toenail.csv")
```

```
library(gtsummary)
toenail %>% select(time, treatn) %>%
  tbl_summary(by=treatn)
```

Characteristic	[1]">**0**, N = 937[1]	[1]">**1**, N = 970[1]
time		
0	146 (16%)	148 (15%)
1	141 (15%)	147 (15%)
2	138 (15%)	145 (15%)
3	132 (14%)	140 (14%)
6	130 (14%)	133 (14%)
9	117 (12%)	126 (13%)
12	133 (14%)	131 (14%)

[1] n (%)

Recap: Generalized Linear Models

```
glm(formula = y ~ treatn * time, family = binomial, data = toenail)
```

```
##
## Call:  glm(formula = y ~ treatn * time, family = binomial, data = toenail)
##
## Coefficients:
## (Intercept)       treatn         time  treatn:time
##    -0.55706      0.02358     -0.17693     -0.07798
##
## Degrees of Freedom: 1906 Total (i.e. Null);  1903 Residual
## Null Deviance:        1980
## Residual Deviance: 1812     AIC: 1820
```

- $\beta_3 = -0.078$ is the interaction effect.
- Borderline significance (p = 0.048) which means that trends might be different in the two treatment groups.
- This analysis is wrong since we have not taken into account the multiple responses per subject.
- We have considered 1907 independent observations (residual degrees of freedom + 1).

```
length(toenail$y)
```

```
## [1] 1907
```

Generalized Estimating Equations (GEE)

- We return our focus on repeated measurements data, namely, repeated categorical data
  - we need to account for the correlations

Reminder: In the marginal models for continuous multivariate data we took account of the correlations by incorporating a correlation matrix in the error terms

- ML and REML

Challenges for Non-Gaussian Data

- For non-Gaussian 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
- Popular alternative approaches, like Generalized Estimating Equations (GEE), Alternating Logistic Regression (ALR), and Pseudo-Likelihood (PL) have been formulated.

Generalized Estimating Equations (GEE)

- GEE introduced by Liang and Zeger (Biometrika, 1986) is a best way to model longitudinal data in the marginal modeling framework for categorical responses.
- 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:

- S_β represents the score vector.
- μ_i is the mean of Y_i, and for dichotomous data $\mu_i = \pi_i$.
- V_i is a diagonal matrix with the variance of Y_i, e.g., for dichotomous data $V_{i-1} = diag(\pi_i(1 - \pi_i))$.

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:

- A_i is a diagonal matrix with the standard deviations $\sqrt{var(Y_i)}$.
- $R_i(\alpha)$ represents a 'working' assumption for the pairwise correlations.
- This approach follows the same form as the full likelihood procedure but restricts the specification to the first moment only.

Generalized Estimating Equations (GEE)

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

$$\hat{\beta} \sim \mathcal{N}(\beta, \mathbf{var}(\hat{\beta}))$$

  where $\mathbf{var}(\hat{\beta}) = V_0^{-1} V_1 V_0^{-1}$ is called the Sandwich or Robust estimator.

  - $V_0 = \sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$
  - $V_1 = \sum_i \frac{\partial \mu_i}{\partial \beta} \mathbf{var}(Y) V_1^{-1} \frac{\partial \mu_i}{\partial \beta}$

  V_0 and V_1 are often referred to as the "bread" and "meat" of the Sandwich estimator.

- GEE provides consistent regression coefficient estimates even if the correlation structure is miss-specified.

- A poor choice of working correlation matrix can affect the efficiency of the estimators of β.

Sandwich/Robust vs Naive/Purely Model-Based Standard Errors

- Software often also reports the Naive/Model-Based standard errors.
- These standard errors assume that the working correlation matrix is correctly specified.
- The Sandwich/Empirically Corrected/Robust standard errors correct for a possible misspecification of the correlation structure, although at the expense of power.

- A correct guess $\Longrightarrow$ likelihood variance.

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

    - It is an estimation method
- No assumptions for the joint distribution of repeated measurements $\Longrightarrow$ Semi-parametric approach
- The method relies solely on assumptions about the mean response
    - 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 class.
- **Fitting algorithm**:
  - Fit a Generalized linear model $\Longrightarrow$ **choose a working correlation matrix** $\Longrightarrow$ update $\hat{\beta}$, the covariance, and the correlation matrix.
- Interest is primarily in the βs, the covariance structure is considered as "nuisance"

  – Assumptions for the correlation are not supposed to be correct.
- This has implications for Hypothesis testing:
  - Likelihood ratio test or score test not applicable.

  Why? (semi-parametric approach) $\Rightarrow$ The Wald test can be used.
- Care needed with **incomplete data**.

GEE in R

- In R there are two main packages for GEE analysis, namely gee and geepack.
- The main function to fit GEEs is geeglm() – this has similar syntax as the glm() function of base R that fits GLMs.
- The major difference between gee and geepack is that geepack contains an ANOVA method that allows us to compare models and perform Wald tests.

Using Toenail data

Variables in the data:

- obs: observation number
- treat: treatment group (0: Itraconazole (group B); 1: Lamisil (group A))
- id: subject identification number
- time: time at which the observation is taken (months)
- response: the response measured (1: severe infection; 0: no severe infection)

Research question: Does treatment have an effect in curing the infection or not?

A function that fits GEE to deal with correlation structures arising from repeated measures on individuals, or from clustering as in family data is:

```
gee(formula, family, data, corStructure = "ar1", clusterID, startCoeff,
    maxit = 20, checks = TRUE, display = FALSE, datasources)
```

- `formula`: a string character which describes the model to be fitted.
- `family`: description of the error distribution: 'binomial', 'gaussian', 'Gamma', or 'poisson'.
- `data`: the name of the data frame that holds the variables.
- `corStructure`: the correlation structure: 'ar1', 'exchangeable', 'independence', 'fixed', or 'unstructured'.
- `clusterID`: the name of the column that holds the cluster IDs.
- `startCoeff`: a numeric vector, the starting values for the beta coefficients.
- `maxit`: an integer, the maximum number of iterations to use for convergence.

To fit GEE in R, you need the following packages first: **geepack, wgeesel, MuMIn**

```
library(geepack)
fit1 <- geeglm(y ~ treatn + time + treatn*time, id = idnum, data = toenail,
               family = binomial, corstr = "exchangeable", scale.fix = TRUE)
fit2 <- update(fit1, corstr = "ar1")
fit3 <- update(fit2, corstr = "unstructured")
summary(fit1)
```

```
##
## Call:
## geeglm(formula = y ~ treatn + time + treatn * time, family = binomial,
##     data = toenail, id = idnum, corstr = "exchangeable", scale.fix = TRUE)
##
##  Coefficients:
##              Estimate   Std.err   Wald Pr(>|W|)
## (Intercept) -0.586598  0.173669 11.409 0.000731 ***
## treatn       0.008405  0.261981  0.001 0.974406
## time        -0.177229  0.031209 32.249 1.36e-08 ***
## treatn:time -0.087144  0.056993  2.338 0.126260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
```

Choosing the Best Model

Random-Effects Models for Discrete Outcomes

- In the previous session, we focused on marginal models (GEE):
  - It can be seen as the extension of the marginal models for continuous data.
- In this session, we will see the analogue of linear mixed models for categorical data.
- **Random-effect models:** $E(Y_{ij})/b_i, x_{ij}$

Generalized Linear Mixed Models (GLMMs)

- GLMMs = GLMs (Logistic, Poisson, etc) with random effects.
- The intuitive idea behind GLMMs is the same as in LMMs, i.e.,
  - The correlation between the repeated categorical measurements is induced by unobserved random effects.
  - The categorical longitudinal measurements of a subject are correlated because all of them share the same unobserved random effect (conditional independence assumption).

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

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta + Z_i b_i, \quad b_i \sim \mathcal{N}(0, D)$$

- **Random effects** account for between-subject variability.

Three-Part Specification for GLMMs

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

2. The conditional mean of y_{iii} depends upon fixed and random effects via the following expression: $\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = x_i^T\beta + z_i^T b_i$

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

Notes 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.
  - This has direct and important implications with respect to the interpretation of the parameters!

Estimation of GLMMs

- 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.
- No REML.
- Nevertheless, there is an important complication in GLMMs.
- The fitting of GLMMs is a computationally challenging task!

Log-Likelihood Expression for GLMMs

- What is the problem?
- The log-likelihood expression for GLMMs has the same form as in LMMs:

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

  where θ are the parameters of the model.

- In linear mixed effects models, both terms in the integrand - $p(y_i/b_i;\theta)$ and $p(b_i;\theta)$ - are densities of (multivariate) normal distributions, and also because y_i and b_i are linearly related.

- The integral in the log-likelihood expression has a closed-form solution.

- **What is the problem?**
- In GLMMs, the two terms of the integrand denote densities of different distributions. For example, in mixed effects logistic regression:
  - $p(y_i/b_i;\theta) \Longrightarrow$ Bernoulli distribution
  - $p(b_i;\theta) \Longrightarrow$ Multivariate Normal distribution
- The implication is that in GLMMs, the same integral does not have a closed-form solution.

The Solutions

To overcome this problem, two general types of solutions have been proposed:

- **Approximation of the integrand:** This entails approximating the product inside the integral (i.e., $p(y_i/b_i;\theta)p(b_i;\theta)$) 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., $\int p(y_i/b_i;\theta)p(b_i;\theta)$) by a sum:
  - Gaussian Quadrature (GQ) & Adaptive Gaussian Quadrature (AGQ)
  - Monte Carlo & MCMC (Bayesian approach)

- From the two alternatives, methods that rely on approximation of the integral (GQ, AGQ) 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 GQ rules, it is the number of quadrature points (nGQ=1) point is equivalent to the Laplace approximation.
    - AGQ needs fewer quadrature points than classical GQ but is more time-consuming.
    - The Laplace approximation is a good choice when dealing with many repeated measures per subject.
- The higher nGQ (nAGQ), the more accurate the approximation will be.
- For more details on this, refer to Molenberghs and Verbeke (2005, Sections 14.3–14.5).

Estimation of Random Effects in GLMMs

- Estimation of the random effects proceeds in a similar manner as in linear mixed models.
- Predictions of random effects can be based on the posterior distribution $f(b_i/Y_i = y_i)$.
  - Empirical Bayes (EB) estimate:
  - May be used when interest is in predicting subject-specific evolutions.
  - Identifying subjects with outlying evolutions.
  - Estimation is based on the posterior distribution of b_i.
  - Posterior mode used as estimate.
- With EB estimates, $\hat{b}_i$, subject-specific probability profile:

$$\widehat{P(Y_{ij} = 1|b_i)} = \frac{\exp(X_{ij}\hat{\beta} + Z_{ij}\hat{b}_i)}{1 + \exp(X_{ij}\hat{\beta} + Z_{ij}\hat{b}_i)}$$

Interpretation of Parameter Estimates

- For the Linear Mixed Model (LMM): $E(Y_{ij}/b_i) = X_i\beta + Z_i b_i$
  - The marginal expectation $E(Y_{ij}) = E\left(E(Y_{ij}/b_i)\right) = X_i\beta$
  - β in LMMs has both conditional and marginal interpretation.
- For the Generalized Linear Mixed Model (GLMM), this does not hold:
  - Expectation of the conditional expectation of the GLMM:

$$E(E(Y_{ij}/b_i)) = E\left[\frac{\exp(X_i\beta + Z_i b_i)}{1 + \exp(X_i\beta + Z_i b_i)}\right] \neq \frac{\exp(X_i\beta + Z_i b_i)}{1 + \exp(X_i\beta + Z_i b_i)}$$

- Parameter vector β now does not have a marginal interpretation!

Conditional Interpretation of β in GLMMs

- β in GLMMs has a **conditional interpretation**.
- The parameters are **conditional on the random effects**.
- Interpretation of the fixed-effects coefficients:
  - For example, e^{β} does not have the interpretation of the average Odds Ratio (OR) for a unit increase in follow-up.
  - The parameters are **conditional on the random effects**.

GLMMs in R

- Packages: 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

GLMMs in R: lme4

- Fits a mixed effects logistic regression for Toenail data with random intercepts and 15 quadrature points for the adaptive Gauss-Hermite rule:

```
library(lme4)
glmmFit <- glmer(y ~ treatn*time + (1 | idnum), family = binomial(),
                 data = toenail, nAGQ = 15)
summary(glmmFit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 15) [glmerMod]
##  Family: binomial  ( logit )
## Formula: y ~ treatn * time + (1 | idnum)
##    Data: toenail
##
##      AIC      BIC   logLik deviance df.resid
##   1257.1   1284.9   -623.6   1247.1     1902
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
##  -2.96  -0.19  -0.09  -0.01  38.15
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  idnum  (Intercept) 16.5     4.07
```

GLMMs in R: GLMMadaptive

- The function that fits GLMMs in GLMMadaptive is `mixed_model()`. To fit the same model as we did above with `glmer()`, the code is:

```
library(GLMMadaptive)
glmmFit2 <- mixed_model(y ~ treatn*time, random = ~ 1 | idnum, family = binomial(),
                        data = toenail, nAGQ = 15)
summary(glmmFit2)
```

```
##
## Call:
## mixed_model(fixed = y ~ treatn * time, random = ~1 | idnum, data = toenail,
##     family = binomial(), nAGQ = 15)
##
## Data Descriptives:
## Number of Observations: 1907
## Number of Groups: 294
##
## Model:
##  family: binomial
##  link: logit
##
## Fit statistics:
##  log.Lik  AIC  BIC
##   -623.6 1257 1276
```

GLMMs in R

- 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.

Model Building

Model building for GLMMs proceeds in the same manner as for LMMs, 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 interaction terms.
- Following that, we build up the random-effects structure, starting from random intercepts, and then potentially including random slopes, quadratic slopes, etc.
- At each step, we perform Likelihood Ratio Tests (LRTs) to determine if including the additional random effect improves the fit of the model.
- After choosing the random-effects structure, we return to the fixed effects and assess whether the specification can be simplified.
- Once again, we start by testing complex terms (i.e., interactions and nonlinear terms), and then proceed to drop explanatory variables if required.
- In practice, 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

Summary: GEE

Coefficients relating Y to X Inference valid in large samples even if distribution of Y and/or variance of Y are incorrectly specified Valid inference if data are Missing Completely At Random (MCAR) even if variance model is wrong

GLMM

Coefficients relating Y to X conditional on b Valid inference generally requires correct specification of distribution of Y and of variance of Y Valid inference if data are Missing At Random (MAR)

GEE: The Jimma Infant Data

The response variable is categorized body mass index.

$$Y_{ij} = \begin{cases} 1 & \text{if } weight \leq 2500 \\ 0 & \text{otherwise} \end{cases}$$

The following model is assumed for the mean structure:

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij}), \text{ for subject } i \text{ and measurement } j,$$

Exchangeable correlation (or CS)

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Gender_i + \beta_3 Gender_i Age_{ij}$$

$Gender_i$ is a gender indicator. Age_{ij} is age of the i^{th} infant at time j (also the time variable).

Fitting GEE model in R

```
fit1 <- geeglm(BMIBIN ~ sex + age + sex * age, id = ind, data = Infant,
               family = binomial, corstr = "exchangeable", scale.fix = TRUE)
```

GEE Modeling with Different Working Correlation Structures

GEE can be modeled using different working correlation structures:

- `ar1` for autoregressive order 1
- `unstructured` for unstructured working correlation structure

Here's the R code to update GEE models using different correlation structures:

```
fit2 <- geeglm(BMIBIN ~ sex + age + sex * age, id = ind, data = Infant,
               family = binomial, corstr = "ar1", scale.fix = TRUE)
```

