

# Longitudinal Data Analysis Using R

Tadesse Awoke Ayele (PhD, Associate Professor)

University of Gonder

College of Medicine and Health Sciences

Institute of Public Health

June 11, 2020



# Contact Detail

- Tadesse Awoke (PhD, Associate Professor)
  - University of Gondar
  - Colleague of Medicine and Health Sciences
  - Institute of Public Health
  - Epidemiology and Biostatistics
  - Department: Epidemiology and Biostatistics
- Qualifications:
  - MSc Statistics, MSc Biostatistics
  - PhD in Biostatistics, PhD in Public Health
- Contacts:
  - Email: [tawoke7@gmail.com](mailto:tawoke7@gmail.com)
  - Mobile: +251910173308
  - Location: University of Gondar
  - Office hours: Available upon request

# Course Description

- This course focuses on the analysis of longitudinal data with continuous outcome variables. Step by step, we will be developing longitudinal data analysis techniques by extending the well-known multiple linear regression model for studies with repeated observations on the same respondents. This will result in the presentation of the mixed effects model (also known as multilevel model, random-effects model, hierarchical linear model, ...), allowing the analysis of change over time.
- Throughout the course lectures, the emphasis will be on understanding the why and how these models by explaining the underlying theory of these multilevel analyses using lots of examples. The application and interpretation of outcome of these techniques will be demonstrated in R.
- Students entering this course should have knowledge of and experience in using basic statistical concepts and techniques, including multiple linear regression analysis and analysis of variance.

# Learning Objectives

- At the end of the course, students will learn
  - Explore (graphically and numerically) longitudinal data and recognise the need for mixed effects models (multilevel models)
  - Understand the theory behind the multilevel model for change
  - Build, examine, interpret, expand and compare mixed effects models
  - Perform all described techniques using R (or other major statistical software packages, see above)
  - Test the assumptions of the models and examine the reliability of the findings
  - Summarize results for publications

# **Course Outline**

## Part I

# Models for Longitudinal Gaussian Data

# Chapter 1: General Introduction

- Research and Statistics
- Recap on Variable and data
- Source of Information, knowledge, and wisdom
- Longitudinal Data structure
- Choice of statistical models based on research questions
- Basic introduction to R

# Chapter 2: Introduction to Longitudinal Data Analysis

- General Introduction
- Motivating Examples
- Cross-sectional versus Longitudinal Data
- Simple Methods

# Chapter 3: Exploratory Data Analysis

- Exploring the Mean Structure
- Exploring the Random Effects
- Exploring the Correlation Structure
- Exploring the Variability of the Observed Data
  - Individual Profiles
  - Average Profile
  - Correlation Matrix

# Chapter 4: Models for Longitudinal Gaussian Data

- Linear Mixed Models
- A 2-stage Model Formulation
- Hierarchical versus Marginal Model
- Components of the Linear Mixed Effects Model
  - The Mean Structure
  - The Random Effects
  - Variance Structure
  - The Correlation Structures

# Chapter 5: Practical Guide Using R

- Exploratory data analysis using R
- Fitting Linear Mixed Effect Models in R
- Multivariate Regression Model and gls function in R
- Model diagnostics
- Extended Linear Modeling Approach

## Part II

# Models for Longitudinal Non-Gaussian Data

# Chapter 6: Introduction to Longitudinal non-Gaussian data

- General Introduction
- Motivating Examples
- Cross-sectional versus Longitudinal Data
- Simple Methods

# Chapter 7: Models for Correlated Binary Data

- Introduction to correlated binary data
- Generalized Estimating Equation
- Fitting Generalized Estimating Equation in R
- Generalized Linear Mixed Effect model
- Fitting Generalized Linear Mixed Effect Model in R
- Model Comparison

# Chapter 8: Models for Correlated Count Data

- Introduction to correlated count data
- Generalized Estimating Equation
- Generalized Linear Mixed Effect model
- Model Comparison
- Intra class correlation

## References

- NSS project in Biostatistics Series: Longitudinal Data Analysis in R
- Andrzej Gałecki and Tomasz Burzykowski (2013): Linear Mixed-Effects Models Using R: A Step-by-Step Approach
- Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data. New York: Springer.
- Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, Geert Molenberghs: Longitudinal Data Analysis: Handbooks of Modern Statistical Methods
- Peter Diggle, Patrick Heagerty, Kung-Yee Liang, Scott Zeger: Analysis of Longitudinal Data
- Brajendra C. Sutradhar: Longitudinal Categorical Data Analysis
- Liu, Xian: Methods and Applications of Longitudinal Data Analysis
- Jos W. R. Twisk. Applied Multilevel Analysis. A Practical Guide
- Fitzmaurice, Garrett M.,: Handbook of Missing Data Methodology
- Jiming Jiang (2007): Linear and Generalized Linear Mixed Models and Their Applications

# Teaching Methods and Evaluation

- Teaching Methods
  - Lecture
  - Exercise
  - Assignment
- Evaluation
  - Participation 5%
  - Assignment 15%
  - Project 20%
  - Final Exam: 60%

# Tentative Schedule

Chapters	Date
General Introduction	Day 1
<b>Part I: Models for Longitudinal Gaussian Data</b>	
Introduction to Longitudinal Data	Day 1-2
Exploratory Data Analysis	Day 2-3
Models for Longitudinal Gaussian Data	Day 3-4
Practical Guide using R	Day 5-6
<b>Part II: Models for Longitudinal Non-Gaussian Data</b>	
Introduction to non-Gaussian Longitudinal Data	Day 6-7
Model for Correlated Binary Data	Day 8-9
Model for Correlated Count Data	Day 10-11

# **Chapter 1**

## **Introduction**

# What is Research?

- Research is defined as the systematic
  - Collection
  - Organization
  - Analysis and
  - Interpretation
- of data for the purpose of
  - Answering a question
  - Solving a problem or
  - Adding body of knowledge
- The ultimate objective of research is to generate valid evidence for
  - Program planning
  - Policy Making
  - Intervention
  - Evaluation
  - Decision making
  - ...

## Illustration

- What do you know about Statistics/Biostatistics?
- Your experience before on biostatistics
- Any attempt of conducting research
- Statistics is the study and use of theory and methods for the analysis of data arising from random processes or phenomena
- It is the study of how we make sense of data

# History of Statistics

- Ancient civilizations counted their populations for taxation and military purposes
- Complete census were first carried out in: Sweden in 1749, USA in 1790, ..., Ethiopia 1984
- Statistics has grown through successive eras:
  - era of censuses
  - era of vital statistics
  - era of descriptive statistics
  - era of probability statistics
  - era of analytic statistics

# History of BioStatistics

- Alexandre Louis (1787-1872) introduced the numerical method in describing medical facts quantitatively
- Karl Pearson (1857-1936) introduced descriptive statistics, hypothesis and errors
- Sir Ronald Fisher (1890-1962) introduced methods for comparison of means, regression, and significant tests
- Francis Galton introduce applied statistical techniques to natural phenomena, described correlation and regression
- Neyman developed the concept of confidence intervals in 1934

# Limitation of Biostatistics

- The statistical conclusion is used with other knowledge to reach a substantive conclusion
- Statistics has several limitations
  - It gives statistical but not substantive answers
  - The statistical conclusion refers to groups and not individuals
  - It only summarizes but does not interpret data
- Statistics can be misused by selective presentation of desired results
- It is a tool that can be used well or can be misused
- The human must be able to intelligently interpret the output from the computer

## Limitation...

- When examining statistical information consider the following:
  - Was the sample used to gather the statistical data unbiased and of sufficient size?
  - Is the methodology (design) used appropriate?
  - Is the analysis technique appropriate for the data?
  - Is the statistical statement ambiguous, could it be interpreted in more than one way?

# Examples

## ① Probability of surviving surgery by physicians

- A patient undergoes examination and told to have surgery. He asked the doctor to tell him the probability of surviving. Then the doctor told him it is 100%. The patient asked how? The doctor said "of 10 patients who went through this operation, 9 died". The 9th patient died yesterday and you are the 10th patient.

## ② Harvard university example

- There were three female students at Harvard university. Of the three, one married her professor. Thus, the information 33% of female students married their professor at HU.

## ③ Soldiers height and crossing the river

# Introduction(1)

- 21st century is the period in which information becomes;
  - Money
  - Power
  - Everything
- Statistics starts with a problem, continues with the collection of data, proceeds with the data analysis and finishes with conclusions.
- Biostatistics provides the most fundamental tools and techniques of the scientific methods for generating information
- Used for:
  - Forming hypotheses
  - Designing experiments and observational studies
  - Gathering data
  - Summarizing data
  - Drawing inferences from data( eg. testing hypotheses)

## Introduction(2)

- The field of statistics can be divided into two:
  - ① Mathematical Statistics: study and develop statistical theory and methods in the abstract
  - ② Applied Statistics: application of statistical methods to solve real problems involving randomly generated data and the development of new statistical methodology motivated by real problems

Later Applied Statistics can be also subdivided in to two branches

- Descriptive statistics: describes what we have in hand (the sample)(sample mean, sample variance, sample proportion, ...)
- Inferential statistics: generalizes the finding from the sample to the population (estimation and hypothesis testing)

## Introduction(3)

- Biostatistics is the branch of applied statistics directed toward applications in the health sciences and biology
- Why biostatistics?
- Because some statistical methods are more heavily used in health applications than elsewhere (eg. survival analysis and longitudinal data analysis)
- Because examples are drawn from health sciences:
  - Makes subject more appealing to those interested in health
  - Illustrates how to apply methodology to similar problems encountered in real life

# Variable, Data and Information

- **Variable:** is a characteristic which takes different value
  - Quantitative variables
    - E.g. Number of children in a family, ...
    - E.g. Weight, height, BP, VL, ...
  - Qualitative variables
    - E.g. Marital status, religion,
    - E.G. Education status, patient satisfaction

# Types of Variable

- Variables can be classified into different types

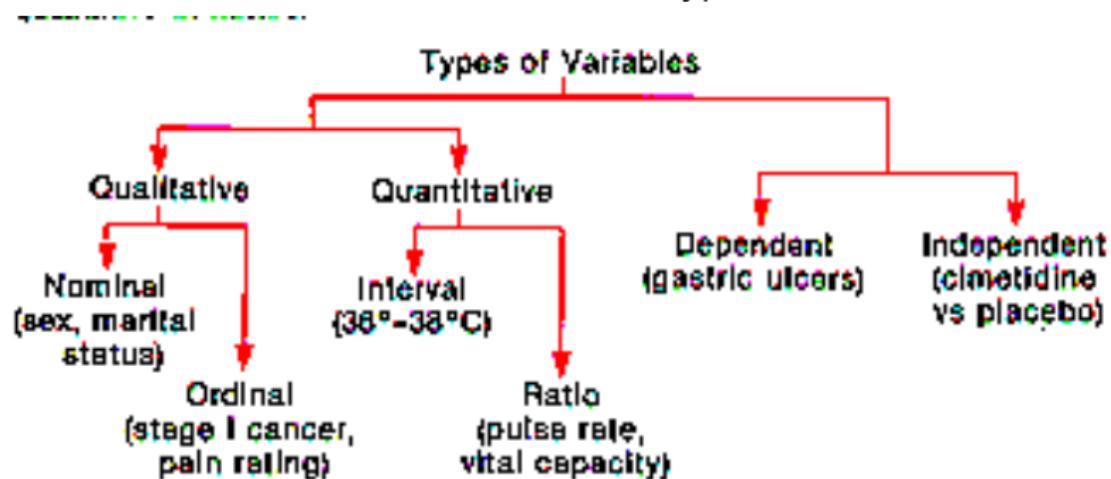


Figure 1: Variable Classification

# Types of Variables

- Another way to distinguish between variables is
  - Qualitative(Categorical)
  - Quantitative (numerical)
- Qualitative variables have values that are intrinsically non-numeric
  - Generally have either nominal or ordinal scales
  - eg. Cause of death, nationality, race, gender, severity of pain (mild, moderate, severe)
  - They can be reassigned numeric values (male=1, female=2 but they are still intrinsically qualitative)

## Quantitative variable

Quantitative variables can be further subdivided into discrete and continuous variables

- Discrete variables have a set of possible values that is either finite or countably infinite
  - eg. number of pregnancies, shoe size, number of missing teeth, e.t.c
- A continuous variable has a set of possible values including all values in an interval of the real line
  - eg. duration of a seizure, body mass index, height, e.t.c

- Variables can be again classified in to two broad categories

## Outcome variable

- Can be also called response or dependent variable
- It is the focus of the research
- Affected by other (independent) variables

## Predictor variables

- Can be also called explanatory or independent variable
- Affects the outcome variable

## Example 1

- In a study to determine whether surgery or chemotherapy results in higher survival rates for a certain type of cancer, whether or not the patient survived is one variable, and whether they received surgery or chemotherapy is the other.
  - Identify the outcome variable
  - Identify the predictor variable

### Solution

- Outcome variable
- Predictor variable

## Example 2

- Global Burden of Noncommunicable diseases and risk factors. They are by far the leading cause of death in the Region, representing 62% of all annual deaths.

NCD risk factors include:

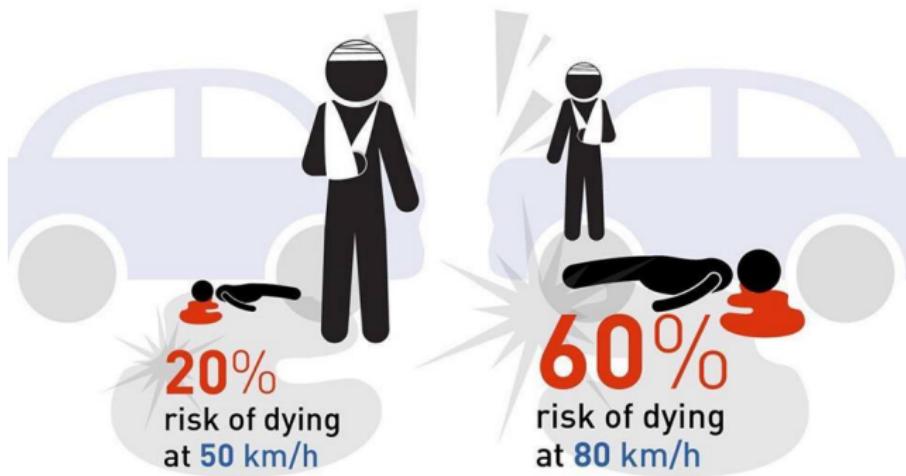
- Tobacco
- Harmful use of Alcohol
- Sedentary behaviour and physical inactivity
- Obesity
- Unhealthy diet.

# Schematic presentation



# Speed and risk of car accident

The higher the speed of the vehicle,  
the higher the **risk** of injury and death for pedestrians



World Health  
Organization



Save Lives  
SlowDown

Figure 3: Speed versus car accident.

- Data: Is a measurement (observation) taken about the variable
- The collection of data is often called dataset
  - Can be quantitative data or
  - Qualitative data

However, data is raw in which the required evidences can not be easily obtained.

## Illustration

- Data is raw, unorganized facts that need to be processed.
- When data is processed, organized, structured or presented in a given context so as to make it useful, it is called information.

### Data and information

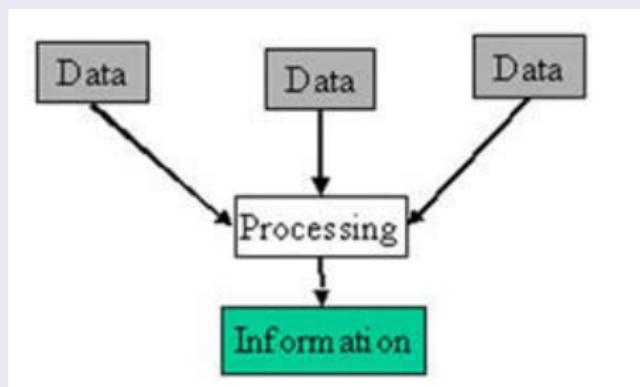


Figure 4: Relationship between data and information

- **Information:** Data that is
  - specific and organized for a purpose
  - presented within a context that gives it meaning and relevance and
  - can lead to an increase in understanding and decrease in uncertainty

Biostatistics/Statistics is the tool which converts data to information.

# Illustration

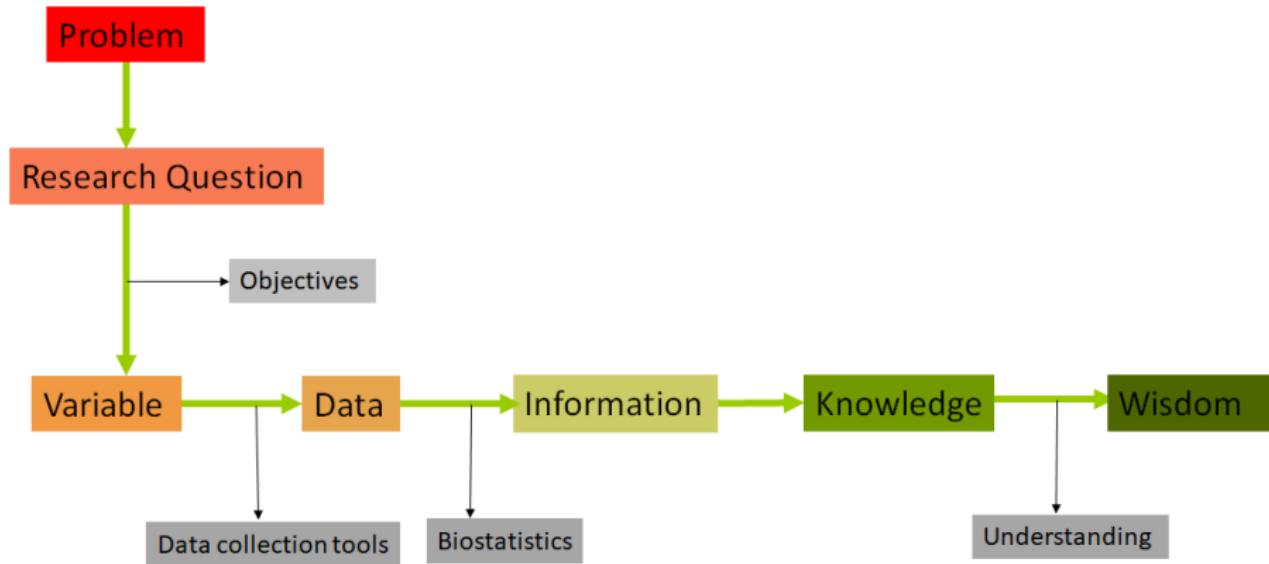


Figure 5: Schematic presentation

## Example 1: Jimma Child Mortality Data

The data were collected to establish risk factors affecting infant survival. Children born in Jimma, Kaffa, and Illubabor were examined for their first year growth characteristics. There were 8050 infants enrolled in the study. The variables included were:

ID	TotPrg	TotBrth	Abortion	StillIB	Gravida	Event	Durtaion	parity	MamAge
1	5	5	2	0	0	0	360	0	30
2	3	3	2	0	0	0	62	0	23
3	8	8	2	0	0	0	360	0	32
4	5	5	2	0	0	0	356	0	30
5	4	2	1	1	0	0	362	1	23
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
8050	0	4	4	2	0	0	361	2	35

- Objectives: To determine survival probability of new born
- Statistical approach.....?

## Example 2: Treatment Effect data

The data was collected on ulcer patients. The patients were randomized into two arms (placebo and treatment)

- The variables in the data are age, duration, treatment, time and result.
- Data

ID	Age	Duration	Treatment	Time	Result
1	48	2	1	7	1
2	73	1	1	12	0
3	54	1	1	12	0
4	58	2	1	12	0
5	56	1	0	12	0
6	49	2	0	12	0
.	.	.	.	.	.
.	.	.	.	.	.
42	61	1	0	12	0
43	33	2	1	12	0

- Objectives of the study: to determine the effect of treatment on ulcer
- Statistical Approach.....

## Example 3: HIV/AIDS data

- Follow-up data among HIV/AIDS in Gondar Hospital. The variables time, age, sex, residence, WHO stage and CD4 cell count were included.
- Data

ID	Months	Sex	Age	Residence	Stage	FCD4
1	1	1	20	1	2	100
1	2	1	20	1	2	64
1	3	1	20	1	2	611
1	4	1	20	1	2	744
2	1	0	39	2	3	166
2	2	0	39	2	3	363
2	3	0	39	2	3	263
2	4	0	39	2	3	164
3	1	0	48	1	2	361
.	.	.	.	.	.	.
.	.	.	.	.	.	.
2550	1	1	30	0	4	441

- Objectives: To assess the change in CD4 count over time.
- Statistical approach.....

## Infant growth data

- Children were followed for 12 months and measurement were taken every two months. Part of the data is given below;

Obs	child	age(months)	weight(grams)	CatBMI
1	1	0	2900	1
2	1	2	3100	0
3	1	4	3180	1
.	.	.	.	.
7	1	12	8000	1
8	2	0	3200	0
9	2	2	3340	0
.	.	.	.	.

- Objectives: To assess the change in weight over time.
- Statistical approach.....

## Gilgel-Gibe Mosquito Data

- A study conducted around Gilgel-Gibe dam for three years. Influence of the dam on mosquito abundance and species composition. Eight 'At risk' and eight 'Control' villages based on distance.

	ID	Village	Time	Gamb	Season	
1.	1	at risk	0	0	wet	
2.	1	at risk	1	0	wet	
3.	1	at risk	2	9	wet	
4.	1	at risk	3	30	wet	
6.	1	at risk	5	0	dry	
.	.	.	.	.	.	.

- Objectives: To compare the incidence of mosquito and characterize the type of species.
- Statistical approach.....

- In all the above examples the variables are presented in the first row
- The body of the table represents the data
- Every data is collected for a purpose (research question)
- Need to be analysis to generate information

# Introduction to R

# Introduction to R

- R is a language and environment for data manipulation, calculation and graphical display
- Initially written by Ross Ihaka and Robert Gentleman at Department of Statistics at the University of Auckland, New Zealand during 1990s
- Since 1997, an international “R-core” team of 15 people with access to common CVS archive was established
- Most functionality is provided through built-in and user-created functions and all data objects are kept in memory during an interactive session.
- Basic functions are available by default. Other functions are contained in packages that can be attached to a current session as needed

## R Interface

- Start the R system, the main window (RGui) with a sub window (R Console) will appear
- In the 'Console' window the cursor is waiting for you to type in some R commands.

# Introduction

- It's free!
- It runs on a variety of platforms including Windows, Unix and Mac
- It provides an unparalleled platform for programming new statistical methods in an easy and straightforward manner
- It contains advanced statistical routines not yet available in other packages
- It has state-of-the-art graphics capabilities

# Introduction

- The R distribution contains functionality for large number of statistical procedures including
  - All kinds of descriptive analysis
  - Linear and generalized linear models
  - Nonlinear regression models
  - Linear mixed effect models
  - Generalized linear mixed effect models
  - Time series analysis
  - Nonparametric tests
- R also has a large set of functions which provide a flexible graphical environment for creating various kinds of data presentations
- Since it is a programming language, generating computer code to complete tasks is required

# R and statistics

- Packaging: a crucial infrastructure to efficiently produce, load and keep consistent software libraries from (many) different sources / authors
- Statistics: most packages deal with statistics and data analysis
- State of the art: many statistical researchers provide their methods as R packages

# Installing R

- select a download site, <https://cran.r-project.org/bin/windows/base/>
- download the base package at a minimum
- Install the software with the base package
- download contributed packages as needed
- Install the required package for each output

# Data Source

- Flat Files



- Excel Files



- Statistical Software



- Databases



- Data from the Web



# Download R

The screenshot shows a web browser window with multiple tabs open. The active tab is titled "Download R-3.5.1 for Windows (32/64 bit)". The URL in the address bar is <https://cran.r-project.org/bin/windows/base/>. The page content includes a large button labeled "Click here to download" and a link "Download R 3.5.1 for Windows (62 megabytes, 32/64 bit)" which is circled in red.

