Statistical Modeling Using Stata

Tadesse Awoke Ayele (PhD, Associate Professor)

University of Gonder

Collage of Medicine and Health Sciences

Institute of Public Health

May 26, 2018

Contact Detail

- Instructor: Tadesse Awoke (PhD, Associate Professor)
- Department: Epidemiology and Biostatistics
- Qualification 1: Biostatistics
- Qualification 2: Public Health
- Email: tawoke7@gmail.com
- Phone: +251910173308
- Time: May 23-May 30, 2018
- Location: University of Gondar
- Office hours: Available upon request

Course Description

- This course is designed to give students an overview of Continuous Data Analysis, Categorical Data Analysis, Count data analysis, Survival Data Analysis, Longitudinal Data Analysis for Gaussian data, Longitudinal Data Analysis for non-Gaussian data, and Time Series Analysis.

- The overall emphasis will be placed on understanding the language of statistics and the art of statistical investigation.

Course Objectives

- The aim of biostatistical study is to provide the numbers that contain information about certain situations and present them in such a way that valid interpretations are possible

Specific objectives

- Describe Continuous variable and method of analyses
- Interpret outputs from t-test ANOVA, and linear regression
- Explain Categorical data and methods of analyses
- Interpret outputs from logistic regression, conditional logistic regression, and multinomial logistic regression
- Describe time-to-event data and methods of analysis
- Explain the outputs from survival analysis
- Describe Longitudinal data and methods of analyses
- Interpret the outputs from fixed and random effect models
- Conclude based on the information obtained from the models

Course Outline

Chapter 1: General Introduction

- Research and Biostatistics
- Variable, Data, Information, knowledge, and wisdom
- Concept of Biostatistics
- Introduction to STATA

Generalized Linear Models (GLMs)

Chapter 2: Analysis of Continuous Variable

- Comparison of means
  - Student t-test
  - Analysis of Variance (ANOVA)
  - Analysis of Covariance (ANCOVA)
  - Multivariate Analysis of Variance (MANOVA)
  - Multivariate Analysis of Covariance (MANCOVA)
  - Repeated Measures ANOVA
- Correlation and Linear Regression

Chapter 3: Analysis of Categorical Data

- Categorical Data Analysis
    - Logistic Regression
    - Conditional Logistic Regression
    - Ordinal Logistic Regression
    - Multinomial Logistic Regression

Chapter 4: Models for Count data

- Poisson Regression
- Negative Binomial Poisson Regression
- Zero inflated Poisson Regression
- Zero-Inflated Negative Binomial Regression

Chapter 5: Survival Data analysis

- Time to event data
- Life Table
- Kaplan-Meier Survival curve
- Cox-proportional hazards regression
- Proportional Hazard Assumption
- Time-varying Covariate
- Parametric Cox-regression model

Chapter 6:Longitudinal analysis for Gaussian Data

- Longitudinal Data Structure and Examples
  - Exploratory Data Analysis
    - Individual Profiles
    - Average Profile
    - Correlation Matrix
  - Linear Mixed Effect Model
    - Fixed Effect Models
    - Random Effect Models
  - Model Comparison

Chapter 7: Longitudinal Analysis for non-Gaussian data

- Longitudinal Data structure and Examples
- Models for correlated Data
  - Generalized Estimating Equation
  - Generalized Linear Mixed Effect model
- Frailty Models
- Model Comparison
- Intra class correlation

References

- Martin Bland. An introduction to Medical Statistics
- Colton T. Statistics in Medicine
- Daniel W. Biostatistics a foundation for analysis in the Health Sciences
- Kirkwood BR. Essentials of Medical Statistics
- Knapp RG, Miller MC. Clinical epidemiology and Biostatistics. Baltimore Williams and Wilkins, 1992
- P. Armitage and G. Berry. Statistical Methods in Medical Research
- Pagano and Gauvereau. Principles of Biostatistics

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
Part I: Basic Models	
Models for Continuous data	Day 1-2
Models for Categorical data	Day 2-3
Models for count data	Day 3-4
Survival Analysis	Day 3-4
Part II: Advanced Models	
Longitudinal Analysis for Continuous Data	Day 5-6
Longitudinal Analysis for categorical Data	Day 6-7
Frailty Models	Day 7

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
- 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 in to different types

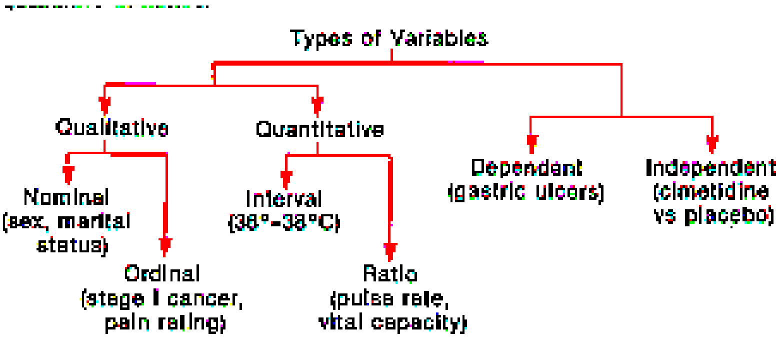

Types of Variables
Qualitative
Quantitative
Nominal (sex, marital status)
Ordinal (stage I cancer, pain rating)
Interval (36°-38°C)
Ratio (pulse rate, vital capacity)
Dependent (gastric ulcers)
Independent (cimetidine vs placebo)

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.

  1. Identify the outcome variable
  2. Identify the predictor variable

Solution

- Outcome variable

- Predictor variable

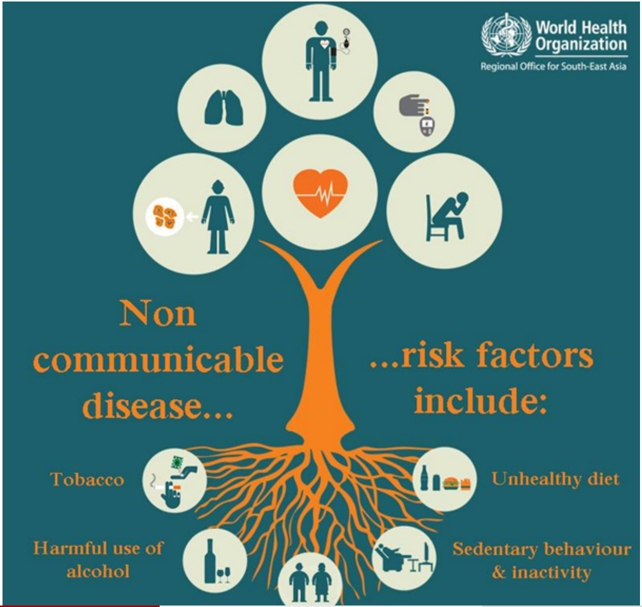
Example 2

- 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


World Health Organization
Regional Office for South-East Asia
Non communicable disease...
...risk factors include:
Tobacco
Unhealthy diet
Harmful use of alcohol
Sedentary behaviour & inactivity

Speed and risk of car accident

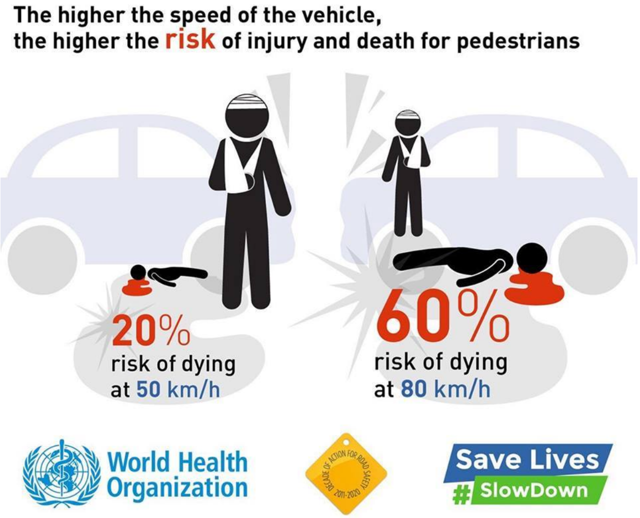

The higher the speed of the vehicle,
the higher the risk of injury and death for pedestrians
20%
risk of dying
at 50 km/h
60%
risk of dying
at 80 km/h
World Health Organization
DECADE OF ACTION FOR ROAD SAFETY 2011-2020
Save Lives
#SlowDown

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

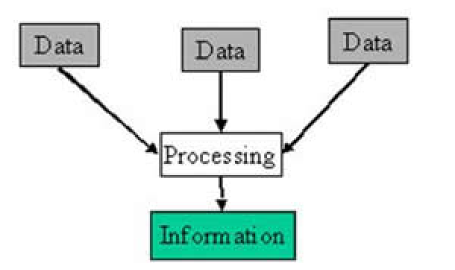


Figure 4: Relationship between data and information

- 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

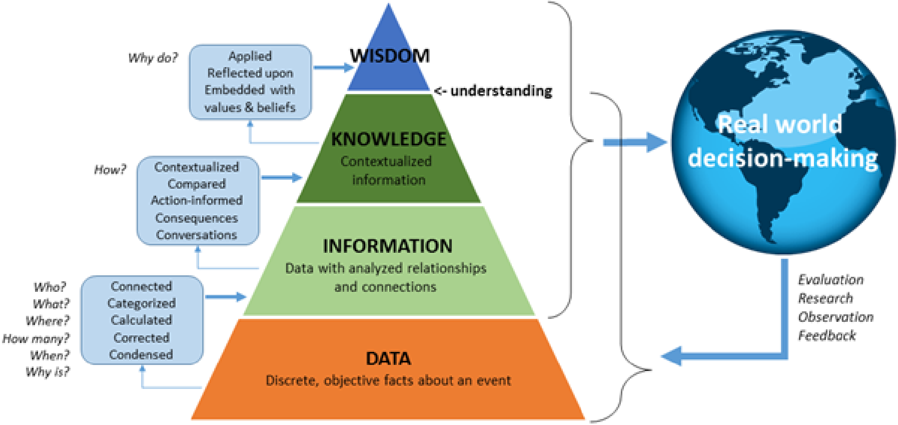


Figure 5: Schematic presentation

Example 1: Jimma Child Mortality Data

The data were collected to established 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	StillB	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 in to 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.........