```
fit3 <- geeglm(BMIBIN ~ sex + age + sex * age, id = ind, data = Infant,
               family = binomial, corstr = "unstructured", scale.fix = TRUE)
```

- There is no effect of gender, age and gender and age interaction,

GLMM: The Jimma Infant Data

A random-effects model for non-Gaussian longitudinal data is applied to the Jimma Infant Data. The following model is assumed for the mean structure:

- $Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$, for subject i and measurement j,
- Gaussian distributed random intercepts b_i, i.e., $b_i \sim N(0, d)$, can be included to capture the correlation.

The logit of π_{ij} is modeled as:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Gender_i + \beta_3 Gender_i Age_{ij} + b_i$$

```r
# GLMM (random intercept)
fitGLMM <- glmer(BMIBIN ~ sex + age + sex * age + (1 | ind),
                 data = Infant, family = binomial(link = "logit"), nAGQ = 25)
summary(fitGLMM)
```

Calculate odds ratio and 95% confidence interval

```
odds_ratio_ci <- confint(fitGLMM, method = "profile", oldNames = FALSE)
odds_ratio_ci
```

Calculate and display odds ratio and 95% confidence interval

```
odds_ratio_ci <- exp(confint(fitGLMM, method = "profile", oldNames = FALSE))
odds_ratio_ci
```

R CODE for odds ratio:

```
exp(cbind(ODDS=coef(fitGLMM), confint(fitGLMM)))
```

Model with Random Intercept and Slope

The model with random intercept and slope can be fitted similarly. The following model is assumed for the mean structure:

- $Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$, for subject i and measurement j
- Gaussian distributed random intercepts b_i, i.e., $(b_{0i}, b_{1i}) \sim \text{N}(0, D)$, can be included to capture the correlation.

The logit of the probability π_{ij} is given by:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 Age_{ij} + \beta_2 Gender_i + \beta_3 Gender_i Age_{ij} + b_{0i} + b_{1i} Age_{ij}$$

R CODE for random intercept and slope model:

```
fitGLMMSlope <- glmer(BMIBIN~ sex + age + sex*age + (1+age|ind),
data = Infant, family = binomial(link = "logit"))
summary(fitGLMMSlope)
```

Day 4: Missing Data Management

- What is missing data – definition, patterns, mechanisms (MCR, MR, NMR)
- Simple methods for handling missing data
- Multiple Imputation (MI) based procedures
- Weighted GEE

Introduction

- Missing data is very common in statistical analysis.
    - is a problem that commonly encounter in research.
- In longitudinal studies, not all individuals are observed at all time points or visits.
- Some individuals may be observed only at one time point, or at first two or three, and so on.
- Having too much missing data can invalidate a study because
    1. It reduces statistical power,
    2. causes bias in estimation of parameters,
    3. reduces the representatives of the samples, and
    4. it complicates analyses of the studies.

What to consider

- Dealing with missing data requires considering the missing data **patterns**, **mechanisms**, **proportion** and the chosen analytic approach.
- **Missing Data Pattern**: It is essential to examine the missing data pattern, such as whether it follows a monotone, intermittent, or arbitrary structure. Understanding the pattern can guide the selection of suitable methods.
- **Proportion of Missing Data**: The proportion of missing data in the dataset is crucial. High levels of missingness may impact the validity of analyses and may require more sophisticated handling techniques.
- **Reasons for Missing Data**: Understanding the reasons why data are missing is critical. Missingness can occur due to various factors like participant dropouts, measurement errors, or system malfunctions. Identifying the reasons can help mitigate potential biases.
- ensure robust analyses in longitudinal studies when there is missing data.

Missing Data in Longitudinal Studies

Monotone Missing Data Pattern

- In longitudinal studies, missing data patterns can have various structures.
- One common pattern is the monotone missing data pattern, where not all individuals are observed at all time points, leading to dropouts in the data.

The table representing the monotone missing data pattern for a longitudinal studies:

Study	Time1	Time2	Time3
1	X	X	X
2	X	X	.
3	X	.	.

Intermittent/Arbitrary Missing Data Pattern

- In longitudinal studies, another common missing data pattern is the intermittent or arbitrary pattern, where data is missing at different time points for various individuals.

The table representing the arbitrary missing data pattern for longitudinal studies:

Study	Time1	Time2	Time3
1	X	X	X
2	X	.	X
3	.	X	X
4	.	.	X
5	.	X	.

- Proper data handling, including data imputation, is crucial to ensure reliable subsequent data analysis.

Types of Missing Data

- Looking carefully the causes of missingness enable us to employee the **appropriate missing data management system**.
    - Estimation of the parameter with missing data depends on the missing data mechanism.
- The missing data mechanism is a probability model for missingness.
- Data is often described in accordance to **the reasons for the missing data**.
- According to the mechanisms of missingness, we assume **three types** of missing data.
    - Missing Completely at Random (MCAR)
    - Missing at Random (MAR)
    - Not Missing at Random (NMAR)

1. Missing Completely at Random (MCAR)

- In MCAR, missingness is assumed to be **independent of both observed and unobserved data**.
- The notation of the MCAR mechanism is expressed as follows:

$$p(R|Y) = p(R|Y^{obs}, Y^{mis}) = p(R)$$

- Where:
  - Y is a vector of partially observed data, that is, $Y = (Y^{obs}, Y^{mis})$.
  - R is a set of missing indicators, i.e., $R = 1$ if the j^{th} element of Y is observed, and $R = 0$ if the j^{th} element of Y is missing (Rubin, 1976).
- MCAR is also known as **ignorable missing** in statistical inference.

Examples of MCAR data include:

- Data that are missing by design.
- Data with missing values due to equipment failure.
- Samples lost in transit.
- Data that is technically unsatisfactory.

Advantages of MCAR:

- The missing data does not introduce any bias in statistical analyses.
- Estimated parameters are not biased as a result of the missing data.

disadvantage of MCAR:

- Statistical power may be decreased due to the loss of information from the missing data.

2. Missing at Random (MAR)

- Missingness depends only on observed components Y^{obs}, not on missing components Y^{mis}.
  - Expressed through the formula:

$$p(R|Y^{obs}, Y^{mis}) = p(R|Y^{obs}).$$

  - The missingness pattern is completely determined by the observed data.
  - The missingness mechanism does not depend on the actual missing values.

Ignore Missing Data Mechanism

- In likelihood inference under MAR, the missing data mechanism can be ignored.
- Conduct likelihood-based inference without explicitly accounting for missing data.
- Missingness pattern does not bias estimation of model parameters.

Unbiased Inference in MAR

- MAR does not produce statistical bias.
- Missingness depends only on observed data.
- The missing data mechanism can be ignored in likelihood inference.
- Consequently, estimates of parameters remain unbiased even with missing data.

Validation of MAR Assumption

- It is crucial to validate the MAR assumption.
- Ensure that the observed data sufficiently inform the missing data mechanism.
- If systematic differences between observed and missing data exist, the MAR assumption may not hold.

- **Example**: In an educational assessment scenario:
  - Data collected: Children's health and test scores.
  - Missingness due to illness: Predictable from health data.
  - Unrelated to test score: Missingness depends only on observed data.
  - Illustrates the Missing at Random (MAR) mechanism.
- **Unbiased Inference**:
  - Under MAR: Unbiased inference is possible.
  - Sufficient observed data: Missingness accounted for.
  - Informative data: Assumption relies on data being informative.
  - Systematic differences: Unaccounted differences may invalidate MAR.
- **Sensitivity Analyses**:
  - Assessing robustness under missing data assumptions.
  - Including MAR assumption.
  - Understanding impact if MAR is violated.
- **Summary**: MAR's value in statistical inferences:
  - Validity despite missing data.
  - Missingness depends solely on observed data.
  - Missing data mechanism ignored in likelihood inference.
  - Essential to validate MAR assumption and explore sensitivity.

3. Not Missing at Random (NMAR)

- NMAR suggests that the probability of a value being missing fluctuates for reasons unknown to us.
- NMAR occurs when the characteristics of missing data do not meet those of MCAR and MAR mechanisms.
- In NMAR, the probability of missingness is influenced by both the observed value (Y^{obs}) and the unobserved missing value (Y^{mis}).
- The NMAR missing data mechanism can be represented by the following equation:

$$p(R|Y^{obs}, Y^{mis}) = p(R|Y^{obs}, Y^{mis})$$

- Unlike MCAR and MAR, where missingness can be explained by observed data alone, NMAR introduces an additional dependency on the unobserved missing values.

Non-Ignorable Missingness

- Due to the presence of this dependency, NMAR is considered a non-ignorable missing data mechanism.
    - **ignorable** missingness refers to the likelihood-based inference can proceed without considering the missing data model.
- However, in NMAR, the missing data mechanism is **non-ignorable**, and it must be explicitly incorporated into likelihood inference.

Example: Student Tutorial Attendance

- A concrete example of NMAR is when a student skips a tutorial lesson because the student knows that the attendance would not be graded.
- In this situation, the probability of missingness (skipping the tutorial) depends on both the observed data (whether the tutorial is graded or not) and the unobserved data (whether the student actually attended the tutorial).
- The missingness is not random and is influenced by the student's decision-making process.

Diagnosing and Handling NMAR

- Diagnosing NMAR can be challenging since the reasons for missingness are unknown and cannot be directly observed.
- Handling NMAR requires careful consideration and specialized methods to address missingness properly.
- Researchers often use sensitivity analyses and model-based imputation techniques to account for the non-ignorable missing data mechanism in NMAR scenarios.
- The goal is to explore how different assumptions about the missing data mechanism may impact the study's results and conclusions.

Common Methods for dealing with missing data

Ad-hoc Methods

- A number of theories were introduced to account missing data.

1. **Listwise Deletion**

- The most commonly used approach that data scientists use to deal with missing data is to simply omit cases with missing data, only analysing the rest of the dataset.
- This method is known as listwise deletion or complete-case analysis.
- The **na.omit()** function in R removes all cases with one or more missing data values in a dataset.
- Let's create a new small dataframe as an example to demonstrate Listwise Deletion

Ad-hoc Methods

Example for *Complete Case Analysis (Listwise Deletion)* example using the `sleepstudy` dataset from the `lme4` package in R:

1. Install and load the required packages and dataset:

```
#install.packages("lme4")
library(lme4)
data(sleepstudy)
set.seed(42)
# Create a new dataframe with missing data
missData <- sleepstudy
missData$Reaction[sample(1:nrow(missData), 50)] <- NA
missData1 <- missData
```

- Then perform Complete Case Analysis:

```
complete_data <- na.omit(missData1)
```

- Now fit a linear mixed-effects model to the complete data:

```
model_complete <- lmer(Reaction ~ Days + (Days | Subject), data = complete_data)
print(model_complete)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: complete_data
## REML criterion at convergence: 1270
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 23.1
##           Days         5.3     0.29
##  Residual             26.3
## Number of obs: 130, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##       253.9         10.2
```

- Remember that CCA (Listwise Deletion) involves removing cases with any missing data, and it may lead to biased results if the missingness is related to the outcome or predictors.
- Consider using more advanced methods like multiple imputation or mixed-effects models to handle missing data more effectively.