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [md5sum](#) of the .exe to the [fingerprint](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

#### Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [RFAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

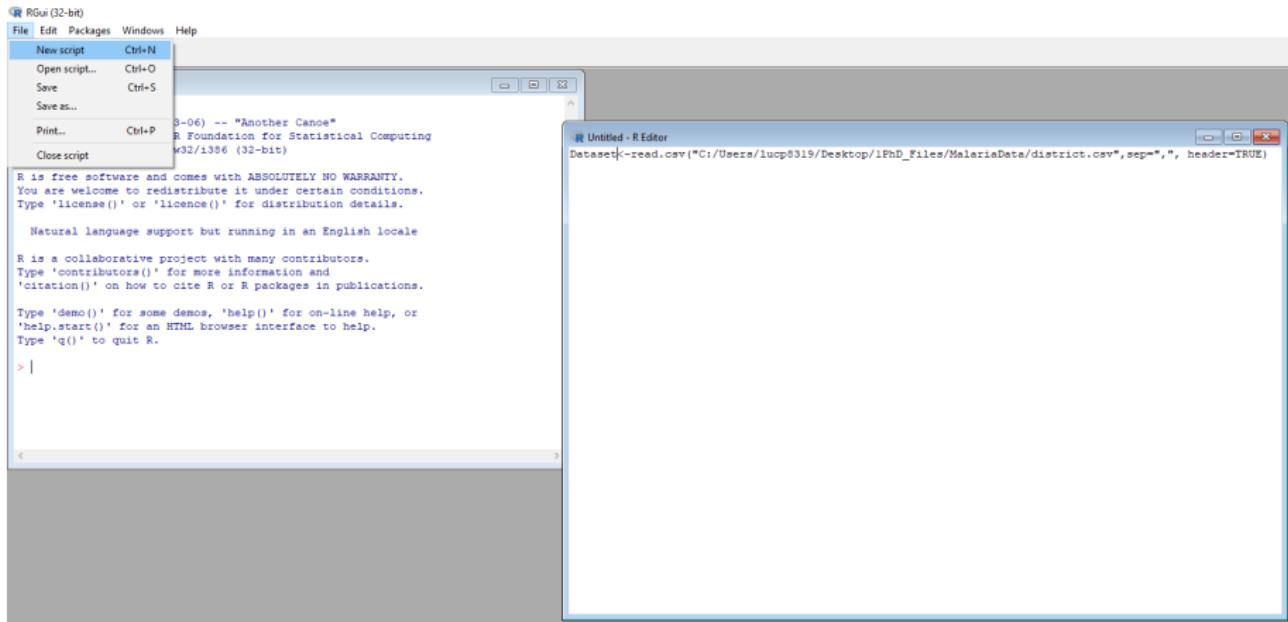
#### Other builds

- Patches to this release are incorporated in the [r-patched.snapshot.build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel.snapshot.build](#).
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is  
<https://CRAN.MIRROR.R-project.org/bin/windows/base/release.htm>

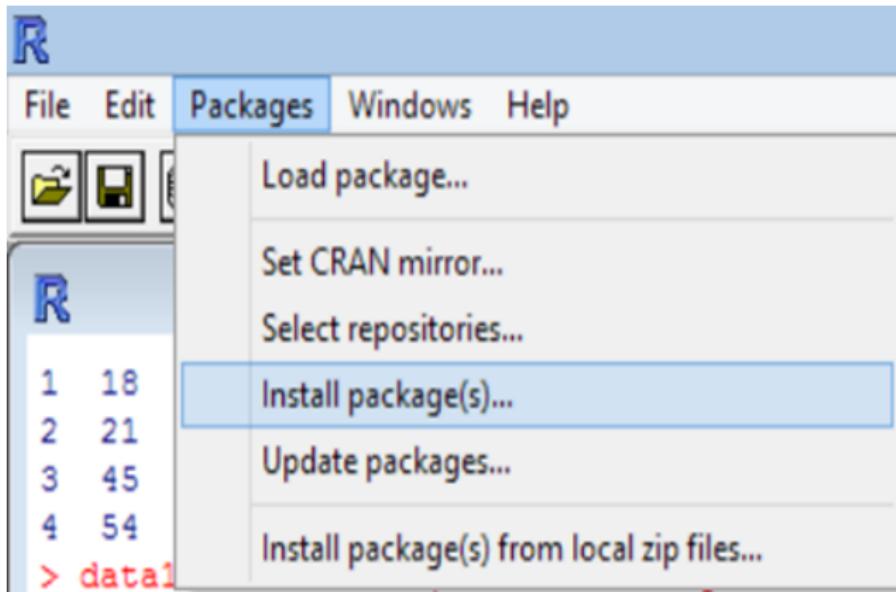
Last change: 2018-07-02

# R Window



## R Packages

- There are many contributed packages that can be used to extend R
- These libraries are created and maintained by the authors.



# Getting Started with R

- Similarly, R need data to perform its magic
- Different dataset are available in the R library
- write "library(" MASS")" and then run data()

R version 3.3.3 (2017-03-06) -- "Another Canoe"  
Copyright (C) 2017 The R Foundation for Statistical Computing  
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()'" on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

```
> library("MASS")
> data()
> |
```

Available datasets:

- fri
- forbes
- galaxies
- gehan
- genotype
- geyser
- glass
- hills
- housing
- injury
- leuk
- mammals
- mcycle
- manarche
- michelson
- insect
- motors
- muscle
- newcomb
- nlschools
- npk
- npri
- oats
- painters
- petrol
- phones
- quine
- road
- rotifer
- ships
- shoes
- shrimp
- shuttle
- snails
- stean
- stormer
- survey
- synth.te
- synth.tz
- topo
- waders
- whiteside
- welches

Measurements of Forensic Glass Fragments  
Forbes' Data on Boiling Points in the Alps  
Velocities for 82 Galaxies  
Remission Times of Leukaemia Patients  
Rat Genotype Data  
Old Faithful Geyser Data  
Soil Nitrogen Uptake in Gilgai Territory  
Record Times in Scottish Hill Races  
Frequency Table from a Copenhagen Housing  
Conditions Survey  
Yields from a Barley Field Trial  
Survival Times and White Blood Counts for  
Leukemic Patients  
Brain and Body Weights for 62 Species of Land  
Mammals  
Data from a Simulated Motorcycle Accident  
Age of Monarchs in Warsaw  
Michelson's Speed of Light Data  
Median Survival Time of Lung Cancer Patients of 1995  
Accelerated Life Testing of Motorcycles  
Effect of Calcium Chloride on Muscle  
Contraction in Rat Hearts  
Newcomb's Measurements of the Passage Time of  
Light  
Fourth-Grade Pupils in the Netherlands  
Classical M, F, K Factorial Experiment  
US Naval Petroleum Reserve No. 1 data  
Data from an Oats Field Trial  
The Painter's Data of oil Paints  
H. C. Prater's Petrol Refinery Data  
Religion-Phone Calls 1960-1973  
Absenteeism from School in Rural New South  
Wales  
Road Accident Deaths in US States  
Number of Notifiers by Fluid Density  
Ship Damage Data  
Spiral wave Pack of Box, Hunter and Hunter  
Percentage of Shrimp in Shrimp Cocktail  
Space Shuttle Autolander Problem  
Small Mortality Data  
The Saturated Steam Pressure Data  
The Stormer Viscometer Data  
Protein Solubility Data  
Synthetic Classification Problem  
Synthetic Classification Problem  
Spatial Topographic Data  
Counts of Waders at 15 Sites in South Africa  
House Insulation: Whiteside's Data  
Weight Loss Data from an Obese Patient

Use 'data(package = .packages(all.available = TRUE))'  
to list the data sets in all \*available\* packages.

# Sample data in R library

- There are numbers of sample datasets available in the R library
- This can be used for exercises



```
R : Copyright 2003, The R Development Core Team
Version 1.7.0 (2003-04-16)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

> library("MASS")
> data()
> data(Cars93)
> Cars93
   Manufacturer      Model     Type Min.Price Price Max.Price MPG.city
1       Acura    Integra Small    12.9  15.9    18.6    25
2       Acura Legend_Midsized  Midsize  29.2  33.9    38.7    18
3       Audi          90 Compact  25.9  29.1    32.3    20
4       Audi         100 Midsize  30.8  37.7    44.6    19
5       BMW        535i Midsize  23.7  30.0    36.2    22
6      Buick    Century Midsize  14.2  15.7    17.3    22
7      Buick    LeSabre Large   19.9  20.8    21.7    19
8      Buick Roadmaster Large   22.6  23.7    24.9    16
9      Buick    Riviera Midsize  26.3  26.3    26.3    19
10     Cadillac DeVille Large   33.0  34.7    36.3    16
11     Cadillac Seville Midsize  37.5  40.1    42.7    16
12    Chevrolet Cavalier Compact  8.5  13.4    18.3    25
13    Chevrolet Corsica Compact 11.4  11.4    11.4    25
14    Chevrolet Camaro Sporty  13.4  15.1    16.8    19
15    Chevrolet Lumina Midsize 13.4  15.9    18.4    21
16    Chevrolet Lumina_APV Van   14.7  16.3    18.0    18
17    Chevrolet Astro Van   14.7  16.6    18.6    15
18    Chevrolet Caprice Large   18.0  18.8    19.6    17
```

# Importing and exporting data

- Most data sets you will be working with will be stored as data frames
- There are many ways to get data into R and out of R.
- Most programs (e.g. Excel), as well as humans, know how to deal with rectangular tables in the form of tab-delimited text files.
- `> x = read.delim("filename.txt")`
- also: `read.table`, `read.csv`
- `> write.table(x, file="x.txt", sep=" ")`

## Reading data

- We can read data file into R as a vector
- The format of the data we want to read in R determine the code
- Consider a dataset named "example1.dat" and we can use the following code to read the dataset
- The number symbol indicates a programmer's comment
- This text is not read by the R application

```
dataset = scan("/Users/Shared/WD/Rdirectory/example1.dat")
#Change pathname to wherever you saved example1.dat
#Print data
> dataset
#Calculate the sample mean
> mean(dataset)
```

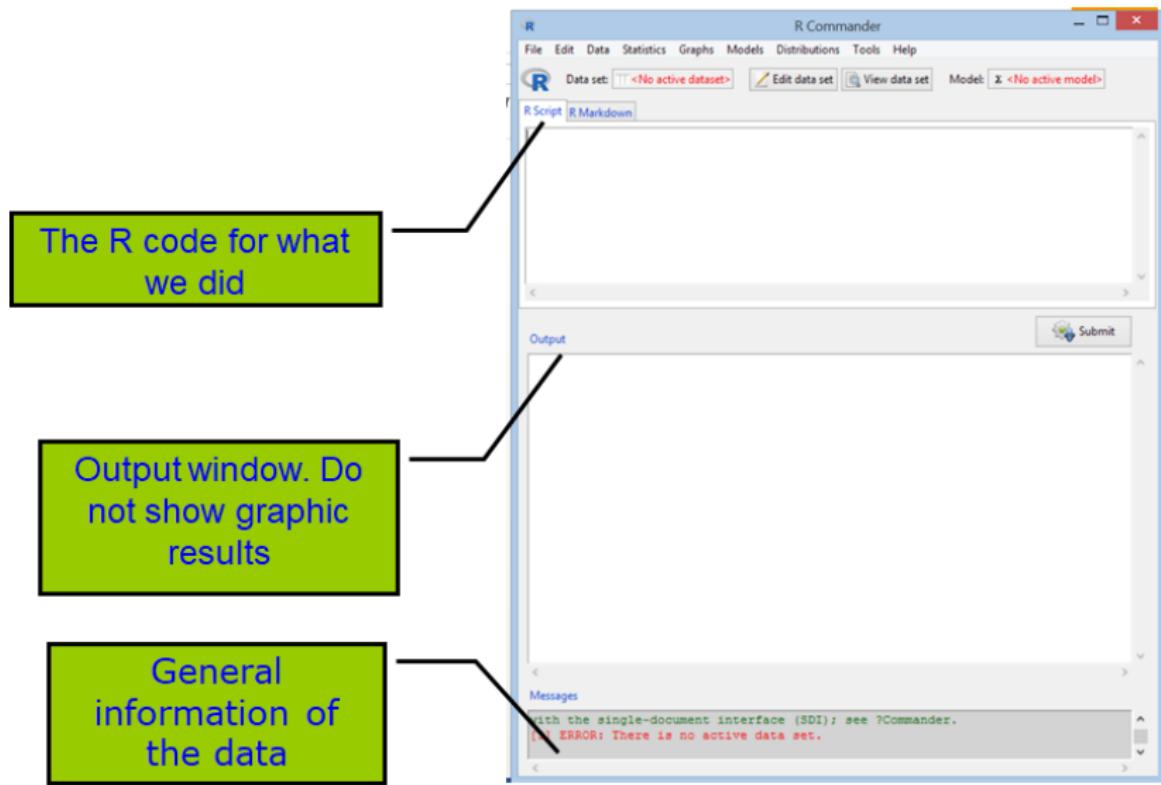
# R Commander

- R provides a powerful and comprehensive system for analyzing data
- R is a command-driven system, and new users often find learning R challenging
- R commander is free statistical software
- R commander was developed as an easy to use graphical user interface (GUI)
- For R to allow the teaching of statistics courses and removing the hindrance of software complexity from the process of learning statistics

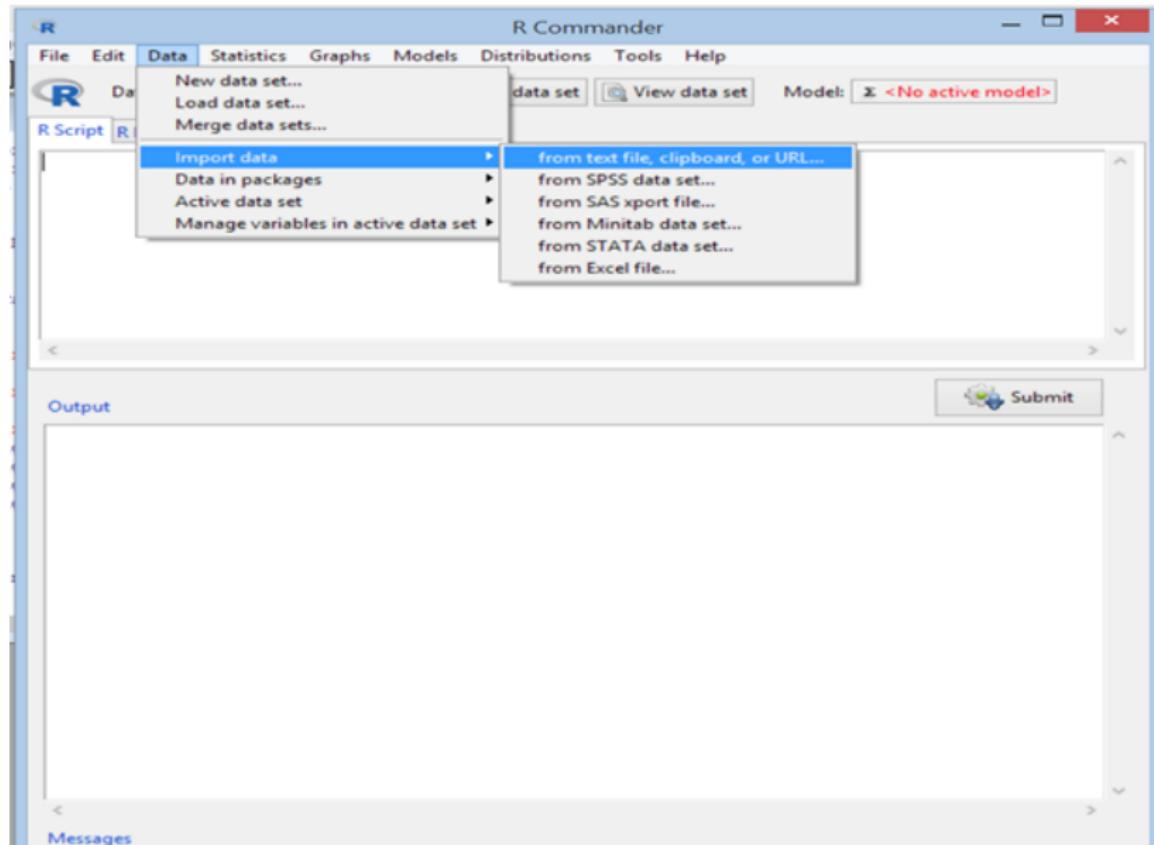
# Installing R Commander

- Once you have installed R, open it by double-clicking on the icon
- A window called “R Console” will open
- At the prompt (the *i* symbol), type the following command exactly and then press enter:
- > install.packages("Rcmdr", dependencies = TRUE)
- R should respond by asking you to select a mirror site, and listing them in a pop-up box.
- Choose a nearby location
- Depending on your connection speed, the installation may take awhile. Be patient and wait until you see the prompt again before you do anything.
- Type the following command at the command line prompt
- > library(Rcmdr) and hit return
- If successful, the GUI should open in a new window

# R Commander



# Importing data to R commander



# Data imported from SPSS

The screenshot shows the R Commander interface with the following details:

- Top Bar:** R Commander, File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, Help.
- Toolbar:** R Script, R Markdown, Data set: `jimma`, Edit data set, View data set, Model: 1 <No active model>.
- R Script Tab:** Contains R code for reading a SPSS file and creating a dataset named `jimma`. It includes setting column labels, fixing the dataset, and displaying it.
- Output Tab:** Shows the R console output corresponding to the script. It includes the same R code and a note indicating the dataset has 8050 rows and 29 columns.
- Submit Button:** Located in the Output tab area.
- Messages Tab:** Displays two messages: [4] NOTE: The dataset `jimma` has 8050 rows and 29 columns. [7] NOTE: The dataset `jimma` has 8050 rows and 29 columns.

# Managing Data

The screenshot shows the R software interface with the 'Data' menu open. The 'Data' menu has the following options:

- New data set...
- Load data set...
- Merge data sets...
- Import data
- Data in packages
- Active data set
- Manage variables in active data set

The 'Manage variables in active data set' option is highlighted with a blue selection bar. A secondary menu is displayed below it, listing the following options:

- Recode variables...
- Compute new variable...
- Add observation numbers to data set
- Standardize variables...
- Convert numeric variables to factors...
- Bin numeric variable...
- Reorder factor levels...
- Drop unused factor levels...
- Define contrasts for a factor...
- Rename variables...
- Delete variables from data set ...

In the background, the R script editor shows some R code. The output pane at the bottom shows the results of the R code execution.

# Summaries

The screenshot shows the SPSS Statistics software interface. The menu bar at the top includes File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, and Help. The Statistics menu is currently active and highlighted in blue. A sub-menu titled "Summaries" is open, listing several statistical analysis options: Active data set, Numerical summaries..., Frequency distributions..., Count missing observations, Table of statistics..., Correlation matrix..., Correlation test..., and Shapiro-Wilk test of normality... . To the left of the menu, a portion of the R Script editor is visible, containing R code related to data analysis.

- Active data set
- Numerical summaries...
- Frequency distributions...
- Count missing observations
- Table of statistics...
- Correlation matrix...
- Correlation test...
- Shapiro-Wilk test of normality...

# Frequency Distribution

The screenshot shows the RStudio interface. The top menu bar has 'Statistics' selected, with a submenu open showing options like 'Summaries', 'Contingency tables', 'Means', etc., and 'Frequency distributions...' highlighted. A red box highlights this option. To the right of the menu, a message says 'No active model'. The main workspace shows an R script with code for printing counts and percentages of parity. The 'Output' pane below shows the run results, with a red circle highlighting the first two lines of the 'counts:' output.

```
cat("\nper\nprint(round\n"))
local({
  .Table <-
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
})
```

Output

```
<11.49  >11.5
77.76  22.24

> local({
+   .Table <- with(jimma, table(parity))
+   cat("\ncounts:\n")
+   print(.Table)
+   cat("\npercentages:\n")
+   print(round(100*.Table/sum(.Table), 2))
+ })
```

counts:

parity	1	2-4	>4
1888	3624	2528	

percentages:

parity	1	2-4	>4
	24.48	45.07	31.44

Frequency distribution

# Contingency Table

The screenshot shows the RStudio interface. On the left is the R Script pane containing R code to generate a contingency table and perform a Chi-square test. On the right is the R Commander window titled "Two-Way Table". The "Data" tab is selected, showing the variables "catagemom", "catbwt", "catgravida", "muac", and "parity" under "Row variable (pick one)" and "catagemom", "catbwt", "catgravida", "muac", and "parity" under "Column variable (pick one)". A subset expression "<all valid cases>" is also present. Below the dialog are buttons for Help, Reset, OK, Cancel, and Apply. The R Console pane at the bottom displays the generated R code, the resulting frequency table, Pearson's Chi-squared test results, and two notes about the dataset. A red oval highlights the frequency table output.

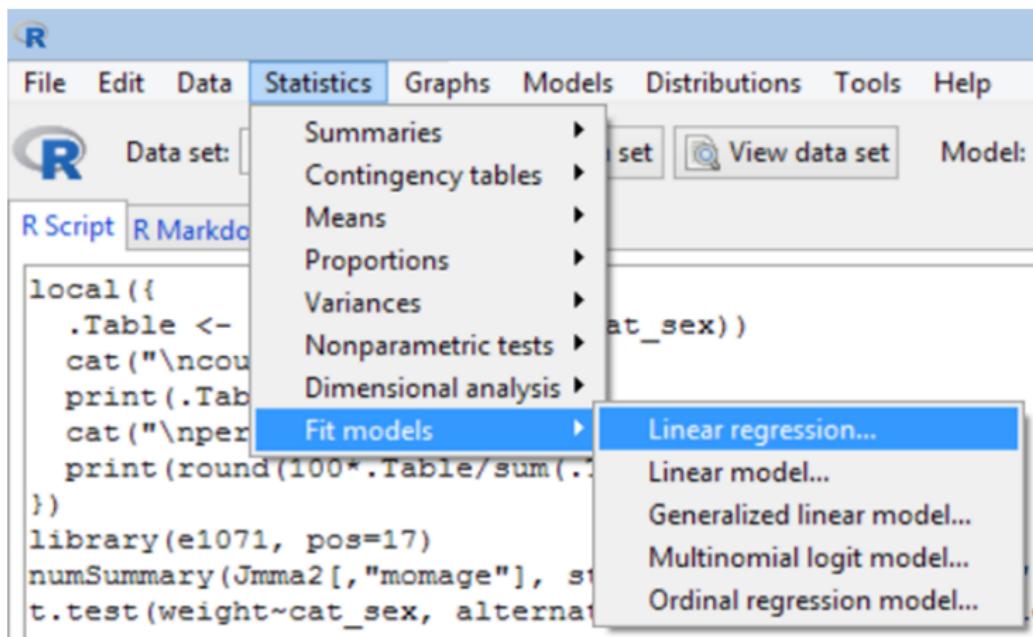
```
R
File Edit Data Statistics Graphs Models
R Data set: jimma Edit data s
R Script R Markdown
showData(jimma, placement=-20+2
summary(jimma)
library(abind, pos=16)
local({
  .Table <- xtabs(~catbwt+catgravida,
  cat("\nFrequency table:\n")
  print(.Table)
  .Test <- chisq.test(.Table, correct=FALSE)
  print(.Test)
})
<
>

Output
> local({
+   .Table <- xtabs(~catbwt+catgravida, data=jimma)
+   cat("\nFrequency table:\n")
+   print(.Table)
+   .Test <- chisq.test(.Table, correct=FALSE)
+   print(.Test)
+ })
>
> Frequency table:
>          catgravida
catbwt
  1      2-4     >4
  200 and above 1532 3283 2392
  <2500        214  239  213
>
> Pearson's Chi-squared test
> data: .Table
> X-squared = 45.517, df = 2, p-value = 1.307e-10
<
>

Messages
[6] NOTE: The dataset jimma has 8050 rows and 29 columns.
[7] NOTE: The dataset jimma has 8050 rows and 29 columns.
```

Chi-square test of association

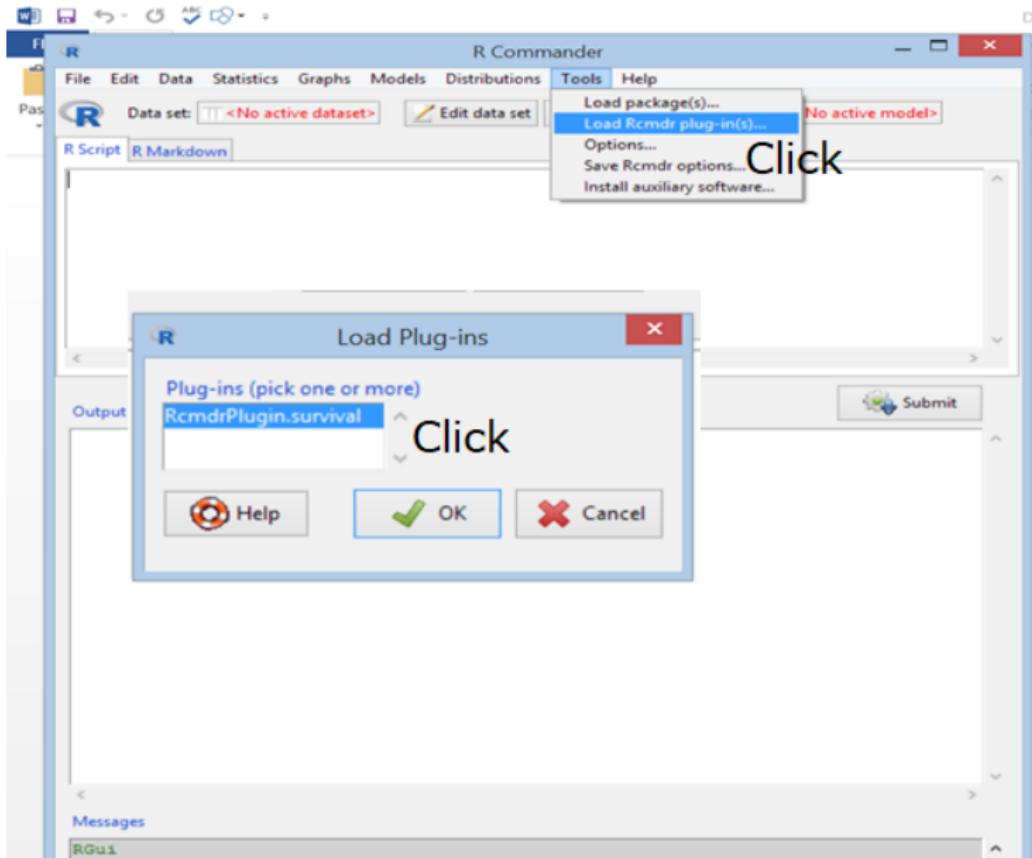
# Fit Models



# Extending the R Commander

- We can extend the R commander by adding packages
- RcmdrPlugin.survival package is for adding Survival Analysis
- RcmdrPlugin.survival package can either be loaded directly, by the command
  - `library("RcmdrPlugin.survival")`, or with the Rcmdr running, via Tools
- RcmdrPlugin.survival package adds a variety of menus and menu items to the R Commander interface:

# Extending the R Commander

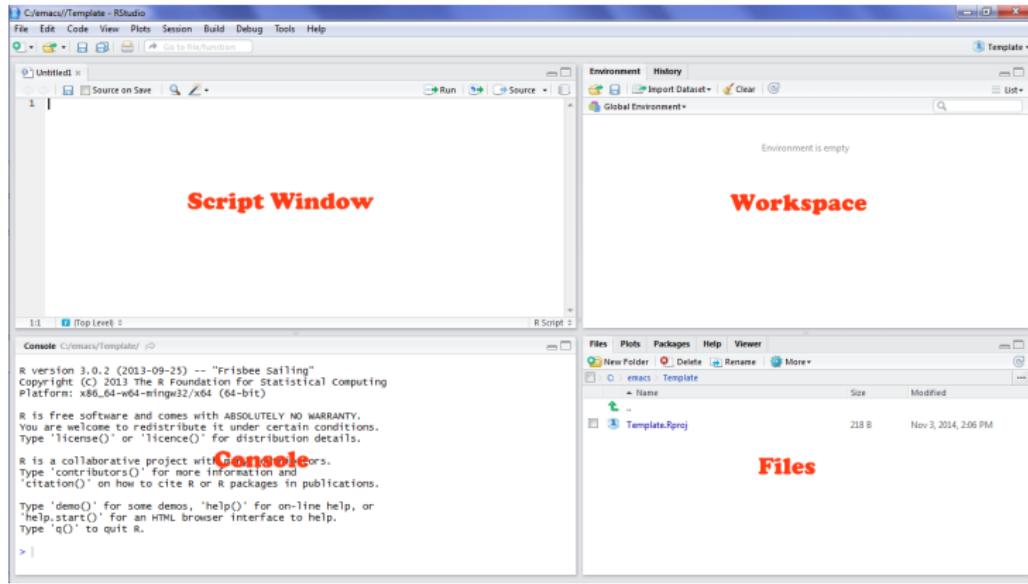


# Rstudio

- RStudio is an active member of the R community
- It makes R easier to use
- In order to install RStudio, first install R
- It can be downloaded here and <http://www.rstudio.com/>
- Once installed, open R Studio then select

# Rstudio layout

- When we open Rstudio, there are 4 windows



## Summary

- R has a wide range of application, from descriptive to LDA and advanced Bayesian models
- Packages are required to perform every analysis
- Recently, R commander package was developed for easy of use for non-statisticians
- Other softwares like Stata and SAS are also widely used to analysis correlated data
- Now a days, Python is used both for data visualization and analysis as well

## Activity 1

- Open an excel file and create the variables ID, age, sex, weight, diastolic blood pressure, blood group, and place of residence (urban versus rural) "UoGdata".
- Create simulated correlated data for 50 sample size and 250 observations
- Make sure the data values are within the possible range
- Imported the dataset "UoGdata" in to R
- Save the data under the name of "UoGdata"

## Part I

# Models for Longitudinal/Cluster Data

## **Chapter 2**

# **Introduction Longitudinal Data Analysis**

# Statistical Models

- Statistics includes the process of finding out about patterns in the real world using data
- When solving statistical problems it is often helpful to make models of real world situations based on observations of data, assumptions about the context, and on theoretical probability
- The model can then be used to make predictions, test assumptions, and solve problems
- These models can be either deterministic or Stochastic
- A deterministic model does not include elements of randomness
- A probabilistic model includes elements of randomness.

# Components of the model

- There are three components in GLM

$$g(y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- Random component

$$y \sim dist()$$

- Systematic Component

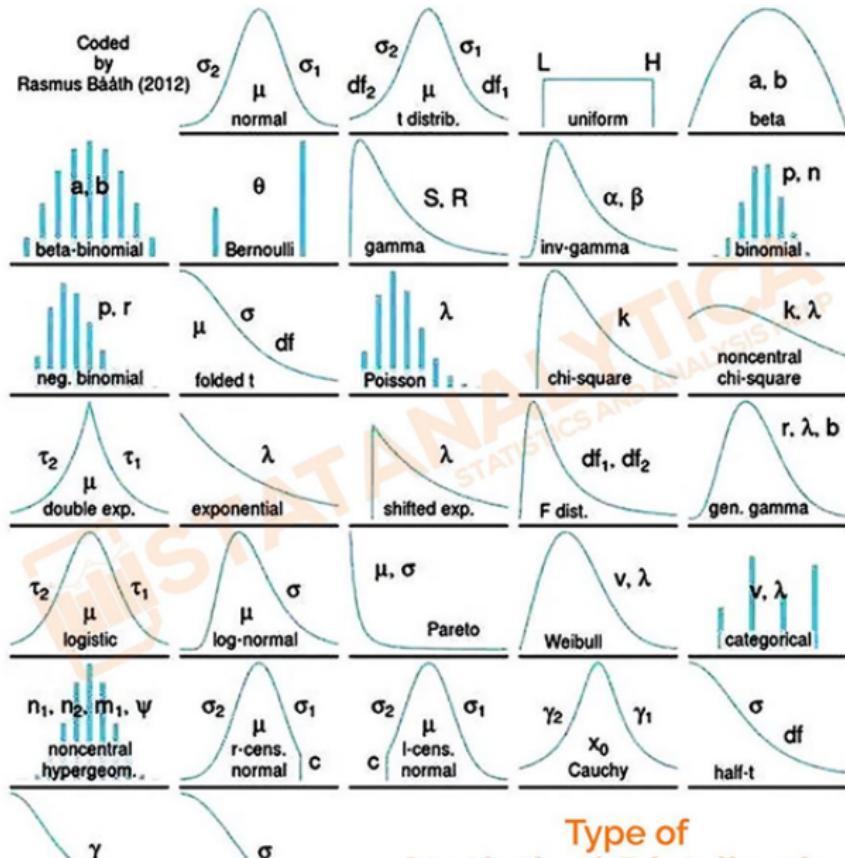
$$\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- Link Function

- The function which links the random component with the systematic component

$$g(y)$$

# Components of the model



# Deterministic Versus Stochastic Models

- Deterministic

- Model parameters are exact

$$E = mc^2$$

- The output of the model is fully determined by the parameter values and the initial conditions
- Stochastic
  - Models possess some inherent randomness

$$g(y) = X\beta + \epsilon_{ij}$$

- The same set of parameter values and initial conditions will lead to an ensemble of different outputs
  - Output has variation
- The natural world is buffeted by stochasticity stochasticity.

## Exercises 2

- Review literatures in your field of study
  - Locate studies and read the method part
  - List the dependent variable
  - List the independent variables
  - What was the statistical methods used?
  - Is the model appropriate for the outcome variable?
- Submission date:

## Correlated Data

- The statistical techniques like analysis of variance and regression have a basic assumption that the residual or error terms are independently and identically distributed
- Many types of studies, however, have designs which imply gathering repeated measurement data that are dependent groups or clusters
- A longitudinal study refers to an investigation where participant outcomes and possibly treatments or exposures are collected at multiple follow-up times

## Repeated Measures or Clustered data

- Clustering arises when data are measured repeatedly on the same unit
- When these repeated measurements are taken over time, it is called a longitudinal or, in some applications, a panel study
- Longitudinal data are special forms of repeated measurements.
- Units could be:
  - Subjects, patients, participants, ...
  - Animals, plants, ...
  - Clusters: families, towns, ...
  - Such repeated measures data are correlated within subjects and thus require special statistical techniques for valid analysis and inference

# Multilevel Data Structure

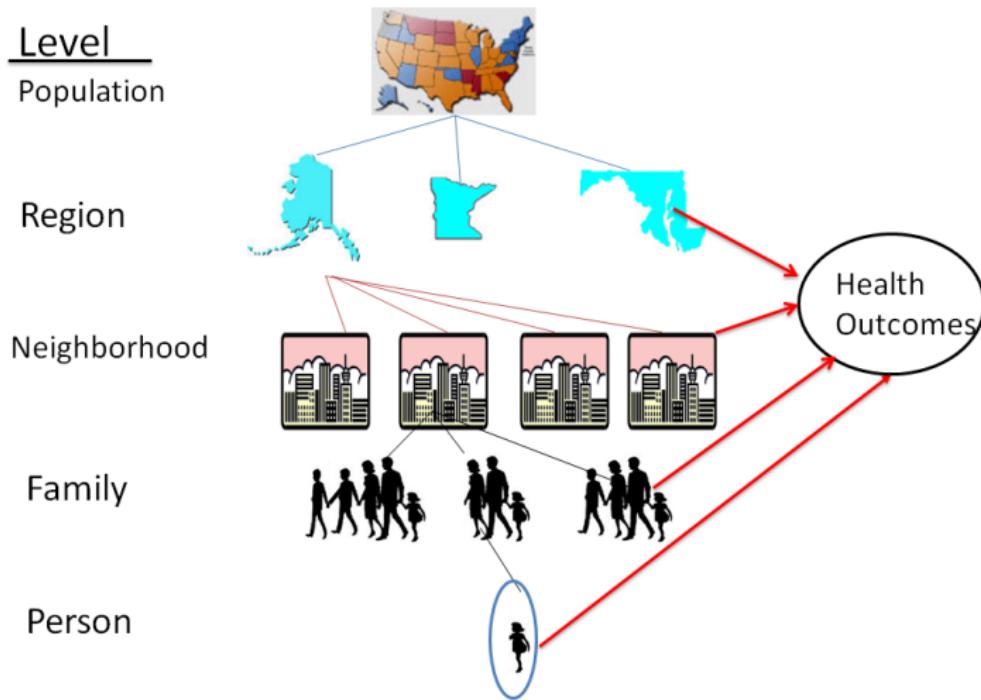


Figure 6: Multilevel Data Structure.

## Types of correlated data

- Repeated Cross-sections: Different samples are taken at each measurement time, to measure trends not individuals experiences
- Examples
  - Ethiopian Demographic and Health Survey (EDHS)
  - Behavioral Risk Factor Surveillance Study (BRFSS)

# Time Series

- Collection of data  $X_t(t = 1, 2, \dots, T)$  with the interval between  $X_t$  and  $X_{t+1}$  being fixed and constant.
- In time-series studies, a single population is assessed with reference to its change over the time
- Here we measure trend, seasonality
- Examples
  - Daily, weekly, or monthly performance of a stock
  - Daily pollution levels in a city
  - Annual measurements of sun spots

## Panel or Multi-level Data

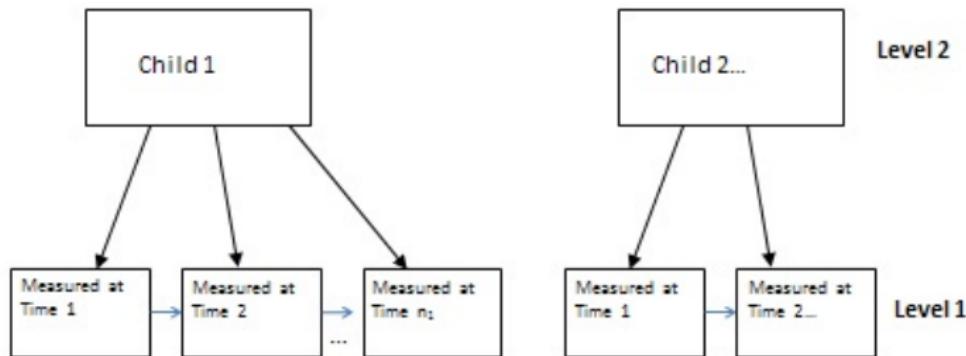
- Same individual/subject/unit is observed over two or more time points.
- 
- Typically large number of observations repeated over a few time points  $i = 1, 2, 3 \dots N$  and  $t = 1, 2, 3 \dots T$
- Examples
  - HIV/AIDS infected individual
  - Patients in chronic care follow-up

## Longitudinal data

- An outcome is measured for the same person repeatedly over a period of time.
- Different subjects may have different numbers of observations which may be taken at different time points.
- Observations made on the same person are likely to be correlated

# Longitudinal data

## Longitudinal Data (Weight Measured Over Time)



Level 1 Variables (Time-Varying): Child weight, Age at each measurement

Level 2 Variables (Time-Invariant): Mother's education, Child's Gender

Figure 7: Two Level longitudinal data.

## Clustered or Hierarchical Data

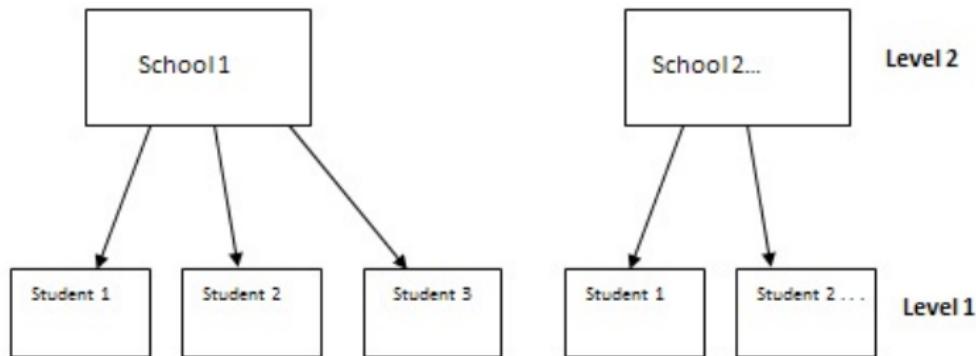
- The observations have a multi-level structure (Same patients ( $i$ ) from facilities ( $k$ ) followed over time ( $t$ ))
  - $k = 1, 2, 3 \dots K$
  - $i = 1, 2, 3 \dots N$
  - $t = 1, 2, 3 \dots T$
- Example
  - Minimum Data Set (MDS) – Quarterly and Annual clinical information on nursing home residents

## Clustered Data

- An outcome is measured once for each subject, and subjects belong to (or are “nested” in) clusters, such as families, schools, or neighborhoods.
- The number of subjects in each cluster may vary from cluster to cluster.
- Outcomes measured for members of these groups are likely to be correlated

# Clustered Data

## Two-Level Clustered Data



Level 1 Variables: Student Achievement Score, Gender, Student's SES....

Level 2 Variables: Public or Catholic School...

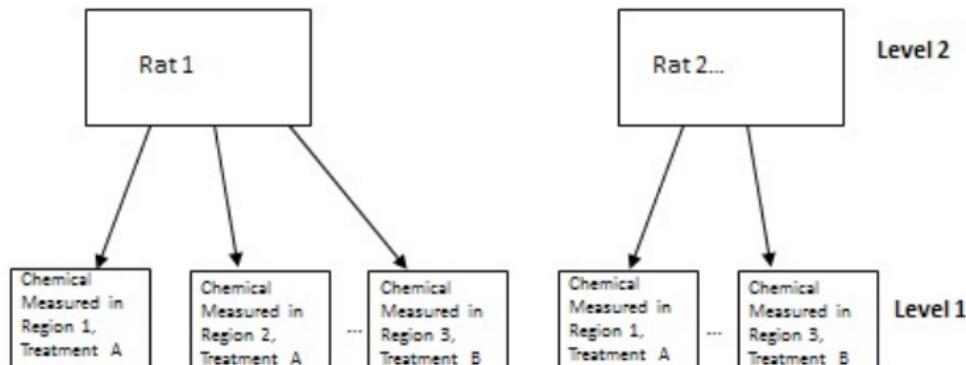
Figure 8: Two Level cluster data.

## Repeated measures data

- Multiple observations are made for the same person over time, space or other dimension.
- Each subject need not have all measurements.
- Outcomes measured for the same person are likely to be correlated.
- Clustered/longitudinal/repeated measures data is more generally known as “multilevel” data.

# Repeated measures data

## Repeated Measures Data (Rat Brain Example)



Level 1 Variables (Varying): Nucleotide bonding measurement, Brain region, Treatment

Level 2 Variables (Invariant): Rat gender

Figure 9: Two Level longitudinal data.

# Types of Responses in Longitudinal Data

- Continuous – Cost of health care
- Discrete – Use or non-use of mental health services
- count – number of outpatient visits
- survival – time from diagnosis to death

## Advantages of modern longitudinal methods

- The methods of analysis used in Longitudinal Data Analysis relax the independence assumption and take into account more complicated data structure.
- You have much more flexibility in research design
  - Not everyone needs the same rigid data collection schedule—cadence can be person specific
  - Not everyone needs the same number of waves—can use all cases, even those with just one wave!
- 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 (those whose values vary over time)
- You can include interactions with time (to test whether a predictor's effect varies over time)

# Challenges in Analyzing Longitudinal Data

- Failure to account for the effect of correlation can result in erroneous estimation of the variability of parameter estimates, and hence in misleading inference.
- Account for dependency of observations
- Both dependent and independent variables change over time—time varying covariates
- Invariable presence of missing data

# Designs of Longitudinal Data

- Equally spaced or balanced panel data
  - When each subject is scheduled to be measured at the same set of times (say,  $t_1, t_2, \dots, t_n$ ), then resulting data is referred as equally-spaced or balanced data
- Unequally spaced or unbalanced data
  - When subjects are each observed at different sets of times there are missing data

# Motivating Datasets

- Five datasets will be considered in this section
- These are:
  - Rat dataset
  - Jimma infant survival data set
  - Orthodontic growth dataset
  - Epileptic dataset
  - Toenail Dataset
  - Gondar HIV/AIDS dataset
  - Gilgel Gibe Mosquito Data

## Rat dataset

- Randomized experiment in which 50 male Wistar rats are randomized to:
  - Control (15 rats)
  - Low dose of Decapeptyl (18 rats)
  - High dose of Decapeptyl (17 rats)
- Research Question: How does craniofacial growth depend on testosterone production?
- Measurements with respect to the roof, base and height of the skull
- Here, we consider only one response, reflecting the height of the skull.

# Rat dataset

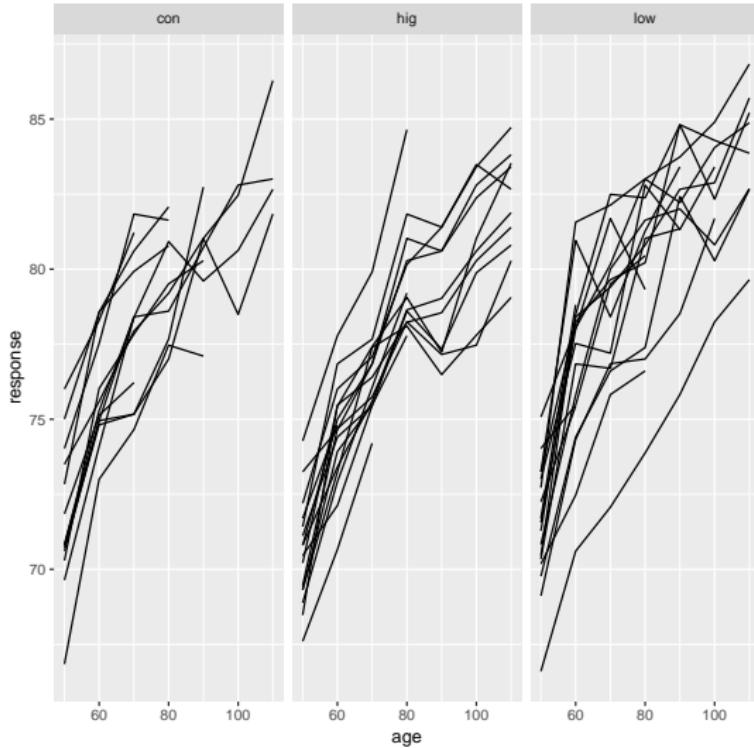


Figure 10: Subject specific profiles of response (height of the skull)

## infant survival data set

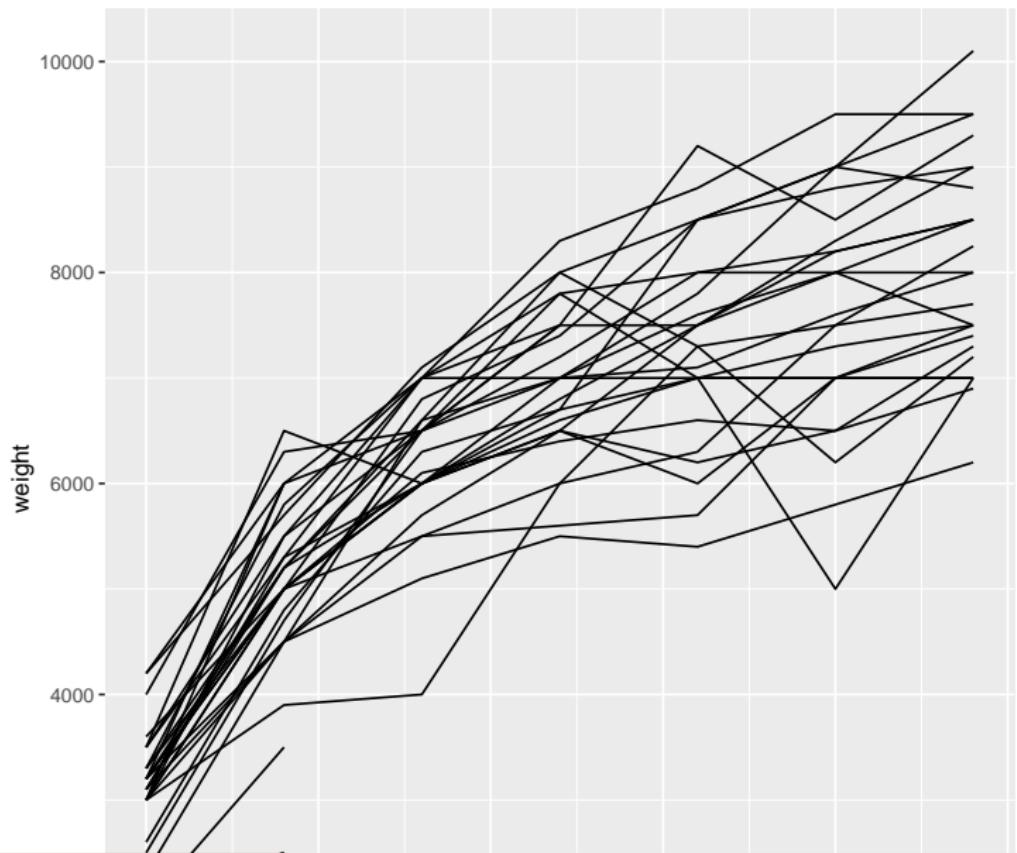
- The data introduced in this section were obtained from a follow-up study of new born infants in Southwest Ethiopia
- Wide ranges of data were collected on the following characteristics:
  - basic demographic information
  - feeding practice
  - anthropometric measurements
  - and others
- Infants were 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 and part of the data can be printed by the following R code

```
# To print the first 16 rows of the data  
head(mydata1, n=16)
```

## Part of the data ...

	ind	sex	place	weight	length	bf	age	numdays	help	BMIBIN
1	1	0	2	3300	49	1	0	0	1	0
2	1	0	2	5000	60	1	2	0	1	0
3	1	0	2	6000	60	1	4	0	1	0
4	1	0	2	6500	66	1	6	0	1	0
5	1	0	2	6200	67	1	8	7	0	0
6	1	0	2	6500	67	1	10	0	1	0
7	1	0	2	7300	70	1	12	0	1	0
8	2	1	2	3000	43	1	0	0	1	1
9	2	1	2	6000	64	1	2	0	1	0
10	3	1	2	4200	53	1	0	0	1	1
11	3	1	2	5700	61	1	2	0	1	0
12	3	1	2	7100	65	1	4	0	1	0
13	3	1	2	8000	70	1	6	7	0	0
14	3	1	2	7300	70	1	8	0	1	0
15	3	1	2	6200	70	1	10	5	1	0
.	.	.	.	.	.	.	.	.	.	.