- 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
.	.	.	.	.
25	3	0	3015	1
26	3	2	3200	0
.	.	.	.	.

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

- 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 Vlge Year Month Time Hous Gamb Season
1 at risk 1 1 0 11 0 wet
1 at risk 1 2 1 11 0 wet
1 at risk 1 3 2 11 9 wet
1 at risk 1 4 3 11 30 wet
.    .       .   .  .    .  .    .
2 at risk 1 1 0 12 0 dry
2 at risk 1 2 1 12 0 dry
2 at risk 1 3 2 12 3 dry
2 at risk 1 4 3 12 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

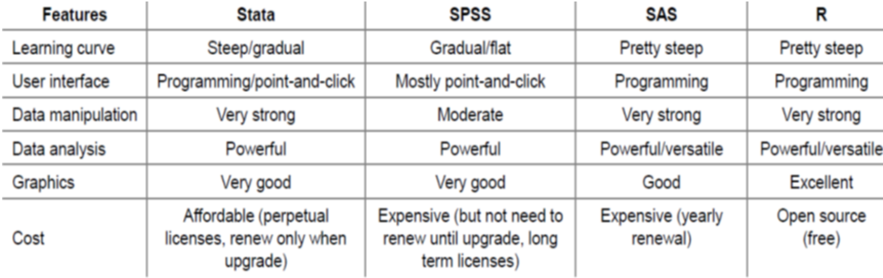
Introduction to Stata

Introduction

- It is a multi-purpose statistical package to help you explore, summarize and analyze datasets
- A dataset is a collection of several pieces of information called variables (usually arranged by columns)
- A variable can have one or several values (information for one or several cases)
- Other statistical packages are SPSS, SAS and R
- Stata is widely used in social science research and the most used statistical software on campus.

Data Management Softwares

- There are different possible data management softwares
- The most common are Stata, SPSS, SAS and R
- They all have different features



Features	Stata	SPSS	SAS	R
Learning curve	Steep/gradual	Gradual/flat	Pretty steep	Pretty steep
User interface	Programming/point-and-click	Mostly point-and-click	Programming	Programming
Data manipulation	Very strong	Moderate	Very strong	Very strong
Data analysis	Powerful	Powerful	Powerful/versatile	Powerful/versatile
Graphics	Very good	Very good	Good	Excellent
Cost	Affordable (perpetual licenses, renew only when upgrade)	Expensive (but not need to renew until upgrade, long term licenses)	Expensive (yearly renewal)	Open source (free)

- For this course, we will focus on Stata

Introduction to Stata

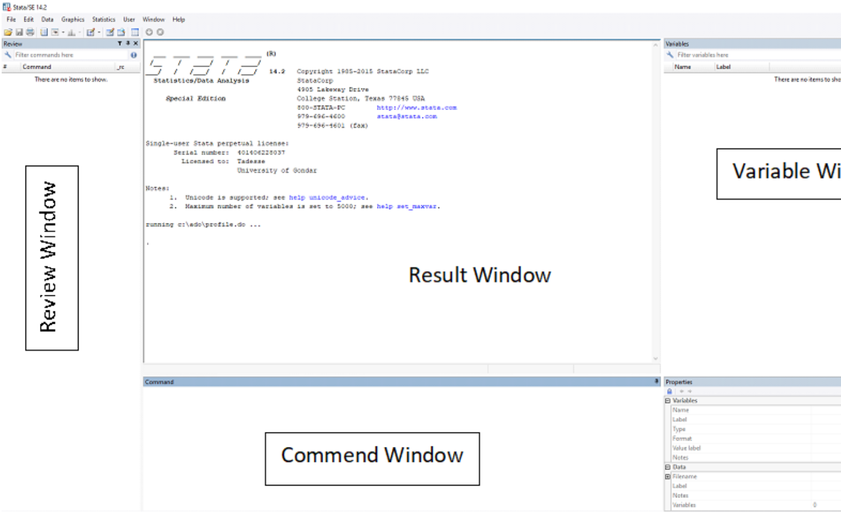
- Stata is a complete, integrated statistical package that provides everything you need for data analysis, data management, and graphics
- It is fast and easy to use
- Current version is Stata 14??
- Standard stata can handle up to 2047 variables
- SE can handle 32766 variables
- Number of observations is limited by your computer (up to 2 billion!)

- Stata has 4 windows:
  - Command: where commands are entered. All commands and variables are case sensitive.
  - Results: where results appear
  - Review: where past commands are listed
  - Clicking a past command in Review window brings it to the command window where it can be modified and re-executed Variable: where

    variables in current dataset are listed

Stata Windows

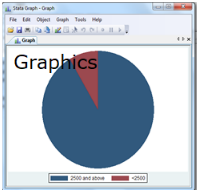
- The four windows of Stata are given here

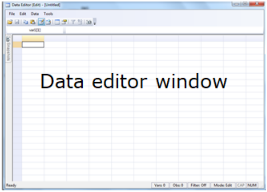


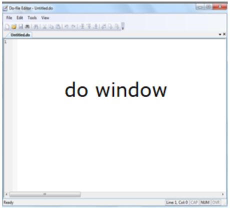
```
                        14.2   Copyright 1985-2015 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  401406228037
         Licensed to:  Tadesse
                       University of Gondar

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

running c:\ado\profile.do ...

.
```

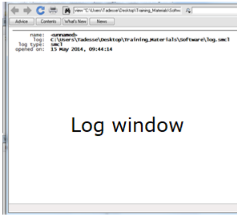
Other windows

- Graphics
- Data editor
- Data viewer
- ‘do’
- Log









```
      name:  <unnamed>
       log:  C:\Users\Tadesse\Desktop\Training_Materials\Software\log.smcl
  log type:  smcl
 opened on:  15 May 2014, 09:44:14
```

Important details

- It is case sensitive!
- First set the memory of Stata: set memory 100m
- You can only have one dataset active at a time
- abbreviations work for commands and variable names!
- Help is very important part
- Two interactive options:
    - help ‘command’
    - help ‘search’

- There are lots of things you can do without data in stata!

```
display 4.1-1.96*0.3       compute the difference

tabi 100 34 \ 17 294       contegency table
tabi 100 34 \ 17 294, col
tabi 100 34 \ 17 294, col row cell chi

cci 100 34 17 294          confidence interval
cci 100 34 17 294, exact

sampsi 0.2 0.5        sample size for two population
```

Inputting Data

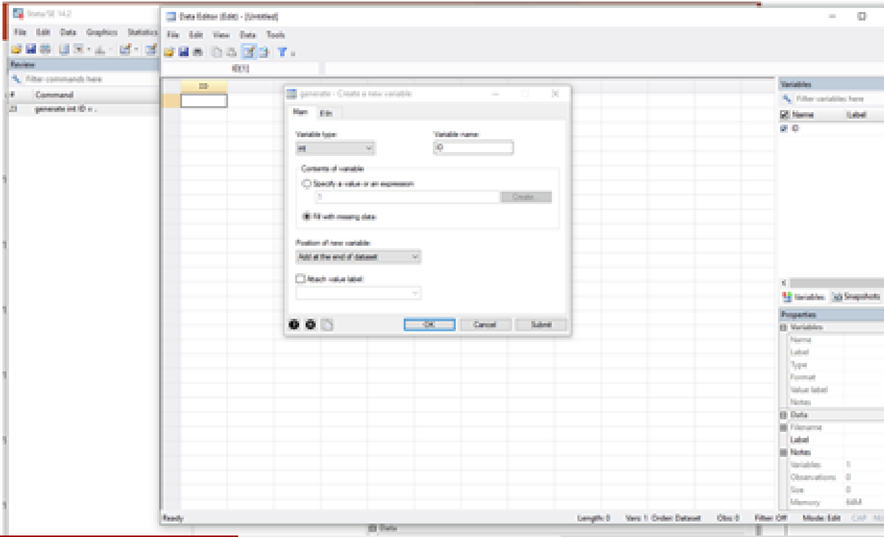
- Many Options:
  - Manually enter data into the Stata Data Editor
  - Copy data into the Data Editor from another source (ex.: Excel)
  - Importing an ASCII (text) file.
  - Reading in an Excel spreadsheet (tab- or comma- delimited text file).
  - Open existing Stata Data file with file extension .dta
  - Use a conversion package (StatTransfer) to read in data from another package (eg, SPSS, SAS data, . . .).

Manual Data entry

- First open the data editor window to enter data manually
- Second define variable names:
  - Note that variables are automatically named var1, var2, . . .
  - Double-click on top of column to view/edit “Variable Properties” and change the name
  - Via command: rename oldvarname newvarname
    - Eg. rename var1 id
- Third write the observation for each variables.
- To move to the second observation, click ”enter”

Data entry

- Variables can be defined in this window



generate int ID = .

generate - Create a new variable

Variable type: int

Variable name: ID

Contents of variable

○ Specify a value or an expression

● Fill with missing data

Position of new variable: Add at the end of dataset

□ Attach value label

Importing Data

- Import: Via drop- menu: File → Import → ASCII data created by a spreadsheet.
- Via drop-menu: File → Open →Scroll to find data
- The clear command is a default command that clears the memory before loading the requested datafile.
- This is necessary because Stata can have only one dataset in memory at a time!