```
model_miss <- lmer(Reaction ~ Days + (Days | Subject), data = missData1)
print(model_miss)
```

Last Observation Carried Forward (LOCF)

- LOCF is an imputation method for longitudinal data, which involves taking the previous observed value to replace missing values in the dataset.
- This method is especially useful for data to be plotted for time series analysis.
- The LOCF can be seen as a good method of choice because it produces a complete dataset and it can be used for cases where we know what the missing values should be.
- To perform the LOCF imputation method, we simply run the
  - **fill()** function from the tidyr package on the dataset and the variable with the missing data.
  - **lna.locf()** function in zoo

```
# Load required packages
library(dplyr); library(tidyr)  # Required for fill() function
set.seed(400)
missData <- sleepstudy
missData$Reaction[sample(1:nrow(missData), 30)] <- NA

# LOCF Imputation using dplyr and tidyr
missData <- missData %>%
  group_by(Subject) %>%
  fill(Reaction, .direction = "down")
```

```r
# Fit a linear mixed-effects model to the imputed data
model_locf <- lmer(Reaction ~ Days + (Days | Subject), data = missData)
print(model_locf)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: missData
## REML criterion at convergence: 1683
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 25.63
##           Days         6.32    -0.04
##  Residual             26.00
## Number of obs: 173, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##       246.5         10.9
```

3. Mean Imputation

- Some data scientists or statisticians may look for a quick fix by replacing missing data with the mean.
- Mode is often used to impute categorical data.
- We use `mice package` with argument **method = mean**, & **m = 1.**

```
# Introduce missing data for demonstration purposes
set.seed(42)
sleepstudy$Reaction[sample(1:nrow(sleepstudy), 30)] <- NA

# Perform mean imputation
mean_value <- mean(sleepstudy$Reaction, na.rm = TRUE)
sleepstudy$Reaction_imputed <- ifelse(is.na(sleepstudy$Reaction), mean_value,
                                      sleepstudy$Reaction)
```

```r
# Fit a linear mixed-effects model to the imputed data
model_mean <- lmer(Reaction_imputed ~ Days + (Days | Subject), data = sleepstudy)
print(model_mean)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction_imputed ~ Days + (Days | Subject)
##    Data: sleepstudy
## REML criterion at convergence: 1778
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 23.27
##           Days         3.55    0.42
##  Residual             29.89
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##      260.22         8.99
```

- To perform mean imputation for missing data in R, the `mice` package is used with `method = "mean"` argument..

- The `method` argument in `mice()` allows you to specify the imputation method to be used. - Setting `method = "mean"` indicates that you want to perform mean imputation for the missing values.

```
library(mice)
library(lme4)
data(sleepstudy)
set.seed(42)
sleepstudy$Reaction[sample(1:nrow(sleepstudy), 30)] <- NA
# Perform mean imputation using mice()
imputed_data <- mice(sleepstudy, method = "mean", m=1, print = FALSE)
model_mean_imputed <- with(imputed_data, lmer(Reaction ~ Days + (Days | Subject)))
```

- Then we can use `summary(pool())` to obtain the pooled results across the multiple imputed datasets.

```r
# Print the summary of the model
print(pool(model_mean_imputed))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
## REML criterion at convergence: 1778
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 23.27
##           Days         3.55    0.42
##  Residual             29.89
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##      260.22         8.99
```