# Jimma infant survival data set



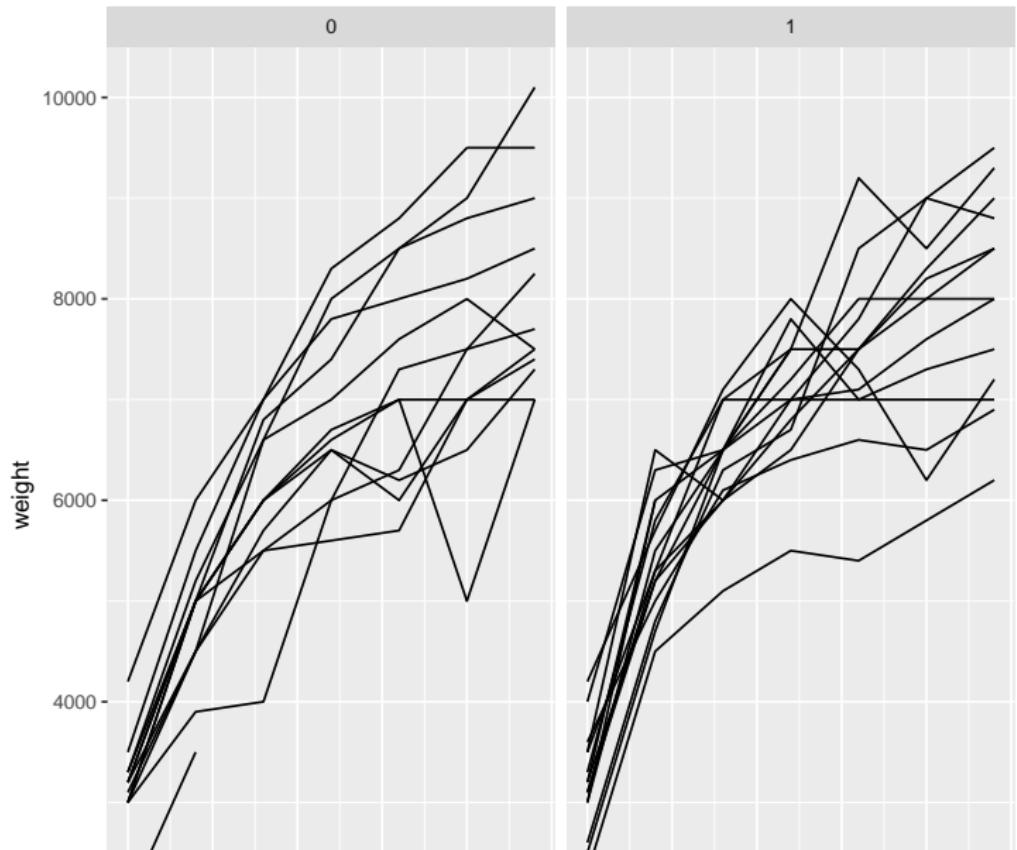
The profile plot is produced by using the following R code.

## R CODE:

```
library(foreign)
View(mysample)
mydata<- read.dta(file="C:/Users/lucp8319/Desktop/Biostat_2018/
LDA/LDA_2018/1dataset/infant.dta")
mydata1 <- mydata[which(mydata$ind<31),] # data only <31 days of follow up

library(ggplot2)
library(methods)
library(labeling)
p <- ggplot(data = mydata1, aes(x = age, y = weight, group = ind))
# simple scatter plot
p + geom_point()
# simple spaghetti plot
p + geom_line()
```

# Jimma infant survival data set



The profile plot by sex groups using the following R code.

### R CODE:

```
## facet (condition) the graph base on the male variable  
library(reshape2)  
p + geom_line() + facet_grid(. ~ sex)
```

- There is an increase in weight overtime for both males and females
- On the average, males appear to have higher mean profile than females
- It is not yet possible to decide on the significance of this difference
- From individual profiles, the variability seems almost the same among the two groups.

## Orthodontic growth dataset

- This data set is taken from Potthoff and Roy, Biometrika (1964)
- A set of measurements of the distance from the pituitary gland to the pterygo-maxillary fissure taken every two years from 8 years of age until 14 years of age on a sample of 27 children, 16 males and 11 females
- The data, were collected by orthodontists from x-rays of the children's skulls
- The individual profile plot and the mean profile for males and females separately are given below
- Research question: **Is dental growth related to gender?**
- Variables in the data include: (1) obs: observation number (2) group: group according to height of the mother (1: small; 2: medium; 3: tall) (3) child: subject identification number (4) age: age at which the observation is taken (years) (5) height: the response measured (cm)

# Orthodontic growth dataset

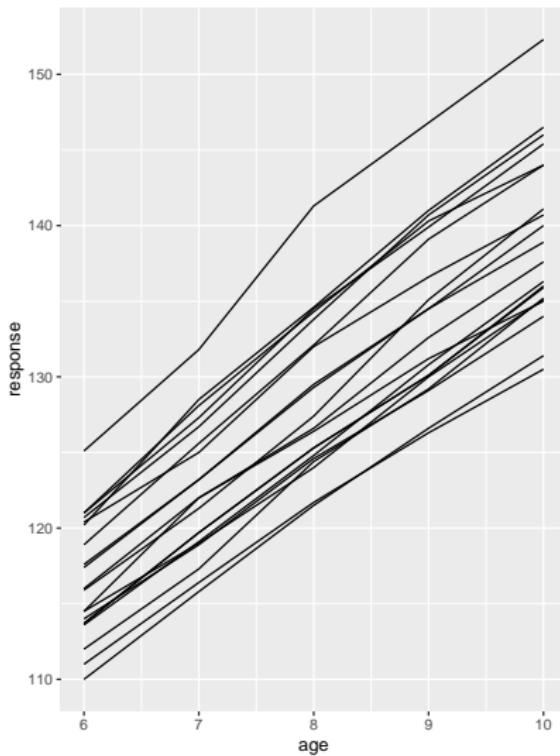


Figure 13: Subject specific profiles of growth

The profile plot by sex groups using the following R code.

## R CODE:

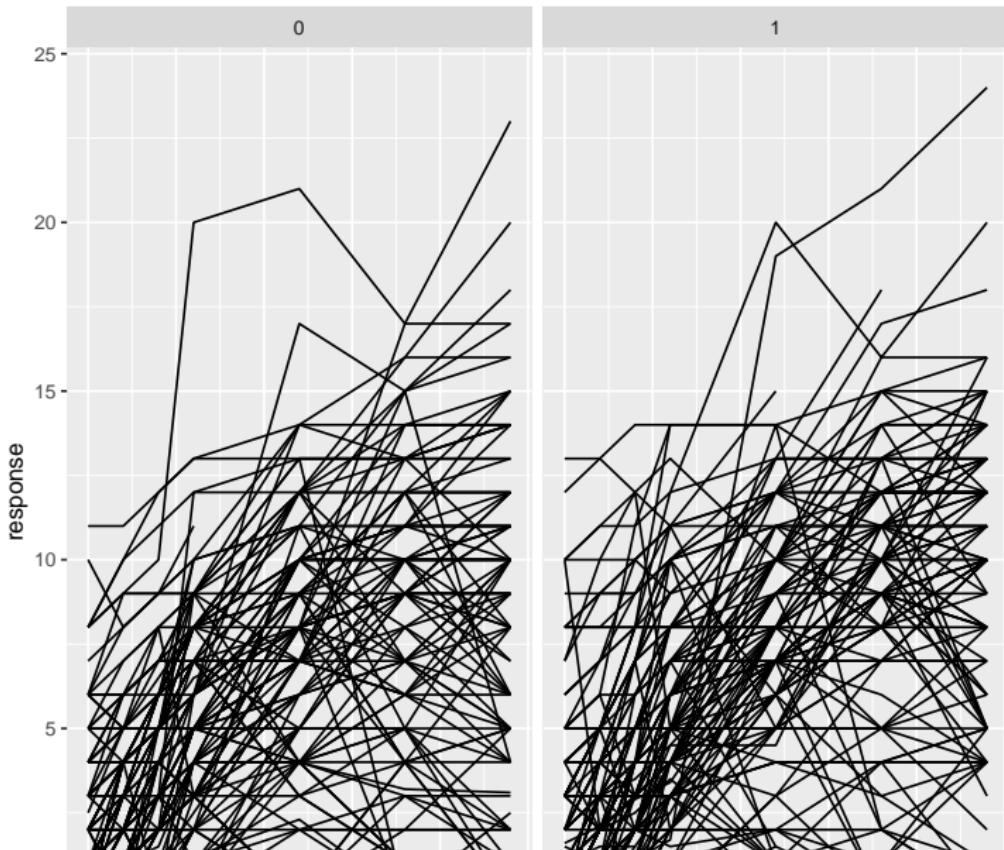
```
##### The Orthodontic Growth DataSet #####
mydata2<- read.csv(file="C:/Users/lucp8319/Desktop/Biostat_2018/
LDA/LDA_2018/1dataset/growth.csv")
library(ggplot2)
library(methods)
library(labeling)
library(reshape2)
d <- ggplot(data = mydata2, aes(x = age, y = response, group = child))
# simple spaghetti plot
d + geom_line()
## facet (condition) the graph base on the male variable
d + geom_line() + facet_grid(. ~ group)
```

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

## Toenail Dataset

- Toenail Dermatophyte Onychomycosis: Common toenail infection, difficult to treat, affecting more than 2% of 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 for the comparison of two such new compounds (A and B) for oral treatment.
- Research question: **Are both treatments equally effective for the treatment of TDO?**
  - $2 \times 189$  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.
  - Response considered here: Unaffected nail length (mm):

# Toenail Dataset



The profile plot by sex groups using the following R code.

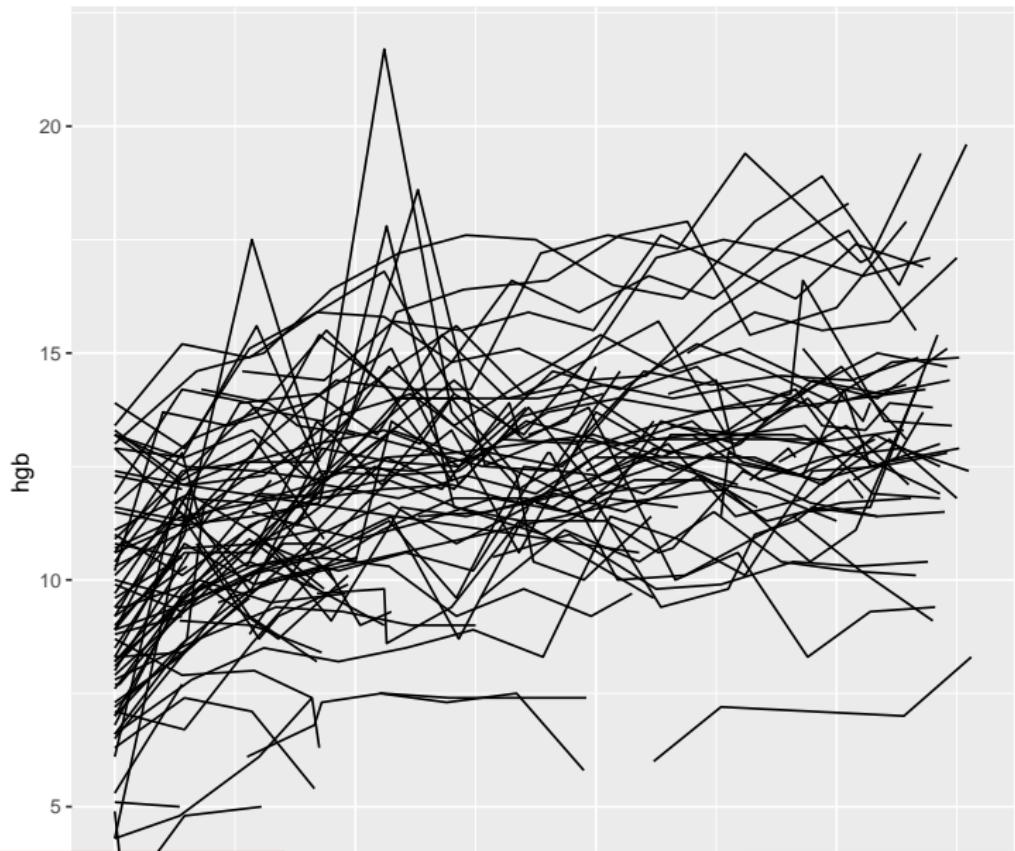
## R CODE:

```
##### Toenail Dataset #####
mydata3 <- read.csv("C:/Users/lucp8319/Desktop/Biostat_2018/LDA/
LDA_2018/1dataset/Toenail_contineous.csv")
d <- ggplot(data = mydata3, aes(x = time, y = response, group = id))
# simple spaghetti plot
d + geom_line()
## facet (condition) the graph base on the male variable
library(reshape2)
d + geom_line() + facet_grid(. ~ treat)
```

# Gondar VL/HIV dataset

- Follow-up data among VL/HIV in Gondar Hospital
- The variables time, age, sex, residence, wbc and hgb were included

# Gondar VL/HIV dataset



The profile plot by sex groups using the following R code.

## R CODE:

```
##### Toenail Dataset #####
mydata4 <- read.csv("C:/Users/lucp8319/Desktop/Biostat_2018/
LDA/LDA_2018/1dataset/VL_HIV.csv")
d <- ggplot(data=mydata4, aes(x = Day, y = hgb, group = id))
# simple spaghetti plot
d + geom_line()
## facet (condition) the graph base on the male variable
library(reshape2)
d + geom_line() + facet_grid(. ~ treat)
```

## Epileptic dataset

- The epileptic data set considered here is obtained from a randomized, multi-center study
- Comparison of placebo with a new anti-epileptic drug (AED)
- In the study, 45 patients were randomized to the placebo group and 44 to the active (new) treatment group
- The number of epileptic seizures were measured on a weekly basis during a 16 weeks period
- After this period, patients were entered into a long-term study up to 27 weeks
- The key research question is whether or not the additional new treatment reduces the number of epileptic seizures

# The Gilgel-Gibe Mosquito Data

- A study conducted around Gilgel-Gibe dam for three years.
- Influence of the dam on mosquito abundance and species composition.
- Eight 'At risk' and eight 'Control' villages based on distance.
- One collection approach: IRC.
- Mosquito species were identified and counted.
- *An. gambiae* was found to be the dominant one (more than 95

# Requirements for Longitudinal Models

- Capture trend over time while taking account of the correlation that exists between successive measurements
- Describe the variation in the baseline measurement and in the rate of change over time
- Explain the variations in baseline measurement and trends by relevant covariates
- A minimum of 4 time points is recommended; With less than 4 time points, it is not possible to identify enough parameters in the growth model to make the model flexible
  - 4 time points give more power
  - With 3 time points restrictions need to be placed on the growth models
  - If only 2, compute change scores, use simple methods

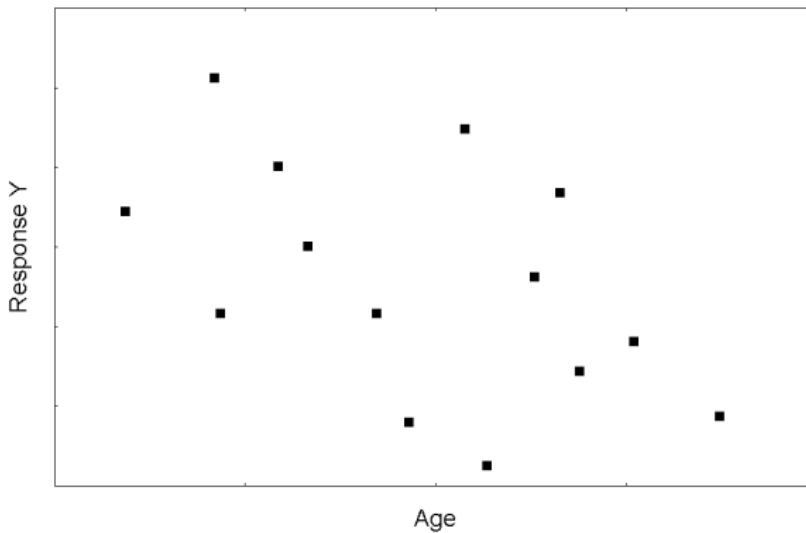
## Cross-sectional versus Longitudinal Data

- Cross sectional data refers to the data collected at a specific point of time.
- Longitudinal data refers to measurements made repeatedly over time to study how the subjects evolve over time
- Observations from cross sectional data are uncorrelated
- In longitudinal study, the measurements made for subjects over a period of time are correlated.

## Cross-sectional versus Longitudinal Data

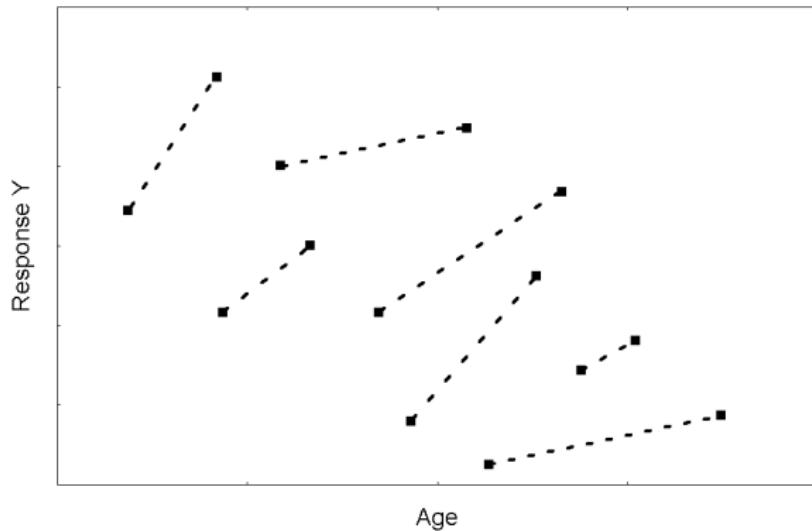
- Recall: Longitudinal data refers to measurements made repeatedly over time to study how the subjects evolve over time
- And, the repeated measures taken from a subjects tend to correlate with each other
- Cross sectional data refers to the data collected at a specific point of time
- Observations from cross sectional data are uncorrelated

- Suppose it is of interest to study the relation between some response  $Y$  and age
- A cross-sectional study yields the following data:



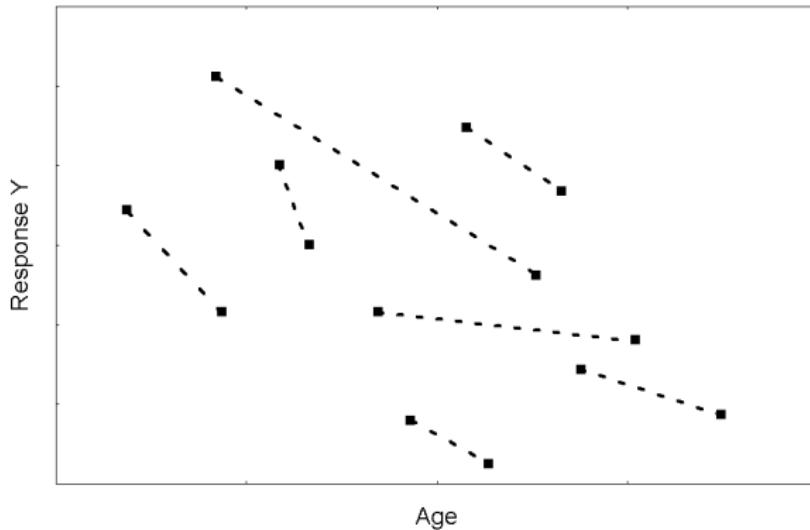
- The graph suggests a negative relation between  $Y$  and age
- Exactly the same observations could also have been obtained in a longitudinal study, with 2 measurements per subject as shown below:

First case:



Are we still inclined to conclude that  $Y$  and Age are negatively related?  
The graph suggests a negative cross-sectional association but a positive longitudinal trend.

## Second case:



The graph now suggests cross-sectional as well as longitudinal trend to be negative.

## Correlation Matrix of Growth Data:

$$\begin{bmatrix} 1.00 & 0.63 & 0.71 & 0.60 \\ 0.63 & 1.00 & 0.63 & 0.76 \\ 0.71 & 0.63 & 1.00 & 0.80 \\ 0.60 & 0.76 & 0.80 & 1.00 \end{bmatrix}$$

- This correlation can not be ignored in the analysis!

- A correct analysis should account for this correlation.
- This is why the classical methods such as ANOVA, linear regression, ... fail for such data
- Usually correlation decreases as the time span between measurements increases
- The simplest case of longitudinal data are paired data
- The paired t-test accounts for this by considering subject-specific differences

# **Chapter 3**

## **Simple Methods**

## Simple Methods

- The reason why classical techniques fail in the context of longitudinal data is that observations within subjects are correlated
- In many cases the correlation between two repeated measurements decreases as the time span between two repeated measurements increases
- Paired t-test and difference in difference (DID) can be used if the repeated measurements are two
- This reduces the number of measurements to just one per subject, which implies that classical techniques can be applied
- In the case of more than 2 measurements per subject, similar simple techniques are often applied to reduce the number of measurements for the  $i^{th}$  subject, from  $n_i$  to 1

# Simple Methods

- Some Examples of Simple Methods
  - Analysis at each time point separately
  - Analysis of Area Under the Curve (AUC)
  - Analysis of endpoints
  - Analysis of increments
  - Analysis of covariance
  - The above methods have limitations such as:

## Analysis at Each Time Point

- The data are analysed at each occasion separately.
- Advantages:
  - Simple to interpret
  - Uses all available data
- Disadvantages:
  - Does not consider 'overall' differences
  - Does not allow to study evolution differences
  - Problem of multiple testing

# Analysis of Area Under the Curve

- For each subject, the area under its curve is calculated

$$AUC_i = (t_{i2} - t_{i1})(y_{i1} + y_{i2})/2 + (t_{i3} - t_{i2})((y_{i2} + y_{i3})/2 + \dots \quad (1)$$

- Advantages

- No problems of multiple testing
- Does not explicitly assume balanced data
- Compare 'Overall' differences

- Disadvantages:

- Uses only partial information :  $AUC_i$

## Analysis of Endpoints

- In randomized studies, there are no systematic differences at baseline.
- Hence, 'treatment' effects can be assessed by only comparing the measurements at the last occasion
- Advantages :
  - ① No problem of multiple testing
  - ② Does not explicitly assume balanced data
- Disadvantages
  - ① Uses only partial information :  $y_{in_i}$
  - ② Only valid for large data sets

# Analysis of Increments

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

$$y_{in_i} - y_{i1} \quad (2)$$

- Advantages
  - ① No problem of multiple testing
  - ② Does not explicitly assume balanced data
- Disadvantages: Uses only partial information :  $y_{in_i} - y_{i1}$ .

# Analysis of Covariance

- Another way to analyze 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 :
  - ① No problems of multiple testing
  - ② Does not explicitly assume balanced data
- Disadvantages :
  - ① Uses only partial information :  $y_{i1}$ . and  $y_{in_i}$
  - ② Does not take into account the variability of  $y_{i1}$

## Summary

- The AUC, endpoints and increments are examples of summary statistics
- Such summary statistics summarize the vector of repeated measurements for each subject separately.
- This leads to the following general procedure:
- **Step 1** : Summarize data of each subject into one statistic , a summary statistic
- **Step2** : 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.
- However, all these methods have the disadvantage that lot of information is lost.
- Further, they often do not allow to draw conclusions about the way the end points have been reached.