Interactive command line

Do file

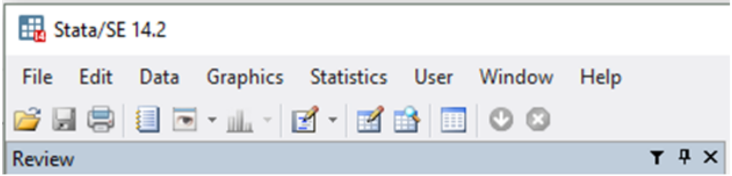
- The commands used to produce the output can be saved in the "do file"
- It can be also executed by highlighting each command line
- Any time we want to run the command, we can open the do file

Log file

- Log file is used to save the outputs produced
- This can bone by clicking begin under Log menu
- File → Log → Begin
- If you want to stop saving the output, then suspend using
- File → Log → Suspend (or End)

Data Management and Analysis

- Stata works in two ways;
  1. Menu driven



  2. Command driven

- The later was used in this material

Data Management

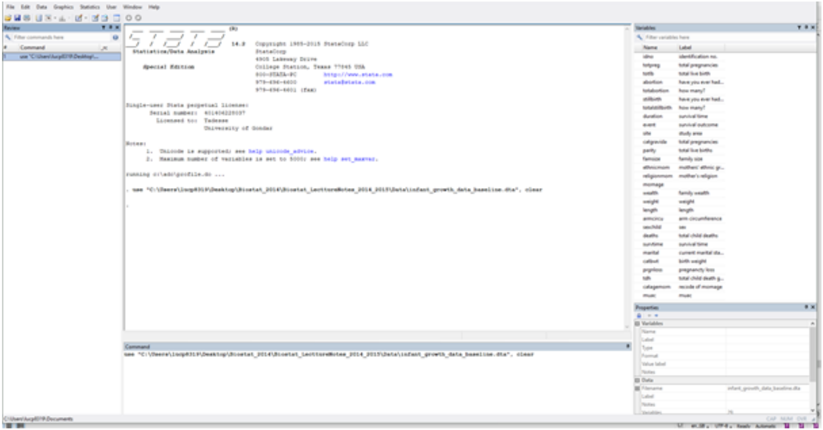
- Once the data were entered or imported, data management will be the next step
- Data should be cleaned, edited, and recoded before analysis started
- cleaning can be done through visual inspection, running frequencies and cross-tabulation
- When error is detected, can be corrected by checking the hard copy and the data collector if necessary
- Some variables may need recoding in order to make clear interpretation and when there is assumption violation

Data Management

- Basic commands for generating new variables or recoding existing variables
  - gen: to generate new variable from existing variable
  - replace: to replace the remaining categories of the new variable
  - recode: to recode continuous variable to categorical
  - egen: to create a new variable that counts the number of "yes" responses
  - merge: variables from a second dataset to the dataset you're currently working with

Data Management

- list, describe, keep, drop, rename and so on.



```
1  use "C:\Users\lucp8319\Desktop\...
```

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   14.2   Copyright 1985-2015 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  401406228037
         Licensed to:  Tadesse
                       University of Gondar

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

running c:\ado\profile.do ...

. use "C:\Users\lucp8319\Desktop\Biostat_2014\Biostat_LectureNotes_2014_2015\Data\infant_growth_data_baseline.dta", clear

.
```

Command

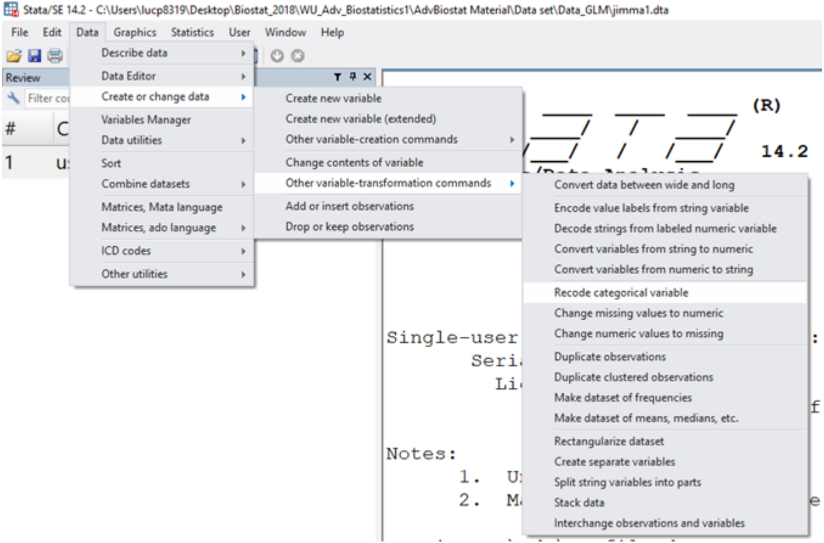
```
use "C:\Users\lucp8319\Desktop\Biostat_2014\Biostat_LectureNotes_2014_2015\Data\infant_growth_data_baseline.dta", clear
```

Name	Label
idno	identification no.
totpreg	total pregnancies
totlb	total live birth
abortion	have you ever had...
totabortion	how many?
stillbirth	have you ever had...
totalstillbirth	how many?
duration	survival time
event	survival outcome
site	study area
catgravida	total pregnancies
parity	total live births
famsize	family size
ethnicmom	mothers' ethnic gr...
religionmom	mother's religion
momage	
wealth	family wealth
weight	weight
length	length
armcircu	arm circumference
sexchild	sex
deaths	total child deaths
survtime	survival time
marital	current marital sta...
catbwt	birth weight
prgnloss	pregnancy loss
tdh	total child death g...
catagemom	recode of momage
muac	muac

Filename: infant_growth_data_baseline.dta

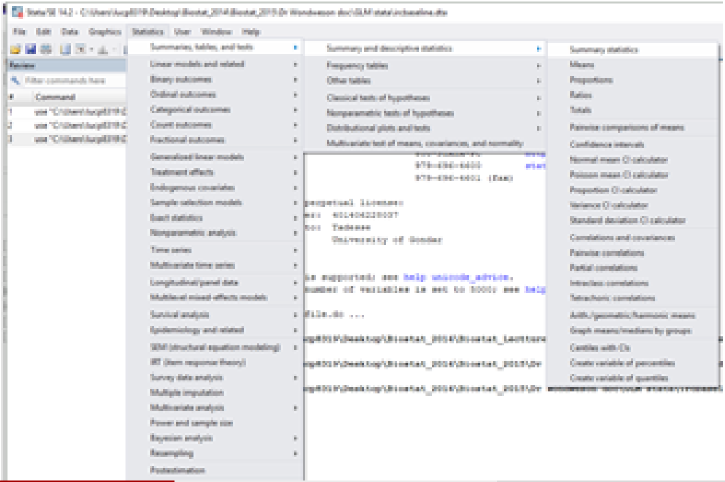
Data Management

- Data management using menu driven


Stata/SE 14.2 - C:\Users\lucp8319\Desktop\Biostat_2018\WU_Adv_Biostatistics1\AdvBiostat Material\Data set\Data_GLM\jimma1.dta
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Filter co
#
Describe data
Data Editor
Create or change data
Variables Manager
Data utilities
Sort
Combine datasets
Matrices, Mata language
Matrices, ado language
ICD codes
Other utilities
Create new variable
Create new variable (extended)
Other variable-creation commands
Change contents of variable
Other variable-transformation commands
Add or insert observations
Drop or keep observations
Convert data between wide and long
Encode value labels from string variable
Decode strings from labeled numeric variable
Convert variables from string to numeric
Convert variables from numeric to string
Recode categorical variable
Change missing values to numeric
Change numeric values to missing
Duplicate observations
Duplicate clustered observations
Make dataset of frequencies
Make dataset of means, medians, etc.
Rectangularize dataset
Create separate variables
Split string variables into parts
Stack data
Interchange observations and variables
(R)
14.2
Single-user
Notes:

Data Analysis

- There are two basic analysis procedures
  - Descriptive versus Analytic
- In stata, these can be done either
  - Menu driven versus command driven



```
                    979-696-4600
                    979-696-4601 (fax)

perpetual license:
er:  401406228037
to:  Tadesse
     University of Gondar

is supported; see help unicode_advice.
number of variables is set to 5000; see help

file.do ...
```

Part A

Generalized Linear Models

Chapter 2

Analysis of Continuous Variable

Introduction

- We may be interested to compare the means of two population
- t-test is appropriate to compare two means from two population
- This test was invented by a statistician working for the brewer Guinness.
- He was called Mr Gosset (1867-1937), but preferred to keep anonymous so wrote under the name Student
- Hence we have Student t test, in memory of Mr Gosset.

There are three different t-tests

- One sample t-test
- Two independent sample t-test
- Paired sample t-test

Assumptions

- The outcome variable is measured in continuous or ordinal scale
- Data is collected from a representative, randomly selected portion of the total population
- The outcome variable follows normal distribution
- Reasonably large sample size is used
- Final assumption is homogeneity of variance

- Example: Consider Jimma infant survival data introduced in chapter 1. Here we want to test if there is significant difference in birth weight of the infant among the sex group of the new born.
- Independent t-test is appropriate

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
Group | Obs     Mean     Std. Err.    Std. Dev.    [95% Conf. Interval]
---------+--------------------------------------------------------------------
0 |   3,861    3055.323    8.099562     503.282    3039.443    3071.203
1 |   4,012    3169.724    8.238749    521.8453    3153.572    3185.877
---------+--------------------------------------------------------------------
combined |7,873  3113.62  5.815093    515.9728    3102.221  3125.02
---------+--------------------------------------------------------------------
diff |          -114.4014    11.55335      -137.049   -91.75373
------------------------------------------------------------------------------
diff = mean(0) - mean(1)                                   t =  -9.9020
Ho: diff = 0           Satterthwaite's degrees of freedom =  7870.96

Ha: diff < 0                  Ha: diff != 0                  Ha: diff > 0
Pr(T < t) = 0.0000     Pr(|T| > |t|) = 0.0000         Pr(T > t) = 1.0000
```

STATA CODE:

```
ttest weight, by(sexchild) unequal
```

- Interpretation
  - The null hypothesis is stated as

$$H_0 : \mu_g - \mu_b = 0$$

  versus

$$H_a : \mu_g - \mu_b \neq 0$$

  - The 95% confidence interval for the difference of means does not contain 0.
  - The p-value is less than 0.05
  - Hence, we conclude that there is significant difference in birth weight among the sex groups.