```r
# Pool the results from multiple imputed datasets
pooled_model <- pool(model_mean_imputed)
# Print the pooled summary of the model
print(pooled_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
## REML criterion at convergence: 1778
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 23.27
##           Days         3.55    0.42
##  Residual             29.89
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##      260.22         8.99
```

Multiple Imputation

- Multiple imputation is a process that is done in 3 main steps:
- This gives the imputed data a valid statistical inference.

Steps for Multiple Imputation:

- Firstly, generate m multiple imputed datasets.
- Secondly, analyze each imputed dataset, then there should be m analyses.
- Lastly, combine the results for the pooled dataset.
- Multiple imputation is robust to small sample sizes or lots of missing data.

- Steps in applying multiple imputation to missing data via the `mice` approach

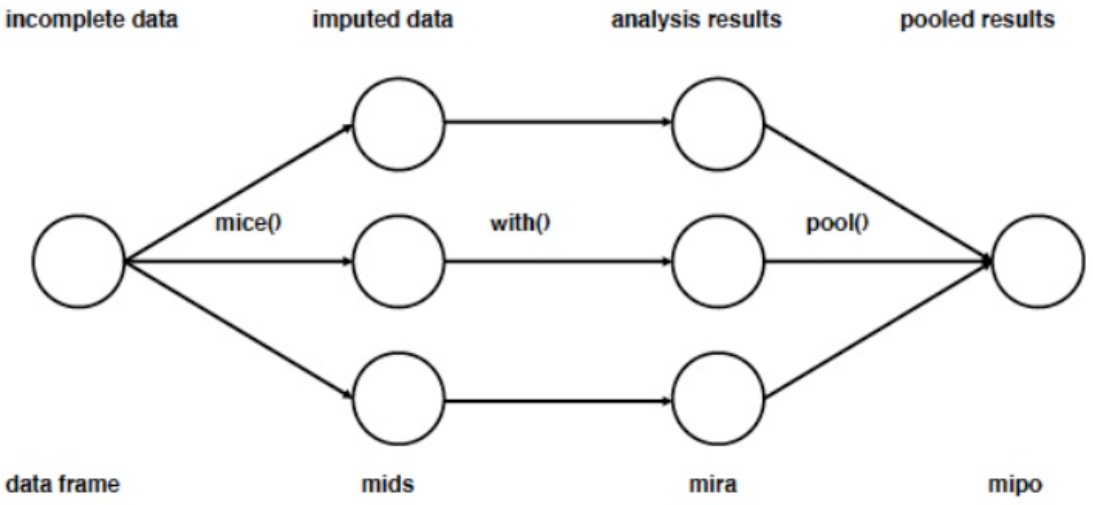


Figure 1: Main steps used in multiple imputation.

- The estimate of the parameter β is simply the average of each parameter estimate β^m obtained over the m imputed datasets (m = 1, ..., M):

$$\hat{\beta}^* = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}^m$$

- The variance of the estimator is partitioned into within imputation variance (sampling variability), and the between imputation variance (estimation variability due to missing data).
- The within imputation variance, W_β, over the m imputed datasets is:

$$W_\beta = \frac{\sum_{m=1}^{M} SE_\beta^2}{M}$$

- The between imputation variance, B_β, over the m imputed datasets is:

$$B_\beta = \frac{\left(\sum_{m=1}^{M} \left(\hat{\beta}^m - \hat{\beta}^*\right)^2\right)}{M-1}$$

- These two variances are combined to provide a single variance, given by

$$T_\beta = W_\beta + \left[\frac{(M+1)}{M}\right] B_\beta$$

- Let's look at it using the mice() function

```
data(sleepstudy); set.seed(42)
sleepstudy$Reaction[sample(1:nrow(sleepstudy), 30)] <- NA
# Perform Multiple Imputation
imputed_data <- mice(sleepstudy, m = 5, method = "pmm", seed = 123, print = FALSE)
# Analyze each imputed dataset separately (e.g., linear mixed-effects model)
for (i in 1:5) {
  model <- lmer(Reaction ~ Days + (Days | Subject), data = complete(imputed_data, i))
  print(print(model))}
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: complete(imputed_data, i)
## REML criterion at convergence: 1751
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 26.23
##           Days         4.85    0.28
##  Residual             26.48
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##         255           10
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: complete(imputed_data, i)
```

```r
# Pool the results from multiple imputed datasets
pooled_model <- pool(with(imputed_data, lmer(Reaction ~ Days + (Days | Subject))))

# Print the pooled summary of the model
print(pooled_model)
```

```r
# Extract imputed data in long format using tidyr
impdata <- complete(imputed_data, action = "long", include = FALSE)
# Pool the results from multiple imputed datasets
pooled_model <- pool(with(impdata, lmer(Reaction ~ Days + (Days | Subject))))

# Print the pooled summary of the model
print(pooled_model)
```

```r
# Load the sleepstudy dataset
data(sleepstudy)

# Introduce missing data for demonstration purposes
set.seed(42)
sleepstudy$Reaction[sample(1:nrow(sleepstudy), 30)] <- NA

# Perform Multiple Imputation
imputed_data <- mice(sleepstudy, m = 5, method = "pmm", seed = 123)

# Pool the results from multiple imputed datasets
pooled_model <- pool(imputed_data)

# Print the pooled summary of the model
print(summary(with(pooled_model, lmer(Reaction ~ Days + (Days | Subject)))))
```

```r
# Load the sleepstudy dataset
data(sleepstudy)
library(broom.mixed)
```

```
## Warning: package 'broom.mixed' was built under R version 4.3.1
```

```r
library(tidyr)
# Introduce missing data for demonstration purposes
set.seed(42)
sleepstudy$Reaction[sample(1:nrow(sleepstudy), 30)] <- NA

# Perform Multiple Imputation
imputed_data <- mice(sleepstudy, m = 5, method = "pmm", seed = 123, print = FALSE)
```

```r
# Extract the models from each imputed dataset and get tidy estimates
tidy_estimates <- lapply(1:5, function(i) {
  complete_data <- complete(imputed_data, action = i)
  model <- lmer(Reaction ~ Days + (Days | Subject), data = complete_data)
  tidy(model)
})
```

```r
# Combine tidy estimates using bind_rows
pooled_tidy_estimates <- bind_rows(tidy_estimates)

# Print the pooled tidy summary of the model
print(pooled_tidy_estimates)
```

```
## # A tibble: 30 × 6
##    effect   group    term                  estimate std.error statistic
##    <chr>    <chr>    <chr>                    <dbl>     <dbl>     <dbl>
##  1 fixed    <NA>     (Intercept)             255.       7.19     35.5
##  2 fixed    <NA>     Days                     10.0      1.33      7.50
##  3 ran_pars Subject  sd__(Intercept)          26.2     NA        NA
##  4 ran_pars Subject  cor__(Intercept).Days     0.278   NA        NA
##  5 ran_pars Subject  sd__Days                  4.85    NA        NA
##  6 ran_pars Residual sd__Observation          26.5     NA        NA
##  7 fixed    <NA>     (Intercept)             253.       6.45     39.3
##  8 fixed    <NA>     Days                     10.4      1.29      8.12
##  9 ran_pars Subject  sd__(Intercept)          22.7     NA        NA
## 10 ran_pars Subject  cor__(Intercept).Days     0.579   NA        NA
## # ℹ 20 more rows
```

Now we can extract the completed dataset using the complete() function.