# **Chapter 4**

## **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
- Introduction

# Exploratory Data Analysis

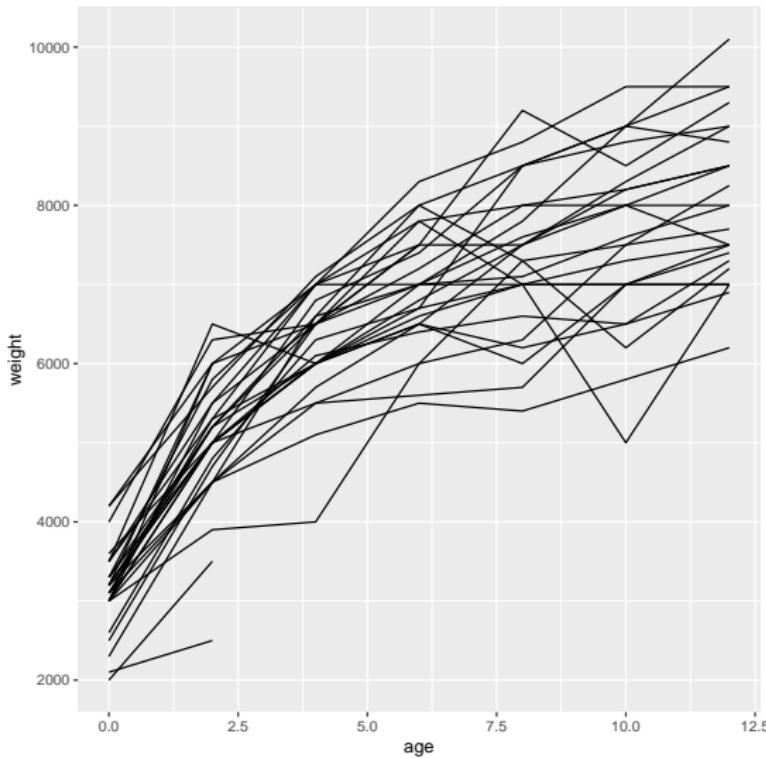
- Longitudinal data analysis, like other statistical methods, has two components which operate side by side:
  - exploratory and
  - confirmatory analysis.
- Exploratory analysis comprises techniques to visualize patterns in the data
- Confirmatory analysis is judicial work, weighing evidence in data 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 new born 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



## Conclusions From the profile

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

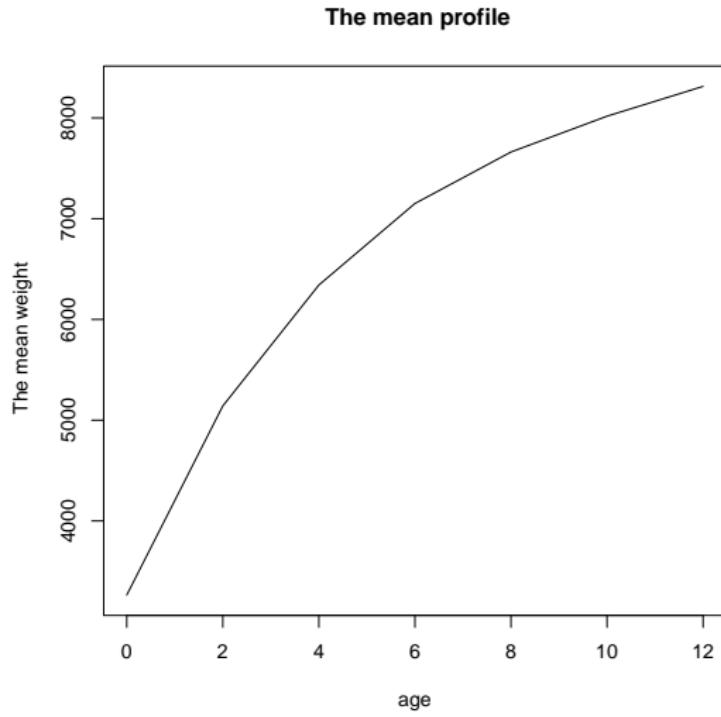
## Mean Profile

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

```
## mean profile  
attach(mydata)  
mean1<-tapply(weight, age, mean)  
age1<-as.numeric(unique(age))  
plot(age1, mean1, type= "l", xlab="age", ylab=" The mean weight",  
lwd=1, main=" The mean profile")
```

# Mean Profile

- The mean profiles

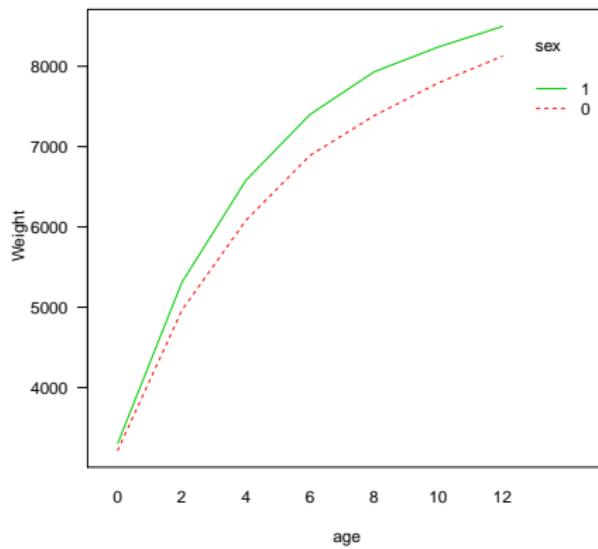


## Mean Profile by Sex

- The mean profiles by sex

```
## mean profile
```

```
interaction.plot(age,sex,weight,fun=mean, col=2:14,  
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 `ImList` 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;

```
> fit<-lmList(weight~age|ind, mydata)
```

Call:

Model: weight ~ age | ind

Data: mydata

Coefficients:

	(Intercept)	age
1	4200.000	2.714286e+02
2	3000.000	1.500000e+03
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

## Choosing the random effect

- 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 *lmeList* are of class *lmeList*, 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  $1-\alpha/2$  will be marked in the plot, we use the R function
- We see that subject 29 has high slope

## Main lmList methods

- The print method displays minimal information about the fitted object

```
augPred # predictions augmented with observed values  
coef # coefficients from individual lm fits  
fitted # fitted values from individual lm fits  
fixef # average of individual lm coefficients  
intervals # confidence intervals on coefficients  
lme # linear mixed-effects model from lmList fit  
logLik # sum of individual lm log-likelihoods  
pairs # scatter-plot matrix of coefficients or random effects  
plot # diagnostic Trellis plots  
predict # predictions for individual lm fits  
print # brief information about the lm fits  
qqnorm # normal probability plots  
ranef # deviations of coefficients from average  
resid # residuals from individual lm fits  
summary # more detailed information about lm fits
```

## R Code

- R code to exploratory the random effect

```
fm.lis<-lmList(response ~ I(age-11)|child, data=mydata2)
summary(fm.lis)
pairs(fm.lis, id = 0.01, adj = -0.5) # scatter plot

intervals(fm.lis) # Confidence intervals

plot(intervals(fm.lis))
```

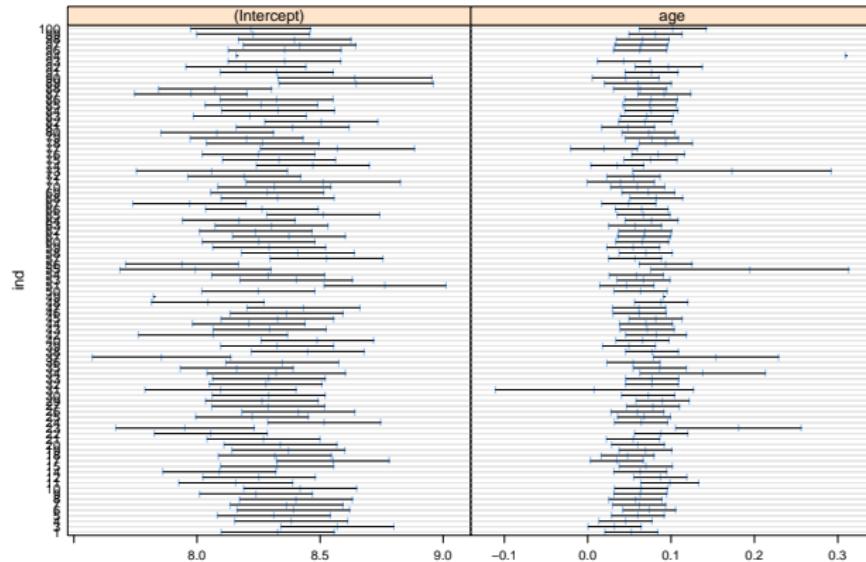
## Exploratory random effect

```
> intervals(fit0.lm)
, , (Intercept)
lower      est.      upper
1  8.105020 8.328701 8.552381
3  8.347202 8.570882 8.794563
4  8.159632 8.383312 8.606992
5  8.089277 8.312957 8.536638
6  8.168702 8.392382 8.616063
7  8.141055 8.364736 8.588416
8  8.179596 8.403276 8.626957
9  8.016647 8.240327 8.464008
10 8.196293 8.419973 8.643654
```

- As often happens displaying the intervals as a table of numbers is not very informative
- It is much more effective to plot these intervals

## Pairs Plot

- These plots are displayed for a subset of the data using interval plots
- 95% CI on intercept and slope for each subject in the infant weight data



# Exploring the Correlation Structure

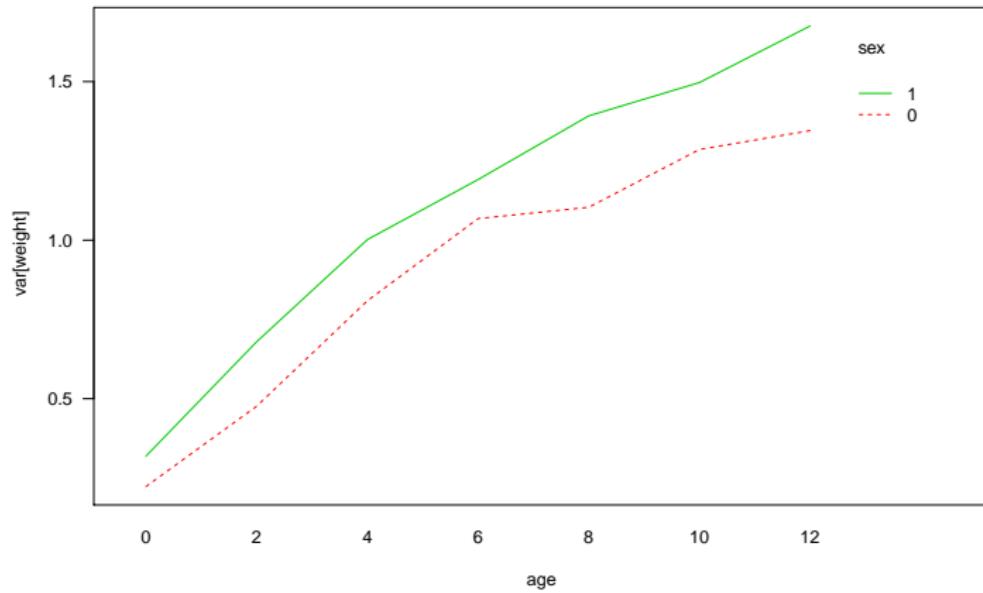
- In longitudinal data analysis we model two key components of the data:
  - Mean structure
  - Correlation structure (after removing the mean structure)
- Modelling the correlation is important to be able to obtain correct inferences on regression coefficients
- Correlation can be formulated in terms of subject-specific models and/or transition models
- Three basic elements of correlation structure:
  - 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

## Jimma Infant Survival Data Set

- Having appropriate model studying the evolution of the variance is very important step of the modeling approach
- The observed variance shows an increase in variability overtime
- 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

## Observed Variance

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

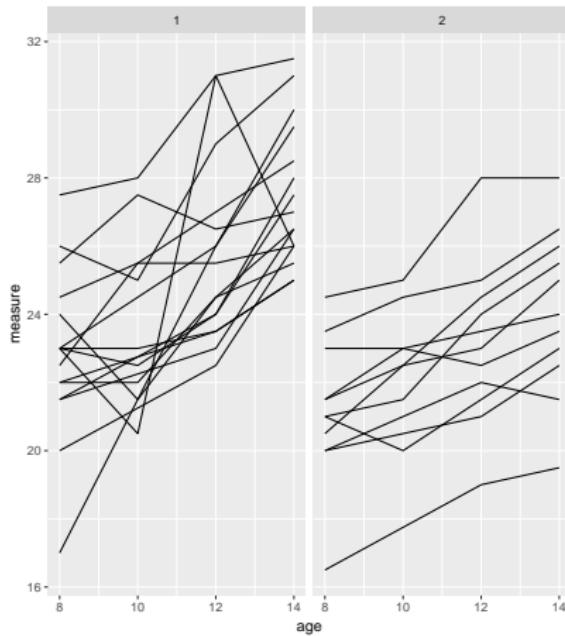


## Growth Data

- 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?**

# Growth Data

- The individual profiles support a random-intercepts model

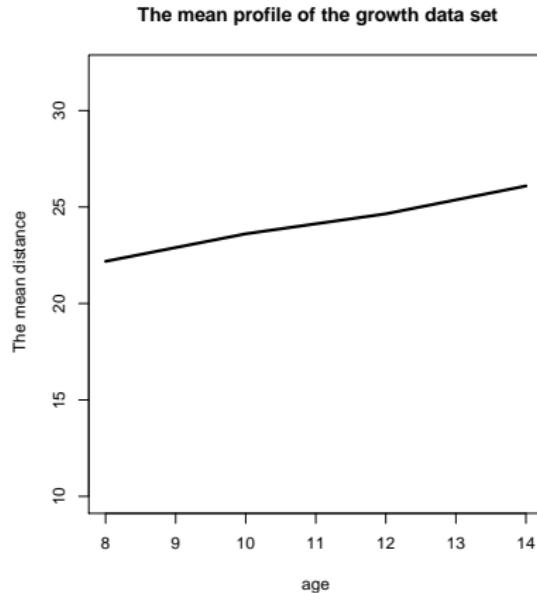


## Growth Data

- 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



## R code

- R code for average profile:

```
mean1<-tapply(response, age, mean)
age1<-as.numeric(unique(age))
plot(age1, mean1, type= "l", ylim=c(100,150), xlab="age",
" The mean distance", lwd=3, main=" The mean profile of the data")
```

- For Balanced longitudinal data, the correlation structure can be studied through the correlation matrix, or a scatter plot matrix
- The correlation matrix for Orthodontic growth data:

```
1.0000000 0.9809247 0.9815512 0.9740140 0.9540426  
0.9809247 1.0000000 0.9858051 0.9803015 0.9584083  
0.9815512 0.9858051 1.0000000 0.9880770 0.9751260  
0.9740140 0.9803015 0.9880770 1.0000000 0.9920014  
0.9540426 0.9584083 0.9751260 0.9920014 1.0000000
```

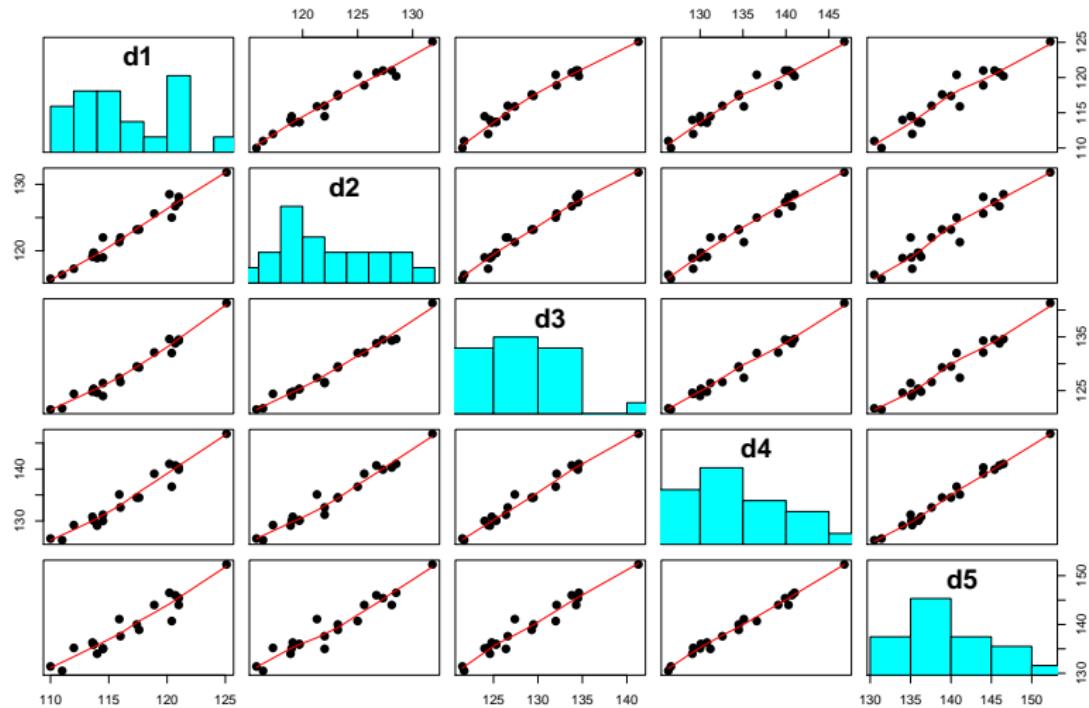
- Data exploration is therefore extremely helpful as additional tool in the selection of appropriate models

# R code

- R code for the correlation matrix:

```
##### R-code for Correlation matrix ##
d1<-response[age==6]
d2<-response[age==7]
d3<-response[age==8]
d4<-response[age==9]
d5<-response[age==10]
response1<-cbind(d1, d2, d3, d4, d5)
cor(response1)
```

# Scatter plot matrix for growth data



## R code for 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 show a considerable within and between subject variability
- This can also be augmented from the variance covariance matrix of the observed data indicated below

# Covariance Matrix

- Covariance Matrix for Growth Data:

$$\begin{bmatrix} 15.81095 & 17.32547 & 20.33874 & 21.64158 & 21.46295 \\ 17.32547 & 19.73063 & 22.81884 & 24.33184 & 24.08595 \\ 20.33874 & 22.81884 & 27.15589 & 28.77184 & 28.74984 \\ 21.64158 & 24.33184 & 28.77184 & 31.22408 & 31.36171 \\ 21.46295 & 24.08595 & 28.74984 & 31.36171 & 32.00997 \end{bmatrix}$$

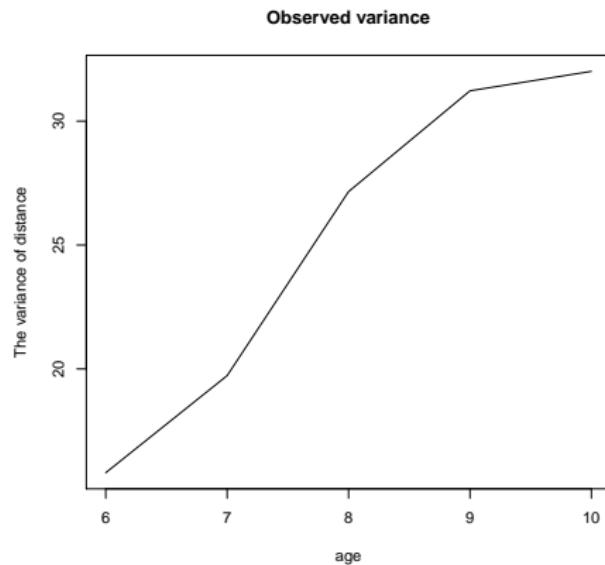
- R code

```
#variance-covariance matrix  
covmatrix = matrix(c(cov(response1)), nrow=5, ncol=5)
```

# Overall variability

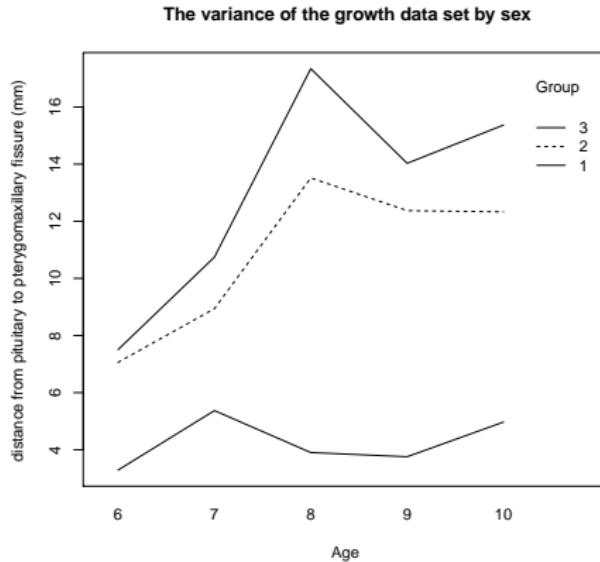
- R code

```
plot(age1, varg, type="l", main = " Observed variance ",  
xlab="age", ylab="The variance of distance", lwd=1)
```



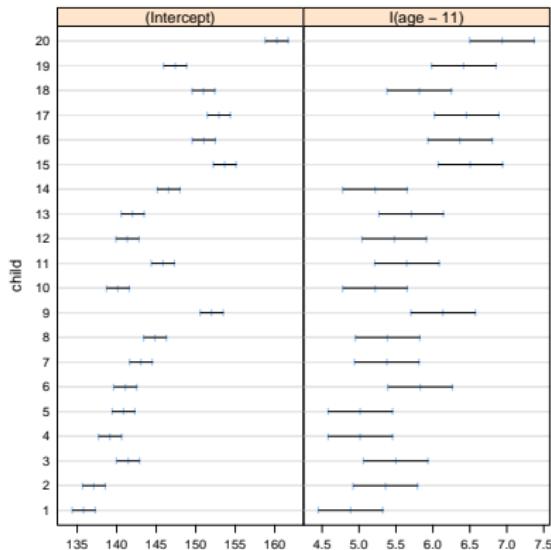
## Variability by group

```
interaction.plot(age, group, response, 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")
```



## Pairs Plot

- Ninety-five percent confidence intervals on intercept and slope for each subject in the orthodontic distance growth data



- The individual CI give a clear indication that a random effect is needed to account for subject-to-subject variability in the intercept

## Exercises 3

- Consider the data simulated above
  - Check the individual profile
  - Check the mean profile
  - Define the mean structure

# **Chapter 5**

## **A Model for Longitudinal Data**

# Introduction

- 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
- LMM can be considered as a further generalization of GLM to better support 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

# Mixed Effect Models

- A mixed effects model for longitudinal or clustered data can be obtained from the corresponding model for cross-sectional data by introducing random effects. Specifically, we have
  - Linear mixed effects (LME) models, which can be obtained from linear regression models by introducing random effects;
  - Generalized linear mixed models (GLMMs), which can be obtained from GLMs by introducing random effects;
  - Nonlinear mixed effects (NLME) models, which can be obtained from nonlinear regression models by introducing random effects;
  - Frailty models, which can be obtained from survival models by introducing random effects.

## Mixed Effect Models

- For mixed effects models, the random effects in the models represent the influence of each individual (cluster) on the repeated observations that is not captured by the observed covariates
- A mixed effects model can be named as multi-level or hierarchical model
  - In longitudinal studies repeated observations from a subject are nested within this subject
  - In multi-center studies observations from a center are nested within this center
- Random effects are used to accommodate the heterogeneity in the data, which may arise from subject or clustering effects or from spatial correlation

## Random Effect Model

- For each unit, baseline value is the result of a random deviation from some mean intercept.
- The intercept is drawn from some distribution for each unit, and it is independent of the error for a particular observation; we just need to estimate parameters describing the distribution from which each unit's intercept is drawn
- Facilities – could be considered as random if they are random sample from a population

## Hierarchical effects

- Hierarchical designs have nested effects.
- Nested effects are those with subjects within groups
- For instance, patients are nested within doctors and doctors are 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)

- In practice: often unbalanced data:
  - unequal number of measurements per subject
  - measurements not taken at fixed time points
- Therefore, multivariate regression techniques are often not applicable
- Often, subject-specific longitudinal profiles can be well approximated by linear regression functions
- This leads to a 2-stage model formulation: i.e.
  - ① Stage 1: Linear regression model for each subject separately
  - ② Stage 2: Explain variability in the subject-specific regression coefficients using known covariates

## Stage 1

- Response  $Y_{ij}$  for  $i^{th}$  subject, measured at time  $t_{ij}$ ,  $i = 1, \dots, N$ ,  
 $j = 1, \dots, n_i$
- Response vector  $Y_i$  for  $i^{th}$
- subject:  $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$
- Stage 1 model:  $Y_i = Z_i\beta_i + \varepsilon_i$
- $Z_i$  is a  $(n_i \times q)$  matrix of known covariates
- $\beta_i$  is a  $q$  dimensional vector of subject-specific regression coefficients
- $\varepsilon_i \sim N(0, \Sigma_i)$
- Often  $\Sigma_i = \sigma^2 I_{n_i}$
- Note that the above model describes the observed variability within subjects

## Stage 2

- Between-subject variability can now be studied from relating the  $\beta_i$  to known covariates

Stage 2 model:

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

- $K_i$  is a  $(q \times p)$  matrix of known covariates
- $\beta$  is a  $p$ -dimensional vector of unknown regression 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  $\widehat{\beta}_i$
- Summary statistics  $\widehat{\beta}_i$  analyzed in second stage
- The associated drawbacks can be avoided by combining the two stages into one model:

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

- General linear mixed-effects model

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \varepsilon_i \\ b_i \sim N(0, D), \quad \varepsilon_i \sim N(0, \Sigma_i), \\ b_1, \dots, b_N, \varepsilon_1, \dots, \varepsilon_N \text{ independent} \end{cases}$$

Terminology:

- Fixed effects:  $\beta$
- Random effects:  $b_i$
- Variance components: elements in  $D$  and  $\Sigma_i$

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,

## Hierarchical versus Marginal Model

- The general linear mixed model is given by:

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \varepsilon_i \\ b_i \sim N(0, D), \quad \varepsilon_i \sim N(0, \Sigma_i), \\ b_1, \dots, b_N, \varepsilon_1, \dots, \varepsilon_N \text{ independent} \end{cases}$$

It can be rewritten as:

$$Y_i | b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i), \quad b_i \sim N(0, D)$$

- It is therefore also called a hierarchical model:
  - A model for  $Y_i$  given  $b_i$
  - A model for  $b_i$
- Marginally, we have that  $Y_i$  is distributed as:

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

- 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 :  $V_i = Z_i D Z_i' + \Sigma_i$
- Note that the hierarchical model implies the marginal one, **NOT** vice versa.

# Components of the Linear Mixed Effects Model

## The Mean Structure

- The  $X_i\beta$  part stands for fixed effects. where  $X_i$  is a set of covariates used for modelling the response.

$$Y_i = X_i\beta + Z_i\mathbf{b}_i + \varepsilon_i$$

# The Random Effects

- In many practical applications, we wish to restrict  $D$  to special forms of variance-covariance matrices that are parametrized by fewer parameters.
- For example we may be willing to assume that the random effects are independent,
  - $D$  would be diagonal, or that, in addition to being independent, they have the same variance, in which case  $D$  would be a multiple of the identity matrix.

## Standard **pdMat** Classes