Comparison of more than 2 groups

- For two normal distributions the two sample means are compared by t-test.
- The means of more than two distributions need to be compared.
- The t-test methodology generalizes to the one-way analysis of variance (ANOVA).
- ANOVA do not tell you which group is different, but only whether a difference exists.
- Consider Jimma infant data
  - Outcome variable: birth weight.
  - Factor variable: residence (urban$=1$, semi-urban$=2$, rural$=3$).
  - Objective: compare weight among the three place categories.

Jimma Infant data:

- For K means (K$\geqslant$ 3).
  - $H_o : \mu_1 = \mu_2 = \ldots = \mu_k$,
  - H_A : at least one of the means is different.
- There is one factor of grouping.
- The variable of interest is the outcome.
- One factor with K groups.
- Notations:
  - x_{ij} is the j^{th} observation in the i^{th} groups.
  - $\bar{x_i}$ is the mean of the i^{th} group.
  - $\bar{x}$ is the grand mean (mean of means).

- Assumptions;
    - The outcome is normally distributed.
    - Population variance is assumed constant among the groups.
    - Independent random samples among the groups.
- One-way ANOVA requires calculation of the following:
    - Between-groups sum of squares $SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$.
    - Within-groups sum of squares $SSW = \sum_{i=1}^{k}(n_i - 1){s_i}^2$.
    - Where ${s_i}^2 = \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{n_i - 1}$.
    - Where ${s_i}^2$ is the sample variance of the i^{th} group.

- Between-groups degrees of freedom $k-1$.
- Within-groups degrees of freedom $n-k$, where $n = n_1 + n_2 + \ldots + n_k$.
- Between-groups mean square $(BMS = \frac{SSB}{k-1})$.
- Within-groups mean square $(WMS = \frac{SSW}{n-k})$.
- The F-statistic is used to test the hypothesis is $F_{cal} = \frac{BMS}{WMS}$.
- $F_{cal} \sim F_{dist}$ with $(k-1, n-k)$ degrees of freedom.
- if $F_{cal} > F_{(1-\alpha)}(k-1, n-k)$ or $p_{value} < \alpha$, then reject H_o.

- Almost all softwares display results in ANOVA table:

SV	SS	df	MS	F_{cal}	pvalue
Between	SSB	$k-1$	MSB	$\frac{MSB}{MSW}$	$p(F_{ta} < F_{cal})$
Within	SSW	$n-k$	MSW		
Total	$SSB+SSW$	$n-1$			

- The variable of interest is the weight.
- There is one factor of grouping (residence): k=3.
- Hypothesis:
  - $H_o : \mu_u = \mu_{su} = \mu_r$,
  - H_A : at least one of the means is different.
- Outputs are ...

```
Analysis of Variance
Source           SS      df     MS         F     Prob > F
---------------------------------------------------------------
Between groups   124172273    2   62086136.5  247.83  0.0000
Within groups  1.9716e+09  7870  250517.627
---------------------------------------------------------------
Total           2.0957e+09   7872   266227.896

Bartlett's test for equal variances: chi2(2)=0.3509 Prob>chi2=0.839
```

STATA CODE:

```
oneway weight place
```

Conclusion:

- We reject the null hypothesis ($p_{value} < 0.05$).
- ... and conclude that at least one of the groups' means differ on body weight.
- Now the question is: which groups are different?
- Answering this question requires multiple comparisons.
- We can use the Bonferroni method.
- Bonferroni method corrects probability of Type I error for the number of tests.

```
Comparison of weight by study area
(Bonferroni)
Row Mean-|
Col Mean |          1          2
---------+----------------------
2 |           -46.0835
|                0.015
|
3 |           -272.528  -226.445
|                0.000     0.000
```

- Interpretation;
- All pairs of comparison are statistically significant at 0.05 level:
    - urban versus semi-urban,
    - urban versus rural,
    - semi-urban versus rural.

STATA CODE:

```
oneway weight place, bonferroni
```

ANCOVA

- ANOVA with continuous covariates as independent variable

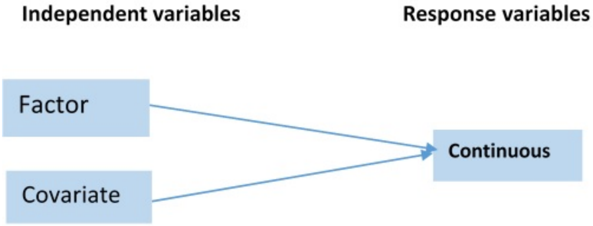


Figure 6: Analysis of covariance structure.

- Example: Consider the dependent variable WBC and independent variable BMI and hgb:

```
mod.11 <- lm(WBC ~ as.factor(BMI1_1)+ Hgb, data=VL)
ancova2 <-anova(mod.11)
ancova2
```

```
> ancova2
Analysis of Variance Table

Response: WBC
Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(BMI1_1)  2   3.227  1.6136  0.3568 0.7028
Hgb                1   3.257  3.2574  0.7203 0.4026
Residuals         31 140.197  4.5225
```

MANOVA

- MANOVA is an ANOVA with two or more continuous response variables

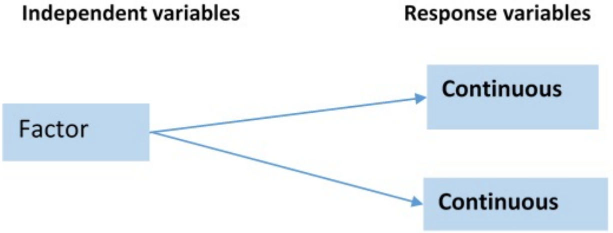


Figure 7: Multivariate Analysis of variance structure.

R CODE:

```
Manova <- manova(cbind(WBC, Hgb, Platelet)~BMI1,data=VL)
Manova
summary(Manova)
```

```
> summary(Manova)
Df   Pillai approx F num Df den Df Pr(>F)
BMI1       1 0.072405  0.78056      3     30 0.5141
Residuals 32
```

MANCOVA

- MANCOVA is when one or more covariates are added to the MANOVA.

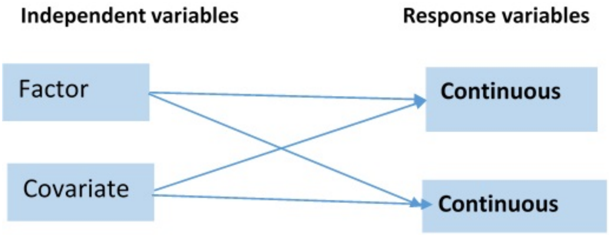
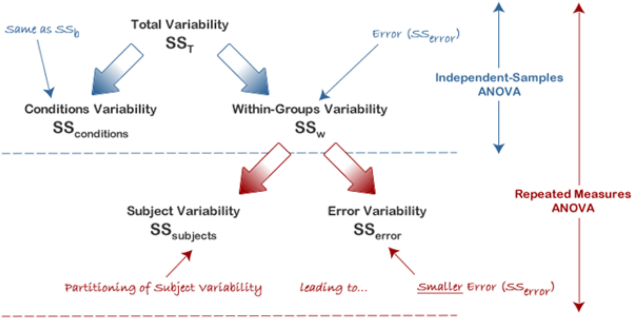


Figure 8: Multivariate Analysis of covariance structure.

Repeated Measures ANOVA

- Repeated measures ANOVA is equivalent to one-way ANOVA, but for related groups, and is the extension of the dependent t-test.
- Also referred to as ANOVA for correlated samples



- Hypothesis:
  - $H_o : \mu_1 = \mu_2 = \mu_3, \ldots \mu_k,$
  - H_A : at least one of the means is different.

Data Structure

- Exercises intervention on weight loss. The weight of study participants were measured at pre, after 3 months and after 6 months.

Subject	pre	3 months	6 months	Sub. Mean
1	60	61	55	58.7
2	66	65	60	63.7
3	71	70	59	66.7
4	62	56	51	56.3
5	59	59	62	60.0
6	80	70	68	72.7
.	.	.	.	.
.	.	.	.	.
Monthly mean	.	.	.	.

- Repeated Measures ANOVA table:

SV	SS	df	MS	F_{cal}	pvalue
Condition	SS_c	$k-1$	MS_c	$\frac{MS_c}{MS_e}$	$p(F_{ta} < F_{cal})$
Subject	SS_s	$n-1$	MS_s	$\frac{MS_s}{MS_e}$	
Error	SS_e	$(k-1)(n-1)$	MS_e		
Total	SS_T	$N-1$			

- The F-statistic found on the first row is the F-statistic that will determine whether there was a significant difference between at least two means or not.