```
#library(tidyr)
#impdata <- mice::complete(impData, action = "long", inc = F)

#View(airquality)
```

The missing values have been replaced with the imputed values in the first of the five datasets. If you wish to use another one, just change the second parameter in the complete() function.

con't

```
#densityplot(impData)
```

- The density of the imputed data for each imputed dataset is shown in magenta while the density of the observed data is shown in blue.

con't

Pooling

- Suppose that the next step in our analysis is to fit a linear model to the data.
- You may ask what imputed dataset to choose.
- The mice package makes it easy to fit a model to each of the imputed datasets and then pool the results together

```
# pool(with(impData, lm(Temp~ Ozone+ Solar.R+Wind)))
#summary(fit)
# summary(pool(fit))
# pool.r.squared(fit, adjusted = TRUE)
```

Multiple Imputation Software

- **`Amelia`** in R (by Gary King and collaborators)
- **`mi`** in R (by Andrew Gelman and collaborators)
- **`mice`** in R (by Stef van Buuren and collaborators)
- SPSS **`(Analyze > Multiple Imputation)`**
- STATA **mi `estimate`**

Outline

- What is Missing Data
  - Definition of Missing Data
  - Missing Data Patterns
  - Missing Data Mechanisms
- Simple Ways of Dealing with Missing Data
- Multiple Imputation of Missing Data

Introduction

- Missing data is very common in statistical analysis.
- However, having too much missing data can invalidate a study as it reduces power and precision.
- Dealing with missing data requires considering the missing data patterns, mechanisms, and the chosen analytic approach.

Common Types of Missing Data

- Survey nonresponse
- Missing dependent variables
- Missing covariates
- Dropouts
- Censoring - administrative, competing events, LTFU
- Non-reporting or delayed reporting
- Non-compliance
- Measurement error

Implications of Missing Data

- Unbalanced data
- Loss of information and reduced inference
- Extent depends on the amount of missing data, missing pattern, mechanism, parameters of interest, and analysis method
- Care is needed to avoid biased inferences

Missing Data Mechanisms

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Examples of Missing Data Mechanisms

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Implications of Missing Data Mechanisms

- Missing Completely At Random (MCAR): Complete Case (CC) analysis is acceptable
- Missing At Random (MAR): No CC; Likelihood-based methods and Generalized Estimating Equations (GEE) are okay
- Missing Not At Random (MNAR): No CC; analysis is difficult - use sensitivity analysis

Missing Data Patterns for Longitudinal Data

- Monotone/dropout missing data pattern: Missing data after the dropout, common in longitudinal studies
- Intermittent/Arbitrary missing data pattern: Missing data before some non-missing data, common in non-longitudinal studies

Common Techniques for Dealing with Missing Data

1. Complete Case Analysis (Listwise deletion)
2. Available Case Analysis (Pairwise deletion)
3. Unconditional Mean Imputation (Mean substitution)
4. Single Imputation (Deterministic Imputation)
5. Stochastic Imputation

Multiple Imputation

- Multiply impute "m" pseudo-complete data sets, typically using a small number of imputations (e.g., 5 < m < 10).
- Combine the inferences from each of the m data sets to acknowledge the uncertainty in the imputation process and the missing data mechanism.

Multiple Imputation: Combining Inferences

- Combine m sets of parameter estimates to provide a single estimate of the parameter of interest.
- Combine uncertainties to obtain valid standard errors.
- Within-imputation variance and between-imputation variance.

Multiple Imputation versus Likelihood Analysis when Data are MAR

- Both multiple imputation and likelihood analysis are valid when data are MAR.
- The choice between them depends on the efficiency and complexity of the models.

What if You Doubt the MAR Assumption?

- Methods for Non-MAR (NMAR) data exist, but they require information and assumptions on pr(Missing I observed, unobserved).
- Sensitivity analysis can assess the stability of findings under various scenarios, setting bounds on the form and strength of the dependence.

First, make sure you have the required packages installed:

```
#install.packages("lme4")     # For the sleepstudy dataset
#install.packages("geepack")  # For weighted GEE
#install.packages("mice")     # For multiple imputation
```

Now, let's load the required libraries and simulate some missing data:

```
library(lme4)
library(geepack)
library(mice)

# Load the sleepstudy dataset
data("sleepstudy")
set.seed(42)

# Simulate missing data (replace 10% of values with NA)
sleepstudy_miss <- sleepstudy
n <- nrow(sleepstudy_miss)
sleepstudy_miss$Reaction[sample(n, n * 0.1)] <- NA
```

Next, we'll use the weighted GEE approach and multiple imputation for comparison:

```
# Weighted GEE approach
# Assuming MAR (missing at random) mechanism, we can use inverse probability weightin

# Create a missing indicator variable
```