```
pdBlocked      # block-diagonal  
pdCompSymm    # compound-Symmetry structure  
pdDiag        # diagonal  
pdIdent       # multiple of an identity  
pdSymm        # general positive definite matrix
```

# Variance Structure

- The **nlme** library provides a set of classes of variance functions, the **varFunc** classes, that are used to specify within-group in the mixed effects model
- Standard **varFunc** classes

```
varFixed           #fixed variance  
varIdent          #different variances per stratum  
varPower          #power of covariate  
varExp            #exponential of covariate  
varConstPower     #constant plus power of covariate  
varComb            #combination of variance functions
```

# The Correlation Structures

- Correlation structures are used to model dependence among observations
- In the context of mixed-effects models, they are used to model dependence among the within-group errors

- **Serial Correlation Structures**

- Are used to model dependencies in data observed sequentially overtime and indexed by a one dimensional time vector
- Some of the most common serial correlation structures used in practice includes:Compound Symmetry, General and Autoregressive-Moving Average.

- **Compound Symmetry**

- This is the simplest serial correlation structure, which assumes equal correlation among all within-group errors pertaining to the same group

$$\text{cor}(\varepsilon_{ij}, \varepsilon_{ik}) = \rho, \forall j \neq k, \quad (3)$$

- where the single correlation parameter  $\rho$  is referred to as the **intraclass** correlation coefficient.

## General

- This structure represents the other extreme in complexity to the compound symmetry structure
- Each correlation in the data is represented by a different parameter, corresponding to the correlation function

$$h(k, p) = \rho_k, k = 1, 2, \dots \quad (4)$$

- As the number of parameters increases quadratically with maximum number of observations per group, the general correlation structure is useful as an exploratory tool to determine a more parsimonious correlation model.

## Autoregressive-Moving Average

- These family of models assume that the data are observed at integer time points
- So lag-1 refers to observations one time unit apart and so on.
- Autoregressive models express the current observation as a linear function of previous observations plus a homoscedastic noise term,  $a_t$ , centered at ( $E[a_t] = 0$ ) and assumed independent of the previous observations.

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \dots + \phi_p \varepsilon_{t-p} + a_t$$

- The AR(1) model is the simplest (and one of the most useful) autoregressive model.

$$h(k, \phi) = \phi^k, k = 0, 1, \dots$$

- The single correlation parameter,  $\phi$ , represents the lag-1 correlation and takes values between -1 and 1.

## Semi-Variogram

- For unbalanced longitudinal data, either the correlation matrix or the scatter plot matrix can be used after discretizing the time scale.
- When the variance function suggest constant variance, the semi-variogram can be used as alternative method.
- Reconsider the general linear mixed model :

$$Y_i = X_i\beta + Z_i\mathbf{b}_i + \varepsilon_{(1)i} + \varepsilon_{(2)i}$$

- In this model we assume that  $\varepsilon_i$  has constant variance and can be decomposed as
- $\varepsilon_i = \varepsilon_{(1)i} + \varepsilon_{(2)i}$  in which
- $\varepsilon_{(2)i}$  is a component of serial correlation and
- $\varepsilon_{(1)i}$  is an extra component of measurement error reflecting variation added due to the measurement process.

# Semi-variogram

- Semi-variance is a measure of the spatial dependence between two observations as a function of the distance between them

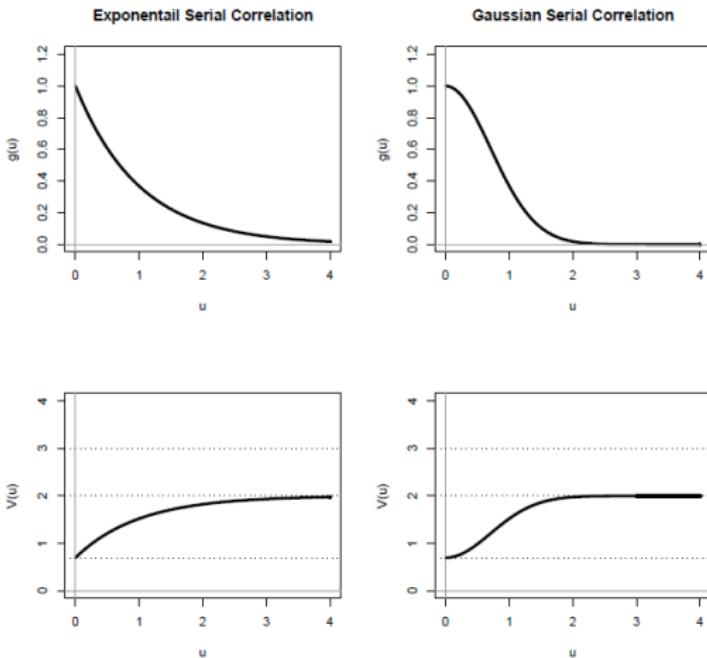


Figure 16: Semi-variogram for Exponential and Gaussian Serial Correlation

## Example: The Rat Data

- Stage 1 Model:  $Y_{ij} = \beta_{1i} + \beta_{2i}t_{ij} + \varepsilon_{ij}, i = 1, \dots, n_i$
- Stage 2 Models:

$$\begin{cases} \beta_{1i} = \beta_0 + b_{1i} \\ \beta_{2i} = \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i} \end{cases}$$

- Combined:  $Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i})t_{ij} + \varepsilon_{ij}$

$$= \begin{cases} \beta_0 + b_{1i} + (\beta_1 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if low dose} \\ \beta_0 + b_{1i} + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if high dose} \\ \beta_0 + b_{1i} + (\beta_3 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if control} \end{cases}$$

# The Rat Data

- Implied marginal mean structure:
  - Linear average evolution in each group
  - Equal average intercepts
  - Different average slopes
- Implied marginal covariance structure ( $\Sigma_i = \sigma^2 I_{n_i}$ )

$$\begin{aligned} \text{Cov}(Y_i(t_1), Y_i(t_2)) &= (1 - t_1) D \begin{pmatrix} 1 \\ t_2 \end{pmatrix} + \sigma^2 + \delta_{t_1, t_2} \\ &= d_{22} t_1 t_2 + d_{12}(t_1 + t_2) + d_{11} + \sigma^2 + \delta_{t_1, t_2} \end{aligned}$$

- The model assumes that the variance function is quadratic over time, with positive curvature  $d_{22}$ .

- A model which assumes that all variability in subject-specific slopes can be ascribed to treatment differences can be obtained by omitting the random slopes  $d_{2i}$  from the above model:

$$Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i) t_{ij} + \varepsilon_{ij}$$

$$= \begin{cases} \beta_0 + b_{1i} + \beta_1 t_{ij} + \varepsilon_{ij}, & \text{if low dose} \\ \beta_0 + b_{1i} + \beta_2 t_{ij} + \varepsilon_{ij}, & \text{if high dose} \\ \beta_0 + b_{1i} + \beta_3 t_{ij} + \varepsilon_{ij}, & \text{if control} \end{cases}$$

- This is the so-called random-intercepts model
- Implied marginal covariance structure ( $\Sigma_i = \sigma^2 I_{n_i}$ )

$$\begin{aligned} \text{Cov}(Y_i(t_1), Y_i(t_2)) &= (1) D(1) + \sigma^2 \delta_{t_1, t_2} \\ &= d_{11} + \sigma^2 + \delta_{t_1, t_2} \end{aligned}$$

- This implied that the covariance matrix is compound symmetry:
  - constant variance  $d_{ii} + \sigma^2$
  - constant correlation  $\rho_i = d_{11}/(d_{11} + \sigma^2)$  between any two repeated measurements within the same rat

# A Model for the Residual Covariance Structure

- Often,  $\Sigma_i$  is taken equal to  $\sigma^2 I_{n_i}$
- Then we obtain conditional independence:
- Conditional on  $b_i$ , the elements in  $Y_i$  are independent
- In the presence of no, or little, random effects, conditional independence is often unrealistic
- For example, the random intercepts model not only implies constant variance, it also implicitly assumes constant correlation between any two measurements within subjects
- Hence, when there is no evidence for (additional) random effects, or if they would have no substantive meaning, the correlation structure in the data can be accounted for in an appropriate model for  $\Sigma_i$

# A Model for the Residual Covariance Structure

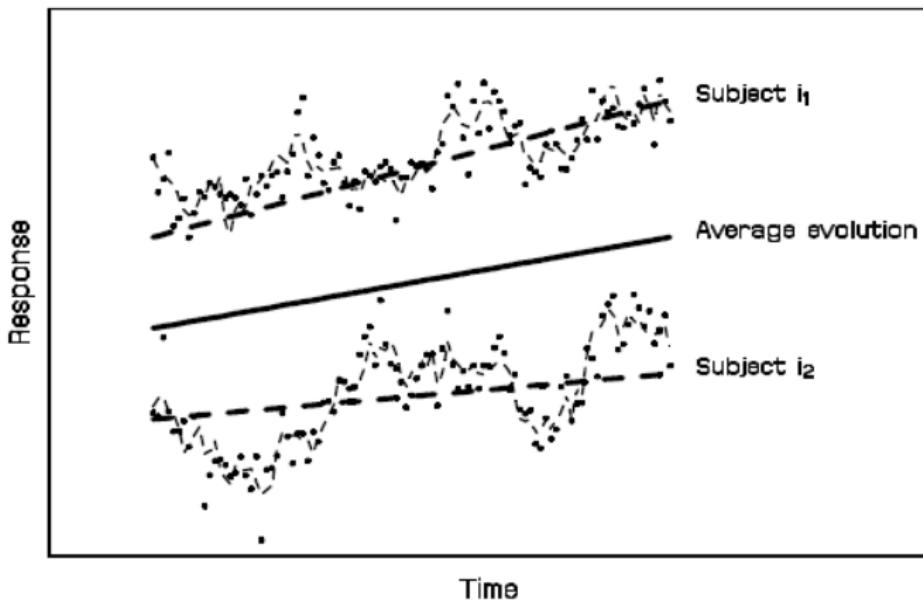
- Frequently used model is:

$$Y_i = X_i\beta + Z_i b_i + \underbrace{\varepsilon_{(1)i} + \varepsilon_{(2)i}}_{\varepsilon_i}$$

- 3 stochastic components:
  - $b_i$ : between-subject variability
  - $\varepsilon_{(1)i}$ : measurement error
  - $\varepsilon_{(2)i}$ : serial correlation component

- $\varepsilon_{(2)i}$  represents the belief that part of an individual's observed profile is a response to time-varying stochastic processes operating within that individual
- This results in a correlation between serial measurements, which is usually a decreasing function of the time separation between these measurements
- The correlation matrix  $H_i$  of  $\varepsilon_{(2)i}$  is assumed to have  $(j, k)$  element of the form  $h_{ijk} = g(|t_{ij} - t_{ik}|)$  for some decreasing function  $g(\cdot)$  with  $g(0) = 1$
- Frequently used functions  $g(\cdot)$ :
  - Exponential serial correlation:  $g(u) = \exp(-\phi u)$
  - Gaussian serial correlation:  $g(u) = \exp(-\phi u^2)$

- Graphical representation of all 4 components in the model:  
Stochastic components in general linear mixed model



## Estimation of the Marginal Model

- In general, the smaller, the stronger is the serial correlation
- Resulting final linear mixed model:

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

*independent* 
$$\begin{cases} b_i \sim N(0, D) \\ \varepsilon_{(1)i} \sim N(0, \sigma^2 I_{n_i}) \\ \varepsilon_{(2)i} \sim N(0, \tau^2 H_i) \end{cases}$$

- Note that inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects

# Estimation of the Marginal Model

- Notation:

- $\beta$ : vector of fixed effects
- $\alpha$ : vector of all variance components in  $D$  and  $\Sigma$
- $\theta = (\beta, \alpha)$ : vector of all parameters in marginal model

- Marginal likelihood function:

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

- If  $\alpha$  were known, MLE of  $\beta$  equals

$$\beta(\alpha) = \left( \sum_{i=1}^N X' V_i X_i \right)^{-1} \sum_{i=1}^N X' V_i y_i$$

- In most cases,  $\alpha$  is not known, and needs to be replaced by an estimated

# Maximum Likelihood Estimation (ML)

- Two frequently used estimation methods for  $\alpha$ :
  - Maximum likelihood
  - Restricted maximum likelihood
- $\bar{\alpha}_{ML}$  obtained from maximizing

$$L_{ML}(\alpha, \beta(\alpha))$$

with respect to  $\alpha$

- The resulting estimate for  $\beta$  will be denoted by  $\hat{\beta}_{ML}$
- $\bar{\alpha}_{ML}$  and  $\hat{\beta}_{ML}$  can also be obtained from maximizing  $L_{ML}(\theta)$  with respect to  $\alpha$  and  $\beta$  simultaneously

# Restricted Maximum Likelihood Estimation (REML)

## Variance Estimation in Normal Populations

- Consider a sample of  $N$  observations  $Y_1, \dots, Y_N$  from  $N(\mu, \sigma^2)$
- For known mean ( $\mu$ ), MLE of  $\sigma^2$  equals:  $\hat{\sigma}^2 = \sum_i^N (Y_i - \mu)^2 / N$
- Here,  $\hat{\sigma}^2$  is unbiased estimator for  $\sigma^2$
- When  $\mu$  is not known, MLE of  $\sigma^2$  equals:  $\hat{\sigma}^2 = \sum_i^N (Y_i - \bar{Y})^2 / N$
- Note that  $\hat{\sigma}^2$  is unbiased for  $\sigma^2$  implies that  $E(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2$
- The bias expression tells us how to derive an unbiased estimate:

$$S^2 = \sum_i^N (Y_i - \bar{Y})^2 / (N - 1)$$

- Estimating  $\mu$  apparently introduce bias in MLE of  $\sigma^2$

- So, we have to try to estimate  $\sigma^2$  without estimating  $\mu$  first. How can we do that?
- The model for all the data simultaneously:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \sim N \left( \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 I_N \right)$$

- we transform  $Y$  such that  $\mu$  vanishes from the likelihood

$$U = A' Y \sim N(0, \sigma^2 A' A)$$

- MLE of  $\sigma^2$  based on  $U$  equals;  $S^2 = \frac{1}{N-1} \sum_i (Y_i - \bar{Y})^2$
- $S^2$  is called REML estimate for  $\sigma^2$ , and  $S^2$  is independent of  $A$
- $A$  defines a set of  $N - 1$  linearly independent error contrasts

# Estimation of Residual Variance in Linear Regression Model

- Consider a sample of  $N$  observations  $Y_1, \dots, Y_N$  from linear regression model

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \sim N(X\beta, \sigma^2 I)$$

- MLE of  $\sigma^2$ :

$$\hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

- Note that  $\hat{\sigma}^2$  is biased for  $\sigma^2$

$$E(\hat{\sigma}^2) = \frac{N-p}{N}\sigma^2$$

- The bias expression tells us how to derive an unbiased estimate:

$$MSE = (Y - X\beta)'(Y - X\beta)/(N - p),$$

- The MSE can also be obtained from transforming the data orthogonal to  $X$ :

$$U = A' Y \sim N(0, \sigma^2 A' A)$$

- The MLE of  $\sigma^2$ , based on  $U$ , now equals the mean squared error, MSE
- The MSE is again called the REML estimate  $\sigma^2$

# REML for the Linear Mixed Model

- We first combine all models

$$Y_i \sim N(X_i\beta, V_i)$$

into one the model

$$Y \sim N(X\beta, V)$$

in which

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, V(\alpha) = \begin{pmatrix} V_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & V_N \end{pmatrix}$$

- Again, the data are transformed orthogonal to X:

$$U = A' Y \sim N(0, A' V(\alpha) A)$$

- The MLE of  $\alpha$ , based on  $U$  is called REML estimate, and denoted by  $\hat{\alpha}_{REML}$
- The resulting estimate  $\hat{\beta}(\hat{\alpha}_{REML})$  for  $\beta$  will be denoted by  $\hat{\beta}_{REML}$
- $\hat{\alpha}_{REML}$  and  $\hat{\beta}_{REML}$  can also be obtained from maximizing

$$L_{REML}(\theta) = \left| \sum_{i=1}^N X' W_i(\alpha) X_i \right|^{-1/2} L_{ML}(\theta)$$

with respect to  $\theta$  ( $\alpha$  and  $\beta$ ) simultaneously

- $L_{REML}(\alpha, \hat{\beta}(\alpha))$  is the likelihood of error contrast  $U$ , and is often called the REML likelihood function
- Note that  $L_{REML}(\theta)$  is not the likelihood for our original data  $Y$ .

# **Chapter 6**

## **Fitting Linear Mixed Models in R**

# 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 the examples that are based on normal theory our focus will be on 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 ( $Female = 0$ ,  $Male = 1$ )
- $b_{0i}$ : is random intercept;  $b_{1i}$ : is 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 =
```
- `fixed` is an argument to define fixed effects portion
- `random` is an argument to define the random effects portion
- `data` 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
- the fixed part and the random parts:
  - The fixed part is  $fixed = distance \sim Sex + Sex * age$
  - The random part is  $random = \sim 1 | Subject$
- If the random part is specified as above it means we will fit a model with random intercept
- Here the response is specified only on fixed part.
- In the random part the model statement begins with just a  $\sim$
- 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:

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

- $D_{ij}$ : Orthodontic distance of the  $i^{th}$  child at the  $j^{th}$  visit.
- $A_{ij}$ : Age of the  $i^{th}$  child at the  $j^{th}$  visit,  $A'_{ij} = A_{ij} - 8$
- $S_i$ : Sex of the  $i^{th}$  child ( $boys = 1, girls = 2$ )
- $b_{0i}$ : is random intercept;  $b_{1i}$ : is 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 = mydata22, random = ~ 1|ind)
```

- 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

## Growth data

```
> print(growth.fit1)
Linear mixed-effects model fit by REML
Data: mydata22
Log-restricted-likelihood: -203.0748
Fixed: measure ~ sex + sex * age
(Intercept)          sex           age       sex:age
15.3820544    0.9177677   1.0846088  -0.2976836
```

Random effects:

```
Formula: ~1 | ind
(Intercept) Residual
StdDev:     1.836499 1.440418
```

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;

	(Intercept)	sex	age	sex:age
1	14.32099	0.9177677	1.084609	-0.2976836
2	15.72939	0.9177677	1.084609	-0.2976836
3	16.13251	0.9177677	1.084609	-0.2976836
4	17.35446	0.9177677	1.084609	-0.2976836
5	15.40437	0.9177677	1.084609	-0.2976836
6	14.05792	0.9177677	1.084609	-0.2976836
7	15.72939	0.9177677	1.084609	-0.2976836
8	16.05440	0.9177677	1.084609	-0.2976836
9	14.05792	0.9177677	1.084609	-0.2976836
10	11.70671	0.9177677	1.084609	-0.2976836
11	18.65453	0.9177677	1.084609	-0.2976836
12	17.80363	0.9177677	1.084609	-0.2976836
13	14.09445	0.9177677	1.084609	-0.2976836
14	14.77016	0.9177677	1.084609	-0.2976836
15	16.82859	0.9177677	1.084609	-0.2976836
16	13.40292	0.9177677	1.084609	-0.2976836
.	.	.	.	.

## Fixed effect parameters

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

```
> fixef(growth.fit1)
(Intercept)          sex          age      sex:age
15.3820544    0.9177677    1.0846088   -0.2976836
```

- 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 = mydata22, 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

Approximate 95\% confidence intervals

Fixed effects:

	lower	est.	upper	
(Intercept)	10.6739882	15.3842162	20.09444414	
sex	-2.3469983	0.9188003	4.18459902	
age	0.7106699	1.0844737	1.45827751	
sex:age	-0.5486437	-0.2977482	-0.04685262	
attr(,"label")				
[1]	"Fixed effects:"			

Random Effects:

Level: ind

	lower	est.	upper	
sd((Intercept))	1.281501	1.759342	2.415358	

Within-group standard error:

	lower	est.	upper	
	1.206178	1.420339	1.672526	

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

```
Linear mixed-effects model fit by maximum likelihood
Data: mydata22
AIC      BIC      logLik
413.3128 428.8835 -200.6564

Random effects:
Formula: ~1 | ind
(Intercept) Residual
StdDev:    1.759342 1.420339

Fixed effects: measure ~ sex + sex * age
Value Std.Error DF t-value p-value
(Intercept) 15.384216 2.4108899 70 6.381136 0.0000
sex          0.918800 1.6187332 25 0.567605 0.5754
age          1.084474 0.1913283 70 5.668131 0.0000
sex:age     -0.297748 0.1284187 70 -2.318573 0.0233

Correlation:
(Intr) sex    age
sex     -0.944
age     -0.881  0.832
sex:age  0.832 -0.881 -0.944

Standardized Within-Group Residuals:
Min        Q1        Med        Q3        Max
-3.331674862 -0.531945925 -0.009607243  0.482544286  3.599008895

Number of Observations: 99
Number of Groups: 27
```

- 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 appears 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  $random = \sim age|subject$

```
library(nlme)
growth.fit2 <- lme(fixed = measure ~ sex+sex*age,
data = mydata22, random = ~ age|ind)
```

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