Output

```
Number of obs =         30    R-squared     =  0.8335
Root MSE      =    3.18736    Adj R-squared =  0.7318

Source | Partial SS          df         MS        F    Prob>F
-----------+-----------------------------------------------
Model |  915.43333          11   83.221212      8.19  0.0001
|
time | 111.8                2         55.9     5.50  0.0137
subject |803.63333       9   89.292593      8.79  0.0001
|
Residual |  182.86667  18   10.159259
-----------+-----------------------------------------------
Total |     1098.3          29   37.872414
```

Conclusion

```
. anova wt time subject, repeated(subject) grouping(time)
```

- The p-value corresponding to the grouping variable is used to make decision
- Based on the p-value, the null hypothesis is rejected and
- At least one mean is significant different from the other

Correlation and Linear Regressions

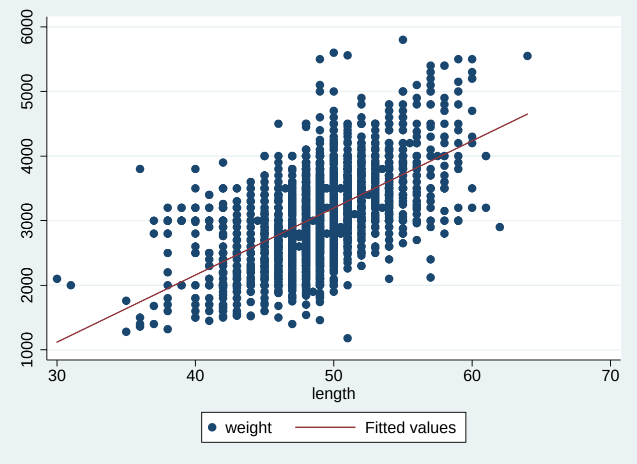
Correlation

- As OR and RR are used to quantify risk between two dichotomous variables, correlation is used to quantify the degree to which two random continuous variables are related, provided the relationship is linear.
- If we are interested in looking at relationship between two continuous random variables, before we conduct any type of analysis we should always create a two way scatter plot for our visual understanding of the relationship
- We now consider ways of describing relationship between numeric variables from a single population.
- consider the relationship between weight and height of children from jimma data.

The correlation coefficient

- Before we conduct any statistical analysis, we should always create a scatter plot of the data.
- These data could be presented in a scatter plot as shown in Figure 1 below.
- One can see from the scatter plot that the mortality rate tends to decrease as the percentage of children immunized increases.
- The correlation coefficient may be used to measure the degree of the association between the two variables.
- Alternatively if we are interested in predicting weight from height the of children we could fit a linear regression model (a straight line) to the data.

Scatter plot showing the above table



- Another way of looking at correlation is that it is a measure of the scatter of the points around the underlying linear trend (the greater the spread of points the lower the correlation

STATA code for scatter plot and lines:

```
twoway (scatter weight length) (lfit weight length)
```

Correlation coefficient

- In the underlying population from which the sample of points (x_i, y_i) is selected, the correlation coefficient between X and Y is denoted by ρ and is given by

$$\rho = \frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}$$

- It can be thought as the average of the product of the standard normal deviates of X and Y

Estimation of Correlation coefficient

- The estimator of the population correlation coefficient is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- r is called the product moment correlation coefficient or Pearson correlation coefficient $-1 < r < 1$
- r close to 1 implies strong direct relationship.
- r close to -1 implies strong inverse.
- $r = 0$ when there is no linear relationship at all Pearson correlation coefficient is a measure of the degree of straight line relationship

Interpretation

- The correlation was found to be 0.5648.

```
. pwcorr weight length

|   weight   length
-------------+-------------------
weight |   1.0000
length |   0.5648   1.0000
```

- It means the correlation between weight and height is positively related

- The p-value is less than 0.05 which indicates that the correlation is statistically significant.

limitations of correlation coefficient

- It quantifies only the strength of linear relationship between two variables.
- If the two variables have non-linear relationship, it will not provide a valid measure of this association.
- Care must be taken when the data contain outliers, correlation coefficient is highly sensitive to extreme values
- The estimated correlation coefficient should never be extrapolated beyond the observed ranges of the variables; the relationship between two variables may change outside of this region
- It must be kept in mind that a high correlation coefficient between two variables does not in itself imply a cause and effect relationship.

Introduction to Linear Regression

- We frequently measure two or more variables on the same individual (case, object, etc)
- We do this to explore the nature of the relationship among these variables. There are two basic types of relationships.

Relationship

- Cause and effect
- Functional

- Function: a mathematical relationship enabling us to predict what values of one variable (Y) correspond to given values of another variable (X)
- Y : is referred to as the dependent variable, the response variable or the predicted variable
- X : is referred to as the independent variable, the explanatory variable or the predictor variable.

Linear Regression

Questions to be answered

- What is the association between Y and X?
- How can changes in Y be explained by changes in X?
- What are the functional relationships between Y and X?
- A functional relationship is symbolically written as:

$$Y = f(X)$$

Example: Y=cholesterol versus X=age

Two kinds of explanatory variables:

- Those we can control
- Those over which we have little or no control

Hypertension Data

- We want to study the effect of age (X_1) on blood pressure (Y).
- For this purpose, consider linear regression.
- Model: $Y = \alpha + \beta X_1 + \varepsilon$.
    - α and β are unknown constants to be estimated.
    - ε stands for measurement error.
    - α stands for the value of Y when X_1 is 0.
    - β is the expected change in mean of Y for a unit change X_1.
    - ε is assumed normally distributed with mean 0 and constant variance.

- Commonly, least squares (LS) or maximum likelihood (ML) techniques are used for estimation.
- α and β are estimated by $\hat{\alpha}$ and $\hat{\beta}$, respectively.
- Estimated model: $\hat{Y} = \hat{\alpha} + \hat{\beta} X_1$.
- Least squares estimators of $\hat{\alpha}$ and $\hat{\beta}$:
    - $\hat{\beta} = \frac{SS_{xy}}{SS_{xx}}$,
    - $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$,
    - As before, $SS_{xx} = \sum_{1=1}^{n} (x_i - \bar{x})^2$, and
    - $SS_{xy} = \sum_{1=1}^{n} (x - \bar{x})(y_i - \bar{y})$.

- Hypertension data: Blood pressure (Y) and age (X_1):

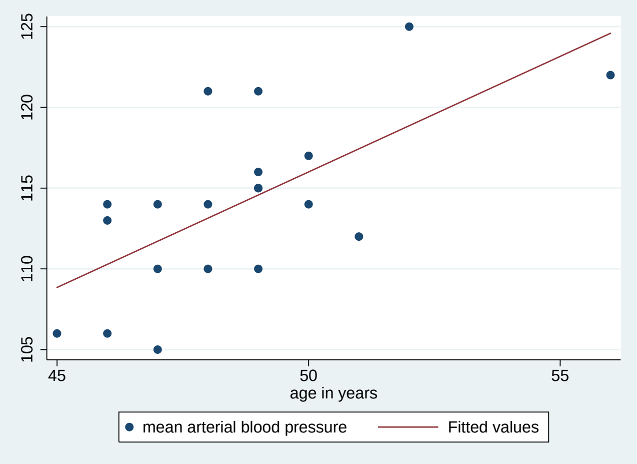


Figure 9: Mean arterial blood pressure versus age.

```
Source |     SS       df     MS          Number of obs =     20
-------------------------------------      F(  1,    18) =   13.82
Model | 243.265993   1  243.265993      Prob > F      =  0.0016
Residual | 316.734007  18  17.5963337      R-squared     =  0.4344
-------------------------------------      Adj R-squared =  0.4030
Total |        560  19  29.4736842      Root MSE      =  4.1948

-----------------------------------------------------------------
y |   Coef.    Std. Err.  t    P>|t|  [95% Conf.Interval]
----------------------------------------------------------------
x1 | 1.430976  .3848601  3.72  0.002   .6224154   2.239537
_cons | 44.45455   18.7277  2.37  0.029   5.109098   83.79999
----------------------------------------------------------------
```

- ANOVA table: It tells the story of how the regression equation accounts for variablity in the response variable.

STATA CODE:

```
regress y x1
```

- Estimate of intercept is: $\hat{\alpha} = 44.45455$.
- Estimate of slope (coefficient of age) is :$\hat{\beta} = 1.43098$.

STATA code for scatter plot and lines:

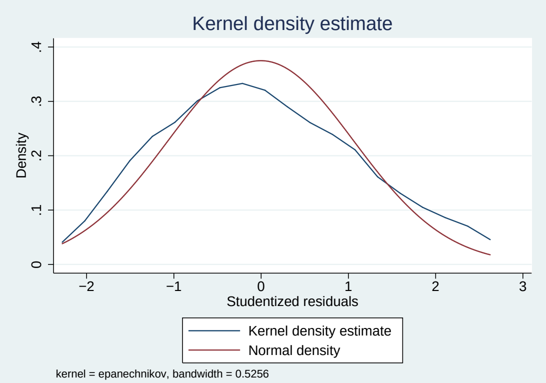
```
twoway (scatter y x1) (lfit y x1)
```

Conclusion

- The ANOVA (F-test) part shows that regression is significant ($p = 0.0016$).
- Adj-R-Sq implies that nearly 40% of the variation in Y is explained by the regression.
- $\hat{\beta}$ is significant ($p = 0.0016$) and implies that for a unit (one year) increase in age, the expected bp increases, on average, by 1.43 mmHg.
- The studentized residuals can be used for identifying outliers.
- We use the 'predict' command with the 'rstudent' option to generate studentized residuals.

Test of Normality

- We can use the ‘predict’ command to create residuals.
- ‘kdensity’, ‘qnorm’ and ‘pnorm’ to check the normality of the residuals.
- kdensity stands for kernel density estimate.
- It can be thought of as a histogram with narrow bins and moving average.