```
ind = pdLogChol(age)
Variance     StdDev     Corr
(Intercept) 8.35519280 2.8905350 (Intr)
age          0.04415048 0.2101202 -0.766
Residual    1.76655402 1.3291178
```

# Inference for the Marginal Model

- Inference for fixed effects:
  - Wald test
  - t-test and F-test
  - Robust inference
  - LR test
- Inference for variance components:
  - Wald test
  - LR test
- Information criteria

## Inference for the Fixed Effects

- Estimate for  $\beta$ :

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

with  $\alpha$  replaced by its ML or REML estimate

- Conditional on  $\alpha$ ,  $\hat{\beta}(\alpha)$  is multivariate normal with mean  $\beta$  and covariance  $Var(\hat{\beta})$

$$\begin{aligned} Var(\hat{\beta}) &= \left( \sum_{i=1}^N X' V_i X_i \right)^{-1} \left( \sum_{i=1}^N X' V_i Var(Y_i) X_i \right) \left( \sum_{i=1}^N X' V_i X_i \right)^{-1} \\ &= \left( \sum_{i=1}^N X' V_i X_i \right)^{-1} \end{aligned}$$

## Approximate Wald Test

- For any known matrix  $L$ , consider testing

$$H_0 : L\beta = 0, \text{ versus } H_A : L\beta \neq 0$$

- Wald test statistics

$$G = \hat{\beta}' L' \left[ L \left( \sum_{i=1}^N X_i' V_i^{-1}(\alpha) X_i \right)^{-1} L' \right]^{-1} \hat{\beta}' L$$

- Asymptotic null distribution of  $G$  is  $\chi^2$  with  $\text{rank}(L)$  degrees of freedom

## Approximate t-test and F-test

- Wald test based on

$$Var(\hat{\beta}) = \left( \sum_{i=1}^N X' V_i X_i \right)^{-1}$$

- Variability introduced from replacing  $\alpha$  by some estimate is not taken into account in Wald tests
- Therefore, Wald tests will only provide valid inferences in sufficiently large samples
- In practice, this is often resolved by replacing the  $\chi^2$  distribution by an appropriate F-distribution (are the normal by a t)
- For any known matrix L, consider testing

$$H_0 : L\beta = 0, \text{ versus } H_A : L\beta \neq 0$$

- F test statistics

$$F = \frac{\hat{\beta}' L' \left[ L \left( \sum_{i=1}^N X_i' V_i^{-1}(\alpha) X_i' \right)^{-1} L' \right]^{-1} \hat{\beta}' L}{\text{rank}(L)}$$

- Approximate null-distribution of F is F with numerator degrees of freedom equal to  $\text{rank}(L)$
- Denominator degrees of freedom to be estimated from the data:
  - Containment method
  - Satterthwaite approximation
  - Kenward and Roger approximation
  - ....
- In the context of longitudinal data, all methods typically lead to large numbers of degrees of freedom, and therefore also to very similar p-values
- For univariate hypotheses ( $\text{rank}(L)=1$ ) the F-test reduces to a t-test

# Likelihood Ratio Test

- Comparison of nested models with different mean structures, but equal covariance structure
- Test Statistics

$$-2\ln\lambda_N = -2\ln \left[ \frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})} \right]$$

- Asymptotic null distribution:  $\chi^2$  with d.f. equal to difference in dimension of  $\Theta_\beta$

## LR Test for Fixed Effects Under REML

- How can the negative LR test statistic be explained?
- Under REML, the response  $Y$  is transformed into error contrasts  $U = A' Y$  for some matrix  $A$  with  $A' X = 0$
- Afterwards, ML estimation is performed based on the error contrasts
- The reported likelihood value,  $L_{REML}(\hat{\theta})$  is the likelihood at maximum for the error contrasts  $U$
- Models with different mean structures lead to different sets of error contrasts
- Hence, the corresponding REML likelihoods are based on different observations, which makes them no longer comparable
- LR tests for the mean structure are not valid under REML

# Inference for the Variance Components

- Inference for the mean structure is usually of primary interest.
- However, inferences for the covariance structure is of interest as well:
  - interpretation of the random variation in the data
  - overparameterized covariance structures lead to inefficient inferences for mean
  - too restrictive models invalidate inferences for the mean structure
- Asymptotically, ML and REML estimates of  $\alpha$  are normally distributed with correct mean and inverse Fisher information matrix as covariance

## Caution with Wald Tests for Variance Components

- Wald test:  $H_0 : d_{33} = 0$
- Under the hierarchical model interpretation, this null-hypothesis is not of any interest, as  $d_{23}$  and  $d_{13}$  should also equals zero when ever  $d_{33} = 0$
- Hence, the test is meaningful under the marginal model only, i.e., when no underlying random effects structure is believed to describe the data
- Boundary Problems
  - The quality of the normal approximation for ML or REML estimate strongly depends on the true value of  $\alpha$
  - Poor normal approximation if  $\alpha$  is relatively close to the boundary of the parameter space
  - If  $\alpha$  is a boundary value, the normal approximation completely fails

## Boundary Problems

- Under the hierarchical model interpretation,  $d_{33} = 0$  is a boundary value
- This imply that the calculation of the above p-value is based on an incorrect null-distribution for the Wald test statistic
- Indeed, how could ever, under  $H_0$ ,  $d_{33}$  be normally distributed with mean 0, if  $d_{33}$  is estimated under the restriction  $d_{33} \geq 0$ ?
- Hence, the test is only correct, when the null-hypothesis is not a boundary value (e.g.,  $H_0 : d_{33} = 0.1$ )
- Note that, even under the hierarchical model interpretation, a classical Wald test is valid for testing  $H_0 : d_{23} = 0$
- Likelihood ratio test for the variance component also suffers with the boundary problem
- Hence, ignoring the boundary problem may invalidate inferences, even for the mean structure

## 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 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
- The model is selected with the highest penalized log-likelihood
- Information criteria are no formal testing procedures
- For the comparison of models with different mean structures, information criteria 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

## Model comparison

- When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over fitting
- Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model
- the penalty term is larger in BIC than in AIC
- Akaike Information Criterion (AIC)

$$AIC = -2\log Lik + 2npar,$$

- Bayesian Information Criterion (BIC)

$$BIC = -2\log Lik + npar \log(N),$$

- Where  $npar$  = number of parameters in the model
- $N$  = Total number of observations used to fit the model
- Both AIC and BIC pick the model with the smallest value
- BIC penalizes model complexity more heavily
- Multilevel models and complicated structural equation models, then the BIC is used more frequently than the AIC

# Inference for the Random Effects

- Empirical Bayes inference
- Best linear unbiased prediction
- Example:
- Shrinkage
- Example: Random-intercepts model
- Example:
- Normality assumption for random effects

# Empirical Bayes Inference

- Random effects  $b_i$  reflect how the evolution for the  $i$ th subject deviates from the expected evolution  $X_i\beta$
- Estimation of the  $b_i$  helpful for detecting outlying profiles
- This is only meaningful under the hierarchical model interpretation:

$$Y_i|b_i \sim N(X_i\beta + Z_i b_i, \Sigma_i) \quad b_i \sim N(0, D)$$

- Since the  $b_i$  are random, it is most natural to use Bayesian methods
- Terminology: prior distribution  $N(0, D)$  for  $b_i$

- Posterior density:

$$F(b_i|y_i) \equiv f(b_i|Y_i = y_i) = \frac{F(b_i|y_i)f(b_i)}{\int F(b_i|y_i)f(b_i)db_i}$$

- Posterior distribution:

$$b_i|y_i \sim N(DZ_i W_i(y_i - X_i\beta), \Lambda_i)$$

for some positive definite matrix  $\Lambda_i$

- Posterior mean as estimate for  $b_i$ :
- Parameters in  $\theta$  are replaced by their ML or REML estimates, obtained from fitting the marginal model.
- $\hat{\beta}_i = \hat{\beta}_i(\hat{\theta})$  is called the **Empirical Bayes** estimate of  $\beta_i$
- Approximate t- and F-tests to account for the variability introduced by replacing  $\theta$  by  $\hat{\theta}$ , similar to tests for fixed effects

## Best Linear Unbiased Prediction (BLUP)

- Often, parameters of interest are linear combinations of fixed effects in and random effects in  $b_i$ ;
- For example, a subject-specific slope is the sum of the average slope for subjects with the same covariate values, and the subject-specific random slope for that subject
- In general, suppose  $u = \lambda_{\beta}'\beta + \lambda_b'b_i$
- Conditionally on  $\alpha$ ,  $\bar{u} = \lambda_{\beta}'\bar{\beta} + \lambda_i'\bar{b}_i$ 
  - linear in the observations  $Y_i$
  - unbiased for  $u$
  - minimum variance among all unbiased linear estimators

## Example: Rate data

- We reconsider the reduced model:

$$Y_{ij} = \beta_0 + \beta_1 \text{Age}_{ij} + \beta_2 \text{Treat}_i + \beta_3 \text{Treat}_i \text{Age}_{ij} + b_{0i} + b_{2i} \text{Age}_{ij} + \epsilon_{ij}$$

- In R the estimates can be obtained using the following code to the random statement:

```
ranef(growth.fit2, condVar=TRUE)
```

- In practice, histograms and scatterplots of certain components of  $b_i$  are used to detect model deviations or subjects with 'exceptional' evolutions over time

## **Conclusion:**

- No statistically significant difference in orthodontic distance among boys and girls at the start
- The evolution over time (rate of change) is higher among males

# Assignment

- Consider Jimma infant growth Data.
- Outcome Variable ..... height of infants measured longitudinally
- Factors..... Possible covariates from the data

# Instruction I

- Compute Summary Statistics
- Fit an appropriate liner mixed effects model and interpret the findings
- Consider linear regression, and compare and contrast with your results in (c)

## **Part II:**

# **Models for Non-Gaussian Longitudinal Data**

# **Chapter 7:**

## **Marginal Models for Non-Gaussian Longitudinal Data**

## Introduction

- Repeated measurement occurs commonly in health-related applications
- In such studies, the response variable for each subject is measured repeatedly, at several times
- Correlated observations can also occur when the response variable is observed for matched sets of subjects
- Observations within a cluster are usually positively correlated
- Analyses should take the correlation into account
- Analyses that ignore the correlation can estimate model parameters well, but the standard error estimators can be badly biased
- As with independent observations, with clustered observations models focus on how the probability of a particular outcome depends on explanatory variables.

## The Toenail Data

- Toenail Dermatophyte Onychomycosis: Common toenail infection, difficult to treat, affecting more than 2% of 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 for the comparison of two such new compounds (A and B) for oral treatment.

- Research question: Severity relative to treatment of TDO?
- $2 \times 189$  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.

## The Jimma Infant Data

- It is of particular interest to identify the risk of overweight in early life through weight and height measurements
- This helps in prevention of overweight and obesity to reduce incidence of several adulthood diseases
- One possible indicator of overweight is age- and sex- specific BMI, with a BMI over the 85th percentile referring to overweight
- The outcome of interest is BMI coded as 0 (normal or underweight) or 1 (over weight)
- The question of interest is whether the percentage of overweight changes over time (age), differs for gender.

## The Epilepsy Study

- The epileptic data set considered here is obtained from a randomized, multi-center study
- Comparison of placebo with a new anti-epileptic drug (AED)
- In the study, 45 patients were randomized to the placebo group and 44 to the active (new) treatment group
- The number of epileptic seizures were measured on a weekly basis during a 16 weeks period
- After this period, patients were entered into a long-term study up to 27 weeks
- The key research question is whether or not the additional new treatment reduces the number of epileptic seizures

# The Gilgel-Gibe Mosquito Data

- A study conducted around Gilgel-Gibe dam for three years.
- Influence of the dam on mosquito abundance and species composition.
- Eight 'At risk' and eight 'Control' villages based on distance.
- One collection approach: IRC.
- Mosquito species were identified and counted.
- An. gambaie was found to be the dominant one (more than 95%).

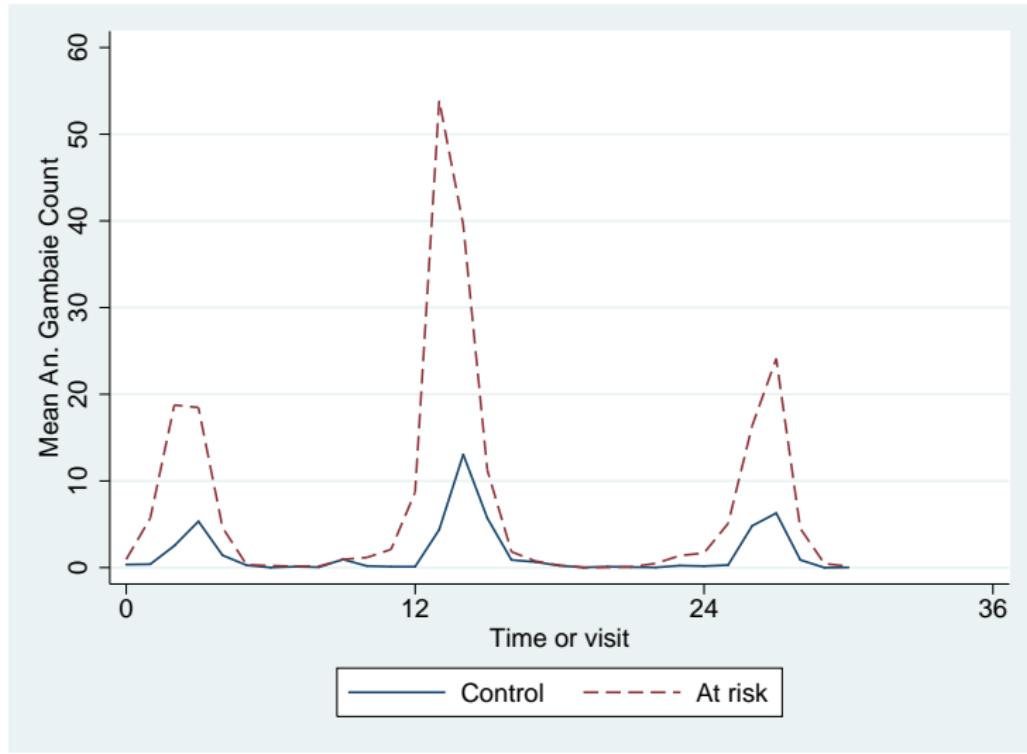


Figure 17: Average evolution of An. gambiae

## The Gilgel-Gibe Mosquito Cont'd

- At-risk seems to be consistently higher.
- There is a clear seasonality pattern.
- Fluctuation between wet and dry season.

# The Generalized Linear Model

- Suppose a sample  $Y_1, \dots, Y_N$  of independent observations is available
- All  $Y_i$  have densities  $f(y_i|\theta_i, \phi)$  which belongs to the exponential family

$$f(y_i|\theta_i, \phi) = \exp(\psi^{-1}[y\theta_i - \Psi(\theta_i)] + c(y, \psi))$$

- $\theta_i$  is the natural parameters
- Linear predictor:  $\theta_i = x_i\beta$
- $\theta$  is the scale parameter (over dispersion parameter)
- $\Psi(\cdot)$  is a function to be discussed next

## Mean and Variance

- We start from the following general property:

$$\int f(y|\theta, \psi) dy = \int \exp(\psi^{-1}[y\theta_i - \Psi(\theta_i)] + c(y, \psi)) = 1$$

- Taking first and second-order derivatives with respect to  $\theta$  yields

$$\begin{cases} \frac{\partial}{\partial \theta} \int f(y|\theta, \psi) dy = 0 \\ \frac{\partial^2}{\partial \theta^2} \int f(y|\theta, \psi) dy = 0 \end{cases}$$

$$\begin{cases} E(Y) = \Psi'(\theta) \\ Var(Y) = \phi \Psi''(\theta) \end{cases}$$

## Example: The Normal Model

- Model

$$Y \sim N(\mu, \sigma^2)$$

- Density Function:

$$\begin{aligned} f(y|\theta, \psi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{\sigma^2}(y - \mu)^2\right) \\ &= \exp\left(-\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) + \left(\frac{\ln 2\pi\sigma^2}{2} - \frac{y^2}{2\sigma^2}\right)\right) \end{aligned}$$

- Exponential family;
    - $\theta = \mu$
    - $\phi = \sigma^2$
    - $\Psi(\phi) = \theta^2/2$
    - $c(y, \phi) = \frac{\ln(2\pi\psi)}{2} - \frac{y^2}{2\psi}$
  - The mean and variance functions:
    - $\mu = \theta$
    - $Var(\mu) = 1$
  - Note that, under this normal model, the mean and variance are not related:
- $$\psi v(\mu) = \sigma^2$$
- The link function is here the identity function:  $\theta = \mu$

# The Bernoulli Model

- Model

$$Y \sim Bernoilli(\pi)$$

- Density function

$$\begin{aligned} f(y|\theta, \phi) &= \pi^y (1-\pi)^{(1-y)} \\ &= \exp(y \ln \pi + (1-y) \ln(1-\pi)) \\ &= \exp\left(y \ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right) \end{aligned}$$

- Exponential family:

- $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$
- $\phi = 1$
- $\psi(\mu) = \ln(1-\pi) = \ln(1 + \exp(-\theta))$
- $c(y, \phi) = 0$

- Mean and variance function:

- $\mu = \frac{\exp\theta}{1+\exp\theta} = \pi$

- $v(\mu) = \frac{\exp\theta}{(1+\exp\theta)^2} = \pi(1 - \pi)$

- Note that, under this model, the mean and variance are related:

$$\phi v(\mu) = \mu(1 - \mu)$$

- The link function here is the logit link:

$$\theta = \ln\left(\frac{\mu}{1 - \mu}\right)$$

# The Poisson Model

- Model:

$$Y \sim \text{Poisson}(\lambda)$$

- Density function:

$$\begin{aligned} f(y|\theta, \phi) &= \frac{e^{-\lambda} \lambda^y}{y!} \\ &= \exp(y \ln \lambda - \lambda - \ln y!) \end{aligned}$$

- Exponential family:

- $\theta = \ln \lambda$
- $\phi = 1$
- $\psi(\theta) = \lambda = \exp \theta$
- $c(y, \phi) = -\ln y!$

- Mean and variance function:

- $\mu = \exp \theta = \lambda$
- $v(\mu) = \exp \theta = \lambda$

- The link function is here the log link:  $\theta = \ln \mu$

# Maximum Likelihood Estimation

- In Generalized Linear model, the parameter  $\beta$  is the corresponding vector of unknown regression parameters, to be estimated from the data.
- Log-likelihood

$$\ell(\beta, \phi) = \frac{1}{\phi} \sum_i [y_i \theta_i - \psi_i(\theta_i)] + \sum_i c(y_i, \phi)$$

- First order derivative with respect to  $\beta$ :

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta} = \frac{1}{\phi} \sum_i \frac{\partial \theta_i}{\partial \beta} [y_i - \psi'_i(\theta_i)]$$

- The score equations for  $\beta$  to be solved:

$$S(\beta) = \sum_i \frac{\partial \theta_i}{\partial \beta} (y_i - \psi'(\theta_i)) = 0$$

- Since  $\mu_i = \psi'(\theta)$  and  $v_i = v(\mu_i) = \psi''(\theta_i)$
- The score equation now becomes

$$S(\beta) = \sum_i \frac{\partial \mu_i}{\partial \beta} v_i (y_i - \mu_i) = 0$$

- Note that the estimation of  $\beta$  depends on the density only through the means  $\mu_i$  and the variance functions  $v_i = v(\mu_i)$
- The score equations need to be solved numerically:
  - iterative (re-)weighted least squares
  - Newton-Raphson
  - Fisher scoring
- Inference for  $\beta$  is based on classical maximum likelihood theory:
  - asymptotic Wald tests
  - likelihood ratio tests
  - score tests

- In some cases,  $\psi$  is a known constant, in other examples, estimation of  $\psi$  may be required to estimate the standard errors of the elements in  $\beta$
- Estimation can be based on  $\text{Var}(Y_i) = \phi v_i$ :

$$\hat{\phi} = \frac{1}{N-p} \sum_i (y_i - \hat{\mu}_i)^2 / v_i(\hat{\mu}_i)$$

- For example, under the normal model, this would yield:

$$\hat{\phi} = \frac{1}{N-p} \sum_i (y_i - x_i' \hat{\beta}_i)$$

- the mean squared error used in linear regression models to estimate the residual variance.

## Marginal Versus Conditional models

- Marginal models are population-average models whereas conditional models are subject-specific
- Interpretation 1: a 1 unit increase in covariate  $x$  is associated with a  $z$ -unit average increase in the outcome variable
- Interpretation 2: Conditional model you would say something like a 1 unit increase in covariate  $x$  is associated with a  $Z$ -unit average increase in response variable, holding each random effect for individual constant

# Generalized Estimating Equations (GEE)

- Marginal model for non-Gaussian longitudinal data
- Extend GLM to accommodate the modeling of correlated data
- Repeated nature of the data is modeled based on ‘working correlation’
- Same form as for full likelihood procedure, but we restrict specification to the first moment only
- GEE analysis IS only suitable for a two-level structure
- When a three-level structure exists in a longitudinal study, only multilevel analysis can be used

- Rather than assuming a particular type of distribution for  $(Y_1, \dots, Y_T)$ , this method only links each marginal mean to a linear predictor and provides a guess for the variance–covariance structure of  $(Y_1, \dots, Y_T)$
- The method uses the observed variability to help generate appropriate standard errors.
- The method is called the GEE method because the estimates are solutions of generalized estimating equations
- These equations are multivariate generalizations of the equations solved to find ML estimates for GLMs

# Generalized Estimating Equations

- Let  $Y_{ij}, j = 1, \dots, n, i = 1, \dots, K$  represent the  $j$ th measurement on the  $i$ th subject
- There are  $n_i$  measurements on subject  $i$  and  $\sum_{i=1}^k$  total measurements
- Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case
- The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled

- Once we have specified a marginal model for each  $Y_t$ , for the GEE method we must:
- Assume a particular distribution for each  $Y_t$ . This determines how  $\text{Var}(Y_t)$  depends on  $E(Y_t)$
- Make an educated guess for the correlation structure among  $Y_t$
- This is called the working correlation matrix
- ML fitting of marginal logit models is difficult
- Model-based version and Empirically-corrected version
- One possible working correlation has exchangeable structure
  - This treats  $\rho = \text{Corr}(Y_s, Y_t)$  as identical (but unknown) for all pairs  $s$  and  $t$

- Let the vector of measurements on the  $i$ th subject be  $Y_i = [Y_{i1}, \dots, Y_{ini}]'$  with corresponding vector of means  $\mu_i = [\mu_{i1}, \dots, \mu_{ini}]'$  and let  $V_i$  be an estimate of the covariance matrix of  $Y_i$
- The Generalized Estimating Equation for estimating  $\beta$  is an extension of the independence estimating equation to correlated data and is given by

$$\sum_{i=1}^k \frac{\partial \mu_i}{\partial \beta} V_i^{-1} (Y_i - \mu_i(\beta))$$

## Correlation Structures

- For the assumed working correlation structure, the GEE method uses the data to estimate the correlations.
- Those correlation estimates also impact the estimates of model parameters and their standard errors.
- When the correlations are small, all working correlation structures yield similar GEE estimates and standard errors.
- Unless one expects dramatic differences among the correlations, we recommend using the exchangeable working correlation structure.

- Even if your guess about the correlation structure is poor, valid standard errors result from an adjustment the GEE method makes using the empirical dependence the actual data exhibit
- That is, the naive standard errors based on the assumed correlation structure are updated using the information the sample data provide about the dependence
- The result is robust standard errors that are usually more appropriate than ones based solely on the assumed correlation structure

## Working Correlation Matrix

- suppose  $V_i(\cdot)$  is not the true variance of  $Y_i$  but only a plausible guess, a so-called working correlation matrix
- Let  $R_i(\alpha)$  be an  $n_i \times n_i$  working correlation matrix that is fully specified by the vector of parameters  $\alpha$
- The covariance matrix of  $Y_i$  is modelled as

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

- $\phi$  is a scale (over dispersion) parameter
- where  $A_i$  is an  $n_i \times n_i$  diagonal matrix with  $v(\mu_{ij})$  as the  $i$ th diagonal element
- If  $R_i(\alpha)$  is the true correlation matrix of  $Y_i$ , then  $V_i$  is the true covariance matrix of  $y_i$ .

- The working correlation matrix is not usually known and must be estimated.
- It is estimated in the iterative fitting process using the current value of the parameter vector  $\beta$  to compute appropriate functions of the Pearson residual

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

- There are several specific choices of the form of working correlation matrix  $R_i(\alpha)$  commonly used to model the correlation matrix of  $Y_i$ .
- A few of the choices are shown below
- The dimension of the vector  $a$ , which is treated as a nuisance parameter, and the form of the estimator of  $a$  are different for each choice

# Types of Working Correlation Matrix

- Consider four repeated measurements from each study participants.  
Some typical choices for the correlation structure are:
  - The independence working correlation structure assumes  $\text{Corr}(Y_s, Y_t) = 0$  for each pair. This treats the observations in a cluster as uncorrelated
    - $(R_i(\rho) = R_0)$ , a fixed correlation matrix
    - For  $R_0 = I$ , the identity matrix, the GEE reduces to the independence estimating equation.

$$R(\alpha) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Exchangeable:
  - $\text{corr}(Y_{ij}, Y_{ik}) = \rho, j \neq k$

$$R(\alpha) = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

## Example: Toenail Data

- Autoregressive (AR-1):

- $\text{corr}(Y_{ij}, Y_{ik}) = \rho^{t_{ij} - t_{ik}}$

$$R(\alpha) = \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^3 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$$

- Unstructured:

- $\text{corr}(Y_{ij}, Y_{ik}) = \rho_{jk}$

$$R(\alpha) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{bmatrix}$$

## Example: Toenail Data

- Variables in the data:
  - obs: observation number
  - treat: treatment group (0: Itraconazol (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)
- The research question is whether treatment has effect in curing the infection or not.

## Fitting GEE model in R

- 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, dat
```

- *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 hold the variables
- *corStructure* the correlation structure: 'ar1', 'exchangeable', 'independence', 'fixed' or 'unstructure'.
- *clusterID* the name of the column that hold the cluster IDs
- *startCoeff* a numeric vector, the starting values for the beta coefficients
- *maxit* an integer, the maximum number of iteration to use for convergence.