STATA CODE:

```
predict r, resid
kdensity r, normal
```

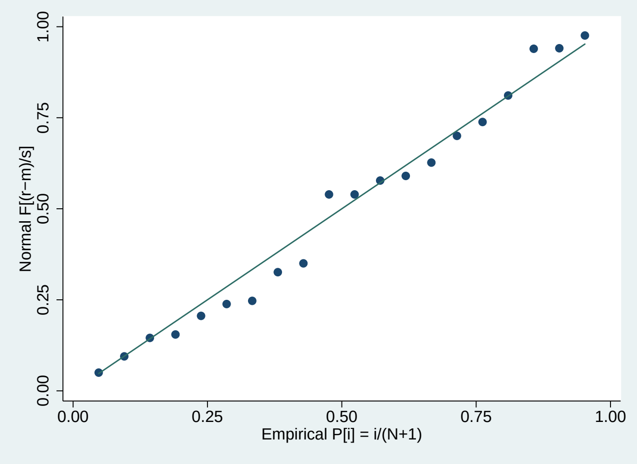


Figure 11: Probability plot of residuals.

- The ‘pnorm’ command graphs a standardized normal probability (P-P) plot against the quantiles of a normal distribution.
- ‘qnorm’ plots the quantiles of a variable against the quantiles of a normal distribution.
- ‘pnorm’ is sensitive to non-normality in the middle range of data.
- ‘qnorm’ is sensitive to non-normality near the tails.

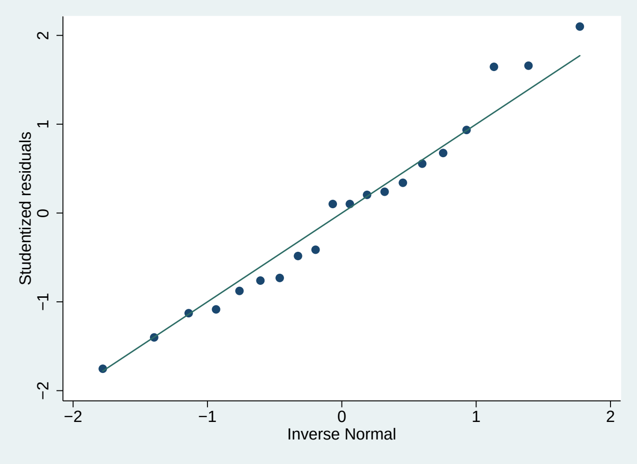


Figure 12: Quantiles of residuals.

STATA CODE:

```
pnorm r
qnorm r
```

Homoscedasticity of Residuals

- Assumption...homogeneity of variance of the residuals.
- If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values.
- If the variance of the residuals is non-constant. it is heteroscedastic.

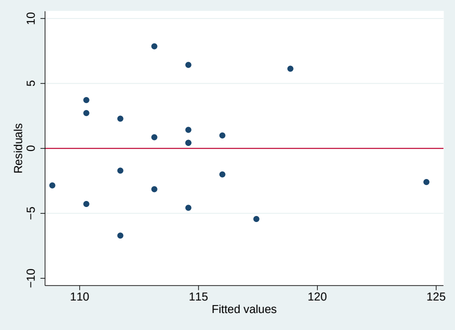


Figure 13: Residual versus Fitted values.

STATA CODE:

```
rvfplot, yline(0)
```

- Cameron and Trivedi's decomposition of IM-test.
- Or hettest is the Breusch-Pagan test.
- p-value < 0.05, reject the hypothesis that states that variance is homogenous.

```
Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
Source |       chi2     df       p
---------------------+-----------------------------
Heteroskedasticity |0.33   2    0.8458
Skewness |       1.22     1    0.2697
Kurtosis |       1.14     1    0.2848
---------------------+-----------------------------
Total |        2.70      4    0.6097
---------------------------------------------------
```

```
Breusch-Pagan / Cook-Weisberg test for
heteroskedasticity

Ho: Constant variance
Variables: fitted values of y

chi2(1)      =     0.01
Prob > chi2  =   0.9324
```

STATA CODE:

```
estat imtest
estat hettest
```

One categorical predictor

- Relationship between a continuous outcome and a single categorical variable.
- Dummy variables for the categories needed.
- Objective: estimation/prediction the effect of the predictor on the response.
- We want to study whether birth weight (Y) varies with sex (X_1; female = 0, male = 1).
- For this purpose, consider liner regression.
- Model: $Y = \alpha + \beta X_1 + \varepsilon$
  - α and β are unknown constants to be estimated.
  - ε stands for measurement error.
  - β is the expected difference in mean of Y among X_1 categories.

```
Source    |   SS         df       MS        Number of obs =    7873
----------------------------------------    F(  1,  7871) =   97.91
Model     | 25750334.5   1    25750334.5    Prob > F      =  0.0000
Residual  | 2.0700e+09  7871  262990.174    R-squared     =  0.0123
-------------+------------------------      Adj R-squared =  0.0122
Total    | 2.0957e+09  7872  266227.896    Root MSE      =  512.83

------------------------------------------------------------------
weight   | Coef.     Std. Err.    t     P>|t|    [95% Conf.Interval]
------------------------------------------------------------------
1.sex   | 114.4014  11.56138   9.90   0.000    91.738   137.0647
_cons   | 3055.323  8.253152  370.20  0.000   3039.144  3071.501
------------------------------------------------------------------
```

STATA CODE:

```
regress weight i.sex
```

- The ‘i.sex’ is used in the code because sex is an indicator variable.
- The ‘1.sex’ stands for males, and hence female is a reference group.
- Estimate of intercept is: $\hat{\alpha} = 3055.32$.
- Estimate of slope (coefficient of age) is :$\hat{\beta} = 114.40$.
- Interpretation?

Multiple linear regression

- Extension of simple linear regression with more than one predictors.
- The predictors could be either a continuous or a categorical variable.
- Objective: estimation/prdiction the effect of the predictors on the response.
- Hypertension Data
  - We want to study the effect of (X_1), (X_2), (X_3), (X_4), (X_5), (X_6) on blood pressure (Y).
  - Predictors: $X_1 =$ age (years); $X_2 =$ weight (kg); $X_3 =$body surface area (sq m); $X_4 =$ duration of hypertension; $X_5 =$ basal pulse (beats/min); $X_6 =$ measure of stress.
  - For this purpose, consider Multiple liner regression.

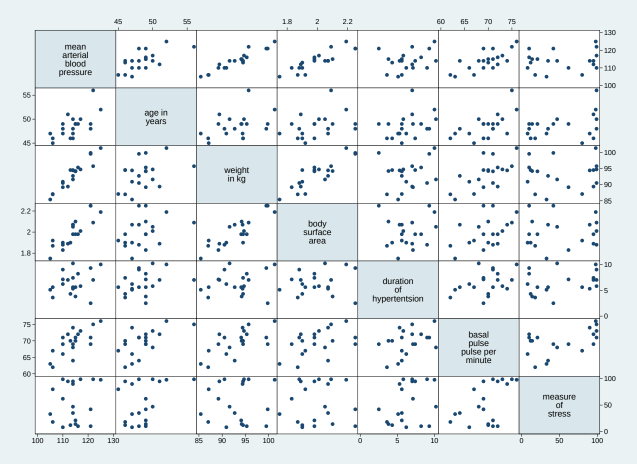


Figure 14: Scatter Plot Matrix

STATA CODE:

```
graph matrix  y x1 x2 x3 x4 x5 x6
```

- What do you observe from the scatter plot matrix?
- Which variables seem to be correlated which other variables?

- Coefficient of determination (R^2):

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}.$$

- The adjusted R^2 statistic is an estimate of the population R^2 taking account of the fact that the parameters were estimated from the data,

$$\text{Adj-}R^2 = 1 - \frac{(n-1)(1-R^2)}{n-p}.$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

- Response: Y=mean arterial blood pressure (mm HG);
- Predictors: X_1 = age (years); X_2 = weight (kg); X_3 = body surface area (sq m); X_4 = duration of hypertension; X_5 = basal pulse (beats/min); X_6 = measure of stress.

```
Source |    SS       df      MS         Number of obs =      20
-------------------------------------     F(  6,    13) =  560.64
Model | 557.844135   6  92.9740225     Prob > F      =  0.0000
Residual |  2.1558651  13  .165835777     R-squared     =  0.9962
-------------------------------------     Adj R-squared =  0.9944
Total |        560  19  29.4736842     Root MSE      =  .40723
```

- Much improvement as compared to the model with only age: say, based on Adj R-sq.

```
-------------------------------------------------------------
y |  Coef.     Std. Err.   t    P>|t|  [95% Conf.Interval]
-------------+-----------------------------------------------
x1 |  .7032596  .0496059  14.18  0.000   .5960926  .8104266
x2 |  .9699192  .0631086  15.37  0.000   .8335815  1.106257
x3 |  3.776502  1.580154   2.39  0.033   .3627878  7.190217
x4 |  .0683829  .0484416   1.41  0.182  -.0362687  .1730346
x5 | -.0844846  .0516091  -1.64  0.126  -.1959792  .0270101
x6 |  .0055715  .0034123   1.63  0.126  -.0018003  .0129433
_cons | -12.87047  2.556654  -5.03  0.000  -18.39378  -7.34715
-------------------------------------------------------------
```

STATA CODE:

```
regress y  x1 x2 x3 x4 x5 x6
```

Variable Selection

- Forward selection,
- Backward elimination,
- Stepwise regression.

Hypertensive Data

Stepwise variable selection using partial F-test with 0.2:

```
begin with full model
p < 0.2000         for all terms in model

Source |   SS        df       MS          Number of obs =      20
------------------------------------        F(  6,    13) =  560.64
Model | 557.844135  6  92.9740225       Prob > F      =  0.0000
Residual |  2.1558651 13  .165835777        R-squared     =  0.9962
------------------------------------        Adj R-squared =  0.9944
Total |     560    19  29.4736842       Root MSE      =  .40723

-------------------------------------------------------------------
y |    Coef.    Std. Err.  t     P>|t|   [95% Conf. Interval]
-------------------------------------------------------------------
x1 |  .7032596  .0496059  14.18  0.000   .5960926    .8104266
x2 |  .9699192  .0631086  15.37  0.000   .8335815    1.106257
x3 | 3.776502  1.580154   2.39  0.033   .3627878    7.190217
x4 |  .0683829  .0484416   1.41  0.182  -.0362687    .1730346
x5 | -.0844846  .0516091  -1.64  0.126  -.1959792    .0270101
x6 |  .0055715  .0034123   1.63  0.126  -.0018003    .0129433
_cons | -12.87047  2.556654  -5.03  0.000  -18.39378    -7.34715
-------------------------------------------------------------------
```

STATA CODE:

```
sw regress y  x1 x2 x3 x4 x5 x6,pr(0.2)
```

- Consider also 0.05 as a cut-off point and refit the model.
- Consider the diagnostic plot. What do you observe?
- Can we further improve model fit?, though the model seems sufficient at it is (Adj R-Sq, Diagnosis).
- Apply transformation on Y, say logarithmic transformation.

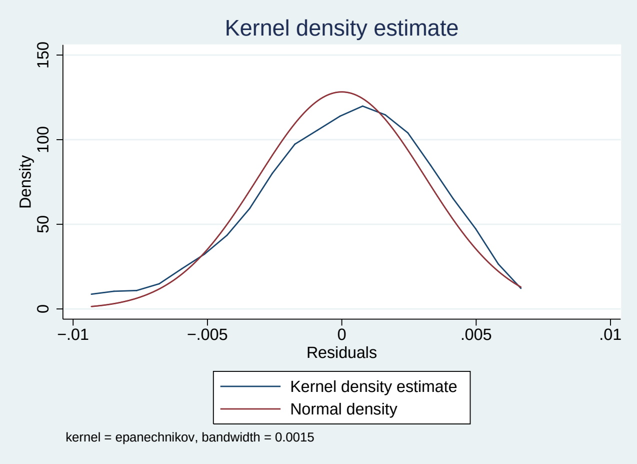


Figure 15: Kernel density of residuals (log y)

- Transformation does not yields substantial improvement.
- Adj R-sq (0.9937) does not seem to improve any more.
- Stick to the first model.
- Interpretation?
- Interpretation of Age in Simple and Multiple linear regression?

Multicollinearity

- When there is a perfect linear relationship among the predictors, the estimates cannot be uniquely computed.
- The term collinearity implies that two variables are near perfect linear combinations of one another.
- The regression model estimates of the coefficients become unstable.
- The standard errors for the coefficients can get wildly inflated.
- We can use the vif command after the regression to check for multicollinearity.
- As a rule of thumb, a variable whose values are greater than 10 may need further investigation.
- Tolerance, defined as 1/VIF, is used by many researchers to check on the degree of collinearity.
- The standard errors for the coefficients can get wildly inflated.

$$\mathrm{VIF}(x_k) = \frac{1}{1 - R_k^2}.$$

- $\mathrm{VIF}(x_k)$ is the variance inflation factor for explanatory variable x_k.
- R_k^2 is the square of the multiple correlation coefficient obtained from regressing x_k on the remaining explanatory variables.

```
Variable   |VIF    1/VIF
-----------+--------------------
x2 |      8.42    0.118807
x3 |      5.33    0.187661
x5 |      4.41    0.226574
x6 |      1.83    0.545005
x1 |      1.76    0.567277
x4 |      1.24    0.808205
-----------+--------------------
Mean VIF | 3.83
```

STATA CODE:

```
regress y  x1 x2 x3 x4 x5 x6
vif
```

- There are different suggestions for the cut-off points
- Some says 5 and some say even 10. So, if we consider 10, multicollinearity does not seem to be a major problem...
- If we take 5, then there are two variables which needs decision
- One solution to dealing with multicollinearity is to remove some of the violating predictors from the model.

Jimma Infant Data

- We want to study the effect of (X$_1$), (X$_2$), (X$_3$), (X$_4$) on birth weight (Y).
- Predictors: (X_1 = sex; female = 0, male = 1); X_2 = place (semi−urban); X_3 = place (rural); X_4 = Age of mother.
- Originally, place was coded as: (urban = 1, semi-urban = 2, and rural = 3). By default, stata considers the first (urban = 1) as a reference.
- For this purpose, consider Multiple liner regression.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

  - Response: Y = birth weight;
  - Predictors: X_1 = sex; female = 0, male = 1); X_2 = place (semi−urban); X_3 = place (rural); X_4 = age of mother (years).

```
Source |       SS      df       MS        Number of obs =    7873
----------------------------------------    F(  4,  7868) =  169.34
Model |   166126585    4   41531646.3    Prob > F       =  0.0000
Residual |   1.9296e+09 7868   245249.036    R-squared      =  0.0793
----------------------------------------    Adj R-squared =  0.0788
Total |   2.0957e+09 7872  266227.896    Root MSE       =  495.23

-------------------------------------------------------------------
weight |    Coef.   Std. Err.     t    P>|t|    [95% Conf.Interval]
-------------------------------------------------------------------
1.sex |  109.2883  11.16942    9.78  0.000    87.39329   131.1833
|
place |
2   | -44.07271  16.25057   -2.71  0.007  -75.92814  -12.21729
3   | -279.7563  13.86682  -20.17  0.000   -306.939  -252.5737
|
momage |  7.809375  .8910886    8.76  0.000    6.062605   9.556145
_cons |  3010.102  26.26505  114.60  0.000    2958.615   3061.588
-------------------------------------------------------------------
```

- $\hat{\alpha} = 3010.102$, $\hat{\beta}_1 = 109.2883$, $\hat{\beta}_2 = -44.07271$, $\hat{\beta}_3 = -279.7563$, $\hat{\beta}_4 = 7.809375$.
- Interpretation?

STATA CODE:

```
.regress weight i.sex i.place momage
. estat vif

Variable |        VIF       1/VIF
-------------+------------------------
1.sex |   1.00    0.999139
place |
2   |       1.53    0.655694
3   |       1.54    0.650037
momage |  1.01    0.988233
-------------+------------------------
Mean VIF |      1.27
```

Assignment 1

- Consider the hematological data we discuss in the first part of the presentation. The outcome variables were WBC, Hgb, platelets. The factor variable is BMI in three categories; 1=severe malnurtioned, 2=Malnurtioned, and 3=Normal
    - Fit ANOVA model for each of the outcome variable with the factor variable and interpret the result
    - Fit ANCOVA and interpret the result
- Consider Jimma survival data. The outcome variable is birth height of new born and others are independent variable.
    - Determine the correlation between mothers age and the height of the new born
    - Test if the correlation is significant
    - Fit linear regression for birth weight with sex, residence, family income and mothers age.
    - Test the assumptions
    - Interpret the result

Chapter 3

Categorical Data Analysis

Categorical Data

- Variables that are measured using Nominal scale and ordinal scale
  - The variable may have only two levels called a dichotomous (E.g. Sex)
  - The variable may have more than two levels called a polygamous (E.g. Blood group)
- Continuous (numeric) variables can be condensed to categorical variables if recoded
  - BMIunderweight, normal. and overweight
  - Income (very poor, poor, rich, very rich)
  - Age ($< 15 years$), 16-24 years, 25-34 years, and $\geq$ 35 years

Contingency Tables

- When working with categorical variables, we often arrange the counts in a tabular format called contingency tables
- If a contingency table involves two dichotomous variables then it is a 2x2 (two way table)

	Diseases Status		
Exposure	Positive	Negative	Total
Exposed	a	b	a+b
Not Exposed	c	d	c+d
Total	a+c	b+d	n

Contingency Tables

- It can be generalized to accommodate into *rxc* contingency table (r-rows and c-columns)

Row Variable	Column Variable						Total
	1	2	.	.	.	c	
1	$R_1 C_1$	$R_1 C_2$	.	.	.	$R_1 C_n$	R_1
2	$R_2 C_1$	$R_2 C_2$	.	.	.	$R_2 C_n$	R_2
3	$R_3 C_1$	$R_3 C_2$	.	.	.	$R_3 C_n$	R_3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
R	$R_n C_1$	$R_n C_2$	.	.	.	$R_n C_n$	R_n
Total	C_1	C_2	.	.	.	C_c	N

Steps in Testing Association

- The test statistic chi-square is used to test the association between two categorical variables

Hypothesis

1. The hypothesis to be tested can be stated as:
   H_0: There is no association H_1: There is association
2. Compute the calculated and tabulated value of the test statistic
3. Compare the two (tabulated and calculated values)
4. Decision

- If the contingency table is $2X2$, then

$$\chi^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

If the contingency table is *rxc*

- The formula for the calculated value of the test statistic is different when the contingency table is *rxc*

$$\chi^2 = \frac{\sum (O_{ij} - E_{ij})^2}{E_{ij}}$$

Where

- O_{ij} is observed value
- E_{ij} is expected value

$$E_{ij} = \frac{R_i \times C_j}{N}$$

- The Chi-squared test measures the disparity between observed frequencies (data from the sample and expected frequencies (probability distribution)

- The Chi-squared test is valid
  - If no observed cell have value 0
  - The values must be counts but not percent or proportion
  - And no 20% of expected cell has less than value of 5

Example

- A cross-sectional survey was conducted to assess the association between head injury and wearing helmet. A total of 795 individuals were included in the study and the data is given below;

	Wearing Helmet		
Head injury	Yes	No	Total
Yes	17	218	235
No	130	428	558
Total	147	646	793

- Test the presence of association between head injury and wearing helmet at 0.05 level of significant

Solution

- hypothesis
  H_O : There is no association between helmet and head injury H_A : There is association between helmet and head injury
- Test statistics

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+d)(c+d)}$$

$$= \frac{(17\times 428 - 218\times 130)^2 793}{235\times 559\times 147\times 646} = 28.26$$