- To fit GEE in R, you need the following packages first

- geepack*
- wgeesel*
- MuMIn*

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

summary(fit1)
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)
Model:
Link:           Logit
Variance to Mean Relation: Binomial
Correlation Structure: Exchangeable
Call:
gee(formula = y ~ treatn + time + treatn * time, id = idnum,
  data = Toenail, family = binomial, corstr = "exchangeable")
Summary of Residuals:
Min      1Q   Median      3Q      Max
-0.3603 -0.2495 -0.1037 -0.0232  0.9768
```

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-0.58406	0.1406	-4.1545	0.1734	-3.3685
treatn	0.00997	0.1952	0.0511	0.2608	0.0382
time	-0.17703	0.0218	-8.1100	0.0311	-5.6903
treatn:time	-0.08668	0.0377	-2.3016	0.0568	-1.5259

Estimated Scale Parameter: 1.09

Number of Iterations: 4

Working Correlation

[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	1.000	0.421	0.421	0.421	0.421	0.421
[2,]	0.421	1.000	0.421	0.421	0.421	0.421
[3,]	0.421	0.421	1.000	0.421	0.421	0.421
[4,]	0.421	0.421	0.421	1.000	0.421	0.421
[5,]	0.421	0.421	0.421	0.421	1.000	0.421
[6,]	0.421	0.421	0.421	0.421	0.421	1.000
[7,]	0.421	0.421	0.421	0.421	0.421	1.000

# Choosing the Best Model

- For GEE
  - QIC(V) – function of V, so can use to choose best correlation structure.
  - QIC<sub>U</sub> – measure that can be used to determine the best subsets of covariates for a particular model.
  - The best model is the one with the smallest value!
- GEE works best if
  - The number of observations per subject is small and the number of subjects is large
  - In longitudinal studies the measurements are taken at the same times for all subjects

- Based on the working correlation structure, there is considerable correlation
- Different correlation structure can be also used
- The best model can be selected using QIC values
- The smaller the better

- QIC was computed for each models

```
> model.sel(fit1,fit3, fit3, rank = QIC)
Model selection table
      (Int)      tim     trt tim:trt corstr qLik   QIC delta weight
fit3  -0.800 -0.122  0.0980 -0.1348 unstrc -911 1832  0.00  0.397
fit31 -0.800 -0.122  0.0980 -0.1348 unstrc -911 1832  0.00  0.397
fit1  -0.587 -0.177  0.0084 -0.0871 exchng -906 1833  1.32  0.206
Abbreviations:
corstr: exchng = 'exchangeable', unstrc = 'unstructured'
Models ranked by QIC(x)
```

- The function *sapply* can be used to compute QIC for the three model
- R code to compute QIC

```
sapply(list(fit1, fit2, fit3), QIC) # QIC for the different mo
```

QIC Results for the three models

```
> sapply(list(fit1, fit2, fit3), QIC)
QIC  QIC  QIC
1833 1832 1832
```

# Conclusion

- The model with the lowest QIC is better
- Autoregressive and exchangeable covariance structure resulted the same QIC

# Alternating Logistic Regression

- ALR seem to offer some of the advantages of marginal models estimated via generalized estimating equations (GEE) and generalized linear mixed models (GLMMs)
- It models the association between pairs of responses by using log odds ratios instead of using correlations, as ordinary GEEs do
- When marginal odds ratios are used to model, association can be estimated using ALR, which is
  - almost as efficient as GEE2
  - almost as easy (computationally) than GEE1

$$\text{logitPr}(Y_{ij} = 1 | x_{ij}) = x_{ij}\beta$$

$$\text{logitPr}(Y_{ij} = 1 | Y_{ik} = y_{ij}) = \alpha_{ij}y_{ik}x_{ik} + \ln \frac{\mu_{ij} - \mu_{ik}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}$$

- Simultaneously regressing the response on explanatory variables as well as modelling the association among responses in terms of pairwise odds ratios
- $\alpha_{ijk}$  can be modelled in terms of predictors
- the second term is treated as an offset
- the estimating equations for  $\beta$  and  $\alpha$  are solved in turn, and the alternating between both sets is repeated until convergence
- this is needed because the offset clearly depends on *beta*

- The R code to fit ALR is
- `install.packages(alr)`

```
a1 <- alr(alr.y ~ alr.x - 1, id=alr.id, depm="exchangeable", ainit=0.01)
```

- Reading assignment: **Alternative Logistic Regression**

# Generalized Linear Mixed Models (GLMM)

- For non-Gaussian data, the well-known generalized linear mixed model is commonly used
- The linear predictor contains random effects in addition to the usual fixed effects
- These random effects are usually assumed to come from a normal distribution
- Responses Correlated:  $g(E(Y|b)) = X\beta + Zb$
- Correlation modeled in part by random effects
- Analysis describes differences in the mean of Y conditional on the patient's specific random effect b
- Most relevant from an individual patient's perspective Often b represent a dimension of frailty-Hence,  $X\beta$  tells about the relationship of Y to X among patients with the same frailty

- Let  $Y_{ij}$  be the  $j$ th outcome measured for subject  $i = 1, \dots, N$ ,  $j = 1, \dots, n_i$  and group the  $n_i$  measurements into a vector  $Y_i$
- Given a vector  $b_i$  of random effects for cluster  $i$ , it is assumed that all responses  $Y_{ij}$  are independent, with density of the form

$$f_i(y_{ij}|\theta_{ij}, \phi) = \exp \left\{ \phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi) \right\}$$

- $\theta_{ij}$  is now modelled as

$$\theta_{ij} = x_{ij}\beta + Z'_{ij}b_i$$

- It is assumed that  $b_i \sim N(0, D)$

- Let  $f_{ij}(y_{ij}|b_i, \beta, \phi)$  denote the conditional density of  $Y_{ij}$  given  $b_i$ , the conditional density of  $Y_i$  equals

$$f_i(y_i|b_i, \beta, \phi) = \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \beta, \phi)$$

- The marginal distribution of  $Y_i$  is given by

$$\begin{aligned} f_i(y_i|\beta, D, \phi) &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \beta, \phi) f(b_i|D) db_i \\ &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \beta, \phi) f(b_i|D) db_i; \end{aligned}$$

where  $f(b_i|D)$  be the density of the  $N(0, D)$  distribution for the random effects  $b_i$ , and  $\phi$  a scale (overdispersion) parameter

- The likelihood function for  $\beta, D$  and  $\phi$  now equals

$$\begin{aligned} L(\beta, D, \phi) &= \prod_{i=1}^N f_i(y_i|\beta, D, \phi) \\ &= \prod_{i=1}^N \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \beta, \phi) f(b_i|D) db_i; \end{aligned}$$

- Under the normal linear model, the integral can be worked out analytically
- In general, approximations are required:
  - Approximation of integrand
  - Approximation of data
  - Approximation of integral
- Predictions of random effects can be based on the posterior distribution

$$f(b_i | Y_i = y_i)$$

- Empirical Bayes (EB) estimate: the prior distribution is estimated from the data
  - Posterior mode, with unknown parameters replaced by their MLE

- For a given formula for how mean depends on the explanatory variables, the ML method must assume a particular type of probability distribution for  $Y$ , in order to determine the likelihood function
- By contrast, the quasi-likelihood approach assumes only a relationship between mean and  $\text{Var}(Y)$  rather than a specific probability distribution for  $Y$
- It allows for departures from the usual assumptions, such as overdispersion caused by correlated observations or unobserved explanatory variables
- To do this, the quasi-likelihood approach takes the usual variance formula but multiplies it by a constant that is itself estimated using the data.

For GEE

- Responses Correlated:  $g(E(Y)) = X\beta$
- Analysis describes differences in the mean of Y across the entire population
- Analysis informative from population perspective; most relevant from perspective of Policy makers
- Providers desiring to optimize outcomes across entire population
- GEE require 50-100 clusters as a fair number of clusters just to get the procedure to run
- QIC (Quasi-likelihood under the independence model criterion).
- CIC (correlation information criterion).

# Laplace Approximation of Integrand

- Numerical Integration
- Penalized quasi-likelihood (PQL)
- Marginal quasi-likelihood (MQL)
- Approximation of Integral
  - Quadrature is a historical mathematical term that means calculating area. Quadrature problems have served as one of the main sources of mathematical analysis
  - Gaussian quadrature
  - Adaptive Gaussian quadrature

## Example: Toenail Data

- $Y_{ij}$  is binary severity indicator for subject  $i$  at visit  $j$
- Model

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

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + b_0 i + \beta_1 T_i + \beta_2 t_{ij} + \beta_3 T_i t_{ij}$$

- Notation:
  - $T_i$ : treatment indicator for subject  $i$
  - $t_{ij}$ : time point at which  $j$ th measurement is taken for  $i$ th subject
- Adaptive as well as non-adaptive Gaussian quadrature can be used

# Generalized Linear Mixed Models in R

- The following table gives family generators and default links:

Family	Defult link	range of $y_i$	$V(Y_i \eta_i)$
Gaussian	identity	$(-\infty, +\infty)$	$\phi$
Binomial	logit	$\frac{0,1,\dots,n_i}{n_i}$	$\mu_i(1 - \mu_i)$
Poisson	log	$0, 1, 2, \dots$	$\mu_i$
Gamma	inverse	$(0, \infty)$	$\phi\mu^2$
Inverse.gaussian	$1/\mu^2$	$(0, \infty)$	$\phi\mu^3$

- For distributions in the exponential families, the variance is a function of the mean and a dispersion parameter  $\phi$  (fixed to 1 for the binomial and Poisson distributions).
- In the *lme4* package
  - `glmm()` : generalized-linear mixed-effects models using a normal mixing distribution computed by Gauss-Hermite integration

- The general formula for GLMM in R is defined as;

```
glmm(formula, family=gaussian, data=list(), weights=NULL,  
offset=NULL, nest, delta=1, maxiter=20, points=10,  
control=glm.control(epsilon=0.0001,maxit=10,trace=FALSE))
```
- *formula* A symbolic description of the model to be fitted.
- *family* A description of the error distribution and link function
- *data* A data frame containing the variables in the model
- *weights* An optional weight vector
- *offset* The known component in the linear predictor
- *nest* The variable classifying observations by the unit (cluster)
- *maxiter* The maximum number of iterations of the outer loop for numerical integration. *points* The number of points for Gauss-Hermite integration of the random effect.

# Required Packages

- Before you run the model, you need to install the required packages
  - installed.package(lme4)*

## R CODE

```
model.fit <- glmer(y ~ treatn + time + treatn*time + (1|idnum),
data = Toenail, family = binomial, control = glmerControl(optimizer
= "bobyqa"), nAGQ = 10)
print(model.fit, corr = TRUE)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermi
nAGQ = 10) [glmerMod]
Family: binomial ( logit )
Formula: y ~ treatn + time + treatn * time + (1 | idnum)
Data: Toenail
AIC      BIC      logLik  deviance df.resid
1257.8392 1285.6057 -623.9196 1247.8392      1902
Random effects:
Groups Name      Std.Dev.
idnum (Intercept) 4.073
Number of obs: 1907, groups: idnum, 294
Fixed Effects:
(Intercept)      treatn        time  treatn:time
-1.6518       -0.1201       -0.4065      -0.1601
```

- Model outputs

```
> summary(model.fit)
```

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.969	-0.189	-0.087	-0.007	38.398

Random effects:

Groups	Name	Variance	Std.Dev.
idnum	(Intercept)	16.59	4.073
Number of obs:	1907, groups:	idnum,	294

Fixed effects:

Estimate	Std. Error	z value	Pr(> z )			
(Intercept)	-1.65177	0.45037	-3.668 0.000245 ***			
treatn	-0.12005	0.59256	-0.203 0.839445			
time	-0.40653	0.04649	-8.744 < 2e-16 ***			
treatn:time	-0.16007	0.07235	-2.213 0.026931 *			
---						
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’	1

Correlation of Fixed Effects:

(Intr)	treatn	time
treatn	-0.633	
time	-0.138	0.212
treatn:time	0.202	-0.296
		-0.547

# Conclusion

- There is no effect of treatment at baseline
- Overtime, the effect of treatment was found to be significant

# Marginal Versus Random-effects Models

- 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.

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

- The following model is assumed for the mean structure:  
 $Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$ , for subject  $i$  and measurement  $j$ ,
- Exchangeable correlation (or CS)

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

$$\text{logit}(\pi_{ij}) = \xi_0 + \xi_1 A_{ij} + \xi_2 G_i + \xi_3 G_i A_{ij}$$

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

## GEE Model

Call:

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

Coefficients:

Estimate	Std.err	Wald Pr(> W )
----------	---------	---------------

(Intercept)	-1.86859	0.11185	279.12	<2e-16 ***
-------------	----------	---------	--------	------------

sex	0.13301	0.15395	0.75	0.39
-----	---------	---------	------	------

age	0.00139	0.01465	0.01	0.92
-----	---------	---------	------	------

sex:age	-0.01660	0.02127	0.61	0.44
---------	----------	---------	------	------

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale is fixed.

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

Estimate	Std.err
----------	---------

alpha	0.145	0.0163
-------	-------	--------

Number of clusters: 990 Maximum cluster size: 7

### R CODE:

```
fit21 <- geeglm(BMIBIN~ sex + age + sex*age, id = ind,
data = Infant, family = binomial, corstr = "exchangeable", scale.fix = TRUE)
```

- GEE can be modeled using different working correlation structure.
  - *ar1* for autoregressive order 1
  - *unstructured* for unstructured working correlation structure

### R code to update GEE

```
fit22 <- update(fit, corstr = "ar1")
fit23 <- update(fit2, corstr = "unstructured")
```

- The three models can be compared using QIC

```
model.sel(fit21, fit22, fit23, rank = QIC)
```

Model selection table

	(Int)	age	sex	age:sex	corstr	qLik	QIC	delta	weight
fit23	-1.85	1.62e-05	0.125	-0.0151	unstrc	-2429	4870	0.00	0.356
fit22	-1.86	-1.54e-04	0.136	-0.0151	ari	-2429	4870	0.11	0.336
fit21	-1.87	1.39e-03	0.133	-0.0166	exchng	-2429	4870	0.29	0.308

Abbreviations:

corstr: exchng = ‘exchangeable’, unstrc = ‘unstructured’

Models ranked by QIC(x)

```
> sapply(list(fit21, fit22, fit23), QIC)
```

QIC QIC QIC

4870 4870 4870

# Conclusion

- There is no effect of gender, age and gender and age interaction
- Based on QIC, there is no difference between the different working correlation structure

## GLMM: The Jimma Infant Data

- Random-effects model for non-Gaussian longitudinal 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.

$$\text{logit}(\pi_{ij}) = \xi_0 + \xi_1 A_{ij} + \xi_2 G_i + \xi_3 G_i A_{ij} + b_i$$

# GLMM: Jimma infant data

- GLMM with only random intercept was fitted

## R code for GLMM

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

## Odds ratio estimates ...

Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Qu  
nAGQ = 25) [glmerMod]

Family: binomial ( logit )

Formula: BMIBIN ~ sex + age + sex \* age + (1 | ind)

Data: Infant

AIC	BIC	logLik	deviance	df.resid
4645	4679	-2318	4635	6094

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.282	-0.362	-0.245	-0.237	2.898

Random effects:

Groups	Name	Variance	Std.Dev.
ind	(Intercept)	1.46	1.21

Number of obs: 6099, groups: ind, 990

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.33284	0.12666	-18.42	<2e-16 ***
sex	0.15899	0.16370	0.97	0.33
age	0.00194	0.01471	0.13	0.90
sex:age	-0.01990	0.02053	-0.97	0.33

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr)	sex	age
sex	-0.688	
age	-0.664	0.510
sex:age	0.479	-0.691

-0.717

## R CODE for odds ratio and 95% confidence Interval:

```
confint(fitGLMM, method="profile", oldNames=FALSE) # CI only  
2.5 % 97.5 %  
sd_(Intercept)|ind 1.0669 1.3624  
(Intercept) -2.5872 -2.0902  
sex -0.1617 0.4807  
age -0.0269 0.0308  
sex:age -0.0602 0.0203  
  
exp(confint(fitGLMM, method="profile", oldNames=FALSE)) # OR a  
2.5 % 97.5 %  
sd_(Intercept)|ind 2.9063 3.906  
(Intercept) 0.0752 0.124  
sex 0.8507 1.617  
age 0.9734 1.031  
sex:age 0.9416 1.021
```

- The odds ratio can be also computed in the same way

## R CODE for odds ratio:

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



- 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.,  $(b_{0i}, b_{1i}) \sim N(0, D)$  can be included to capture the correlation.

$$\text{logit}(\pi_{ij}) = \xi_0 + \xi_1 A_{ij} + \xi_2 G_i + \xi_3 G_i A_{ij} + b_{0i} + b_{1i} A_{ij}$$

# Random intercept and slope

R CODE for random intercept and slope model:

```
##### GLMM (random intercept & slope)
fitGLMMSlope <- glmer(BMIBIN ~ sex + age + sex*age + (1+age|ind),
data = Infant, family = binomial(link = "logit"))

summary(fitGLMMSlope)
```

- The model outputs are

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
Formula: BMIBIN ~ sex + age + sex * age + (1 + age | ind)
Data: Infant
AIC      BIC      logLik deviance df.resid
4569     4616     -2278      4555     6092
Scaled residuals:
Min      1Q Median      3Q      Max
-1.092 -0.305 -0.206 -0.198   2.966
Random effects:
Groups Name      Variance Std.Dev. Corr
ind    (Intercept) 3.6413   1.91
age       0.0577   0.24     -0.71
Number of obs: 6099, groups: ind, 990
Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.68488   0.18577  -14.45 <2e-16 ***
sex          0.17336   0.19908    0.87   0.38
age         -0.00181   0.02516   -0.07   0.94
sex:age     -0.02428   0.02774   -0.88   0.38
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation of Fixed Effects:
(Intr) sex     age
sex     -0.572
age     -0.769  0.439
sex:age  0.429 -0.757 -0.553
```

## Epilepsy Data

- Let  $Y_{ij}$  represent the number of epileptic seizures patient  $i$  experiences during week  $j$  of the follow-up period
- Let  $t_{ij}$  be the time-point (treatment week) at which  $Y_{ij}$  has been measured,  $t_{ij} = 1, 2, \dots$  until at most 27
- An indicator variable of treatment group the  $i^{th}$  subject receives is denoted by  $treat_i$  ( $0 = placebo$ ,  $1 = treated$ )
- The correlation in the data can be modeled by using 'exchangeable' correlation structure.
- Assuming that counts are generated from a Poisson-normal process with mean  $\lambda_{ij}$

$$\ln(\lambda_{ij}) = \xi_0 + \xi_1 treat_i + \xi_2 t_{ij} + \xi_3 treat_i t_{ij}$$

- GEE was fitted for the count data using the following R code

**R CODE:**

```
##### GEE
fit21 <- geeglm(BMIBIN~ sex + age + sex*age, id = ind,
data = Infant, family = binomial, corstr = "exchangeable", scale.fix = TRUE)
summary(fitgee1)
```

- The model can be updated for the different correlation structure

```
fit22 <- update(fit21, corstr = "ar1")
fit23 <- update(fit22, corstr = "unstructured")
```

- QIC can be computed for model comparison

```
model.sel(fit21, fit22, fit23, rank = QIC)
```

```
sapply(list(fit21, fit22, fit23), QIC)
```

```
Call:  
geeglm(formula = nseizw ~ trt + studywee + trt * studywee, family = poisson(link  
= log), data = epilepsy, id = id, corstr = "exchangeable",  
scale.fix = TRUE)  
Coefficients:  
Estimate Std.err Wald Pr(>|W|)  
(Intercept) 1.31567 0.18008 53.38 2.8e-13 ***  
trt 0.01704 0.29320 0.00 0.95  
studywee -0.01477 0.01684 0.77 0.38  
trt:studywee 0.00339 0.02015 0.03 0.87  
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
Scale is fixed.  
Correlation: Structure = exchangeable Link = identity  
Estimated Correlation Parameters:  
Estimate Std.err  
alpha 0.5 0.163  
Number of clusters: 89 Maximum cluster size: 27
```

## Model comparison

Model selection table

	(Int)	std	trt	std:trt	corstr	qLik	QIC	delta	weight
fitgee1	1.32	-0.01477	0.017	0.00339	exchng	7.23e+02	-1456	0.0	0.999
fitgee2	1.17	0.00388	0.137	-0.01657	ar1	7.20e+02	-1442	13.6	0.001
fitgee3	56.23	-3.65800	0.939	0.07399	unstrc	-1.16e+25	-1403	52.8	0.000

Abbreviations:

corstr: exchng = ‘exchangeable’, unstrc = ‘unstructured’

Models ranked by QIC(x)

```
> sapply(list(fitgee1, fitgee2, fitgee3), QIC)
  QIC   QIC   QIC
-1456 -1442 -1403
```

## Epilepsy Data

- Let  $Y_{ij}$  represent the number of epileptic seizures patient  $i$  experiences during week  $j$  of the follow-up period
- Let  $t_{ij}$  be the time-point (treatment week) at which  $Y_{ij}$  has been measured,  $t_{ij} = 1, 2, \dots$  until at most 27
- An indicator variable of treatment group the  $i^{th}$  subject receives is denoted by  $treat_i$  ( $0 = placebo$ ,  $1 = treated$ ).
- $b_i$  are subject specific random intercepts assumed to have Gaussian distribution with mean 0 and variance  $d$ .
- Assuming that counts are generated from a Poisson-normal process with mean  $\lambda_{ij}$

$$\ln(\lambda_{ij}) = \xi_0 + b_i + \xi_1 treat_i + \xi_2 t_{ij} + \xi_3 treat_i t_{ij}$$

# Random Intercept Mixed Effect Poisson

- The model with random intercept can be fitted by the following R code

```
##### GLMM (random intercept)
fitGLMMInt <- glmer(nseizw ~ trt + studywee + trt*studywee + (1|id),
data = epilepsy, family = poisson(link = "log"))

summary(fitGLMMInt)
```

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [']  
Family: poisson ( log )  
Formula: nseizw ~ trt + studywee + trt \* studywee + (1 | id)  
Data: epilepsy  
AIC BIC logLik deviance df.resid  
6282.5 6308.7 -3136.2 6272.5 1414  
Scaled residuals:  
Min 1Q Median 3Q Max  
-4.6148 -0.8545 -0.4130 0.5323 14.9969  
Random effects:  
Groups Name Variance Std.Dev.  
id (Intercept) 1.154 1.074  
Number of obs: 1419, groups: id, 89  
  
Fixed effects:  
Estimate Std. Error z value Pr(>|z|)  
(Intercept) 0.817901 0.167317 4.888 1.02e-06 \*\*\*  
trt -0.170354 0.238204 -0.715 0.47451  
studywee -0.014288 0.004385 -3.258 0.00112 \*\*  
trt:studywee 0.002289 0.006140 0.373 0.70929  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
Correlation of Fixed Effects:  
(Intr) trt studyw  
trt -0.702  
studywee -0.209 0.146  
trt:studywe 0.149 -0.211 -0.714

- Include a random slope assuming subjects have different evolution over time.
- Both  $b_{i1}$  and  $b_{i2}$  are jointly normally distributed and possibly correlated.
- The variance-covariance matrix can then be ‘unstructured’.

$$\ln(\lambda_{ij}) = \xi_0 + b_{i1} + \xi_1 treat_i + \xi_2 t_{ij} + \xi_3 treat_i t_{ij} + b_{i2} t_{ij}$$

# Random Intercept and Slope GLMM

- The R code for random intercept and Slope model

```
##### GLMM (random intercept & slope)
fitGLMMSlope <- glmer(nseizw~ trt + studywee + trt*studywee + (1+studywee|id),
data = epilepsy, family = poisson(link = "log"))

summary(fitGLMMSlope)

      AIC      BIC    logLik deviance df.resid
6081.2   6118.0  -3033.6    6067.2     1412
```

## Model Comparison

- We compare the two models using AIC
  - 6282.5 versus 6081.2
- Log-likelihood can also be used
  - -3136.2 versus -3033.6
- In both cases the model with random intercept and random slope is better

## The Gilgel-Gibe Mosquito Data

- Let  $Y_{ij}$  represent the number of An. gambaie counts in house  $i$ .
- Let  $t_{ij}$  be the time-point (in months) at which  $Y_{ij}$  has been measured,  $t_{ij} = 1, 2, \dots$  until at most 32.
- An indicator variable of village group the  $i^{th}$  house belongs is denoted by  $village_i$  ( $0 = control$ ,  $1 = atrisk$ ).
- An indicator variable of season type a specific month belongs is denoted by  $season_{ij}$  ( $0 = dry$ ,  $1 = wet$ ).
- $b_i$  are subject specific random intercepts assumed to have Gaussian distribution with mean 0 and variance  $d$ .
- Assuming that counts are generated from a Poisson-normal process with mean  $\lambda_{ij}$

$$\ln(\lambda_{ij}) = \xi_0 + b_i + \xi_1 village_i + \xi_2 season_{ij} + \xi_3 t_{ij} + \xi_5 village_i t_{ij}.$$

# Missing Data

- When applying multilevel analysis to longitudinal data, there is no need to have a complete dataset, and, furthermore, it has been shown that multilevel analysis is very flexible in handling missing data.
- It has even been shown that applying multilevel analysis to an incomplete dataset is even better than applying imputation methods (Applied Multilevel Analysis)

# Missing Mechanisms

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MANR)

# Handling Mechanisms

- Complete case analysis
- Available Case Analysis
- Last observed carried forward
- Imputation
  - Mean imputation
  - Multiple imputation

# Conclusions

- For correlated data, assuming independence my result biased result
- The dependence between observations can be accounted by fitting
  - Linear Mixed Effect Model
  - Generalized Estimating Equation
  - Generalized linear mixed effect model (GLMM)

## Intra-Class Correlation

- The ratio of the between cluster variance to the total variance is called ICC
- It tells us the proportion of the total variance in the response variable that is accounted for by the cluster
- It can also be interpreted as the correlation among observations within the same cluster

$$\text{cov}(Y_{ij}, Y_{ij'}) = u_i = \sigma_0^2$$

- $u_i \sim iidN(0, \sigma_0^2)$  for subject  $i$  and
- $\epsilon_{ij} \sim iidN(0, \sigma^2)$  for outcome  $j$

$$\text{corr}(Y_{ij}, Y_{ij'}) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}$$

## Intra-Class Correlation

- It can help to determine whether or not a mixed model is even necessary
- If the correlation is zero that means the observation within cluster are no more similar than the observations from different cluster
- It can be theoretically meaningful to understand how much of the overall variation in the response is explained by clustering
- The choice  $icc=0$  is obvious, but is rarely zero
- As a rule of thumb, it seems to recall to use 0.1
- After fitting conditional models (mixed effect), use the following stata command

```
estat icc
```

# Project

- Consider EDHS2016 data and outcome variable
  - Outcome variable 1: ANC visit
  - Outcome variable 2: Number of children per mother
  - Outcome variable 3: Place of delivery
- Independent variable: Sex, age, residence, income, religion, ...
- Fit GEE for the data
- Interpret the result

**The End**