- The critical value of the test statistic

$$\chi^2_{0.05} = 3.84$$

- Since 28.26$>$ 3.84, reject the null hypothesis

Example 2

- A sample of 263 students who bought lunch at a school canteen were asked whether or not they developed gastroenteritis. The response is given below

Ate Sandwich	Gastroenteritis		Total
	Yes	No	
Yes	109	116	225
No	4	34	38
Total	113	150	263

- Test the presence of association between sandwich and gastroenteritis at 0.05 level of significant

Solution

- Step 1. H_O : There is no association
- Step 2. H_a: There is association
- Step 3. Test statistics $\chi^2 = 17.6$
- Step 4: Critical value of $\chi^2_{1(0.05)} = 3.84$
- Step 5: decision: since $17.6 > 3.84$ then reject the null hypothesis and decide as there is association between eating sandwich and gastrointestinal pain

Probability Distribution

- Binomial Distribution
- Under the assumption of n independent, identical trials, Y has the binomial distribution with index n and parameter π
- The probability of outcome y for Y equals

$$P(Y = y) = \frac{n!}{y!(n-y)!} p^{y}(1-p)^{(n-y)}$$

Maximum Likelihood Estimation

- In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given the data

- MLE begins with writing a mathematical expression known as the Likelihood Function of the sample data

- The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimates

Example: Logistic regression

- Let p_j be the probability that a subject in stratum j has a success
- Then the probability of observing y_j events in stratum j is

$$P(Y_j = y_j) = \binom{n_j}{y_j} p_j^{y_j} (1 - p_j)^{n_j - y_j}$$

- We know p_j is a function of covariates
- assume we are interested in two covariates, x_j,such that

$$p_j = \frac{e^{\alpha_0 + \beta x_j}}{1 + e^{\alpha_0 + \beta x_j}}$$

Example: Logistic regression

- Then, the likelihood function can be written as

$$L(y) = \prod_{J}^{j+1} \binom{n_j}{y_j} p_j^{y_j} (1 - p_j)^{n_j - y_j}$$

$$= \prod_{J}^{j+1} \left(\frac{e^{\alpha_0+\beta x_j}}{1 + e^{\alpha_0+\beta x_j}}\right)_j^{y} \left(\frac{1}{1 + e^{\alpha_0+\beta x_j}}\right)^{(n_j - y_j)} \times \binom{n_j}{y_j}$$

$$= \prod_{J}^{j+1} \frac{(e^{\alpha_0+\beta x_j})^{y_j}}{(1 + e^{\alpha_0+\beta x_j})^{n_j}} \times \binom{n_j}{y_j}$$

$$= \frac{(e^{\alpha_0 t_0+\beta t_1})}{\prod_J^{j+1} (1 + e^{\alpha_0+\beta x_j})^{n_j}} \times \prod_{J}^{j+1} \binom{n_j}{y_j}$$

Inference for proportion

- The point estimate E(p)=π
- Confidence interval for proportion

$$p - z_{\alpha/2}SE(P), p + z_{\alpha/2}SE(P)$$

- Where

$$SE(p) = \sqrt{\frac{p(1-p)}{n}}$$

Contingency Tables 2

- Partial Tables: a two-way cross-sectional slice of the three-way table where X and Y are cross classified at separate levels of the control variable Z
- it shows the effect of X on Y while controlling for Z
- Marginal table: a two-way contingency table obtained by combining the partial tables
- The marginal table ignores Z rather than controlling for it.

Example

- Prevalence study: A total of 715 births, cross classified by clinic (Z), prenatal care (X) and outcome (Y)

Clinic	Prenatal Care	Died	Lived
1	less	3	176
	more	4	293
2	less	17	197
	more	2	23

- Construct the corresponding partial tables and the marginal table.

Confounding

- When the partial and marginal associations are different, there is said to be CONFOUNDING.
- Confounding occurs when two variables are associated with a third in a way to obscure their relationship.

Joint, Marginal, and Conditional Probabilities

- Probabilities for contingency tables can be of three types: joint, marginal, or conditional
- Consider the variable X and Y from randomly chosen subjects
- Let $\pi_{ij} = P(X = i, Y = j)$ denote the probability that (X, Y) falls in the cell in row i and column j
- The probabilities π_{ij} is the joint distribution of X and Y

Marginal Probabilities

- The marginal distributions are the row and column totals of the joint probabilities

Variable 1	Diseases	Non Diseases	Total
Exposed	n_{11}	n_{12}	n_{1+}
Not Exposed	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

- $\pi 1+ = \pi_{11} + \pi_{12}$
- It can be calculated as $p_{ij} = \frac{n_{ij}}{n}$

Conditional Probabilities

- It is informative to construct a separate probability distribution for Y and X at each level of Z.

- Consider The variables sex, smoking and lung cancer

Sex	Smoking	Diseases	Non Diseases	Total
Male	Yes	n_{11}	n_{12}	n_{1+}
	No	n_{21}	n_{22}	n_{2+}
Female	Yes	n_{31}	n_{32}	n_{3+}
	No	n_{41}	n_{42}	n_{4+}
Total		n_{+1}	n_{+2}	n_{++}

- Simpson's Paradox: The result that a marginal association can have different direction from the conditional associations

Test of Independence

- Two variables are statistically independent if the joint probabilities equal the product of their marginal probabilities:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$

for all i and j

Application for screening test

- Sensitivity: the proportion of diseased individuals detected as positive by the test.
- Specificity: the proportion of healthy individuals detected as negative by the test.
- Positive predictivity: given that the test result is positive, what is the probability that, in fact, the disease is present?
- Negative predictivity: given that the test result is negative, what is the probability that, in fact, the disease is not present?

Example

- Cytological test to screen women for cervical cancer

Diseases	negative	positive	Total
Negative	23362	362	23724
Positive	225	154	379
Total	23587	516	24103

Example

- Sensitivity=154/379=40.6%
- Specificity=23362/23724=98.5%
- Positive predictivity=154/516=29.8%
- Negative predictivity=23362/23587=99.0%

Analysis of Categorical Data

- Descriptive;
  - Summary statistics including percentages, Odds ratios, Risk ratios, Relative risk.
  - Graphical techniques: bar chart, pie-chart.
- Inferential Statistics.

Jimma Infant Data...

```
birth |
weight |      Freq.     Percent        Cum.
------------+-----------------------------------
0 |      7,207       91.54       91.54
1 |        666        8.46      100.00
------------+-----------------------------------
Total |      7,873      100.00
```

STATA CODE:

```
tab bwt
```

- Cross tabulation of bwt and gender ($male = 1$).
- The option “row” can be used to obtain the row percentages.

```
birth |             sex
weight |         0          1 |     Total
-----------+----------------------+----------
0 |     3,466      3,741 |     7,207
|     48.09      51.91 |    100.00
-----------+----------------------+----------
1 |       395        271 |       666
|     59.31      40.69 |    100.00
-----------+----------------------+----------
Total |     3,861      4,012 |     7,873
|     49.04      50.96 |    100.00
```

STATA CODE:

```
tab bwt sex, row
```

- Consider the following contingency table:

```
Diseased
Yes     No |     Total
--------------------------------------
yes|     a      b  |     a+b
exposed ------------------------------
No |     c      d  |     c+d
--------------------------------------
Total |    a+c    c+d |   a+b+c+d
```

Odds

- The odds of being diseased among exposed is:

$$\frac{a}{a+b}.$$

- The odds of being diseased among unexposed is:

$$\frac{c}{c+d}.$$

Odds Ratio

- The ratio of the odds of being diseased among exposed to unexposed is:
$$\frac{a}{a+b} \div \frac{c}{c+d} = \frac{ad}{bc}.$$
- Odds ratio compares the risk of being diseased in the exposed versus the unexposed.
$$\text{Var}(\text{ln}(\hat{OR})) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$
- Confidence interval: $\hat{O}R\text{exp}[\pm Z_{1-\alpha/2}\sqrt{\hat{V}ar(In\hat{O}R)}]$.

Jimma Infant Data:

- Outcome-bwt ‘low bwt’ versus ‘normal’.
- Exposure-gender ‘female’ versus ‘male’.

```
Proportion
|   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
Cases |       395         271  |        666       0.5931
Controls |      3466        3741   |      7207       0.4809
-----------------+------------------------+------------------------
Total |      3861        4012  |       7873       0.4904
|                        |
|      Point estimate    |    [95% Conf. Interval]
|------------------------+-------------------------
Odds ratio |         1.573211       |    1.334496    1.854786
+--------------------------------------------------
```

STATA CODE:

```
cci 395 271 3466 3741
```

Test of homogeneity

Consider partial table for variables X, Y, Z and θ denotes odds ratio for each partial table

- Homogeneous association if any conditional odds ratio formed using two levels of X and two levels of Y is the same at each level of Z
- There is homogeneous $X - Y$ association in a $2x2xK$ table when $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(k)}$

Example

- Suppose X=smoking (yes,no),
- Y =lung cancer (yes,no),
- Z=age ($< 45, 45 - 65, > 65$),
- Then from the partial table, we found
- $\theta_{XY(1)} = 1.2, \theta_{XY(2)} = 2.8, and \theta_{XY(2)} = 6.2.$

  Hence, the conditional odds ratio seems to change across levels of the third variable (homogeneous association does not exist):

Introduction to Generalized Linear Models (GLM)

- GLMs are often employed for modeling a non-Gaussian data
- It extends ordinary regression models for non-Gaussian data
- Statistical models that relate outcome variables such as, binary, counts, rates etc, to a linear combination of predictor variables
- It uses different link function to connect the outcome variable to the predictor variables

Components of GLM

- Three components specify GLM:
    - The random component identifies a vector of observations of Y and its probability distribution;
    - The systematic component is a specification for the predictor variables
    - And the link function specifies the function of $\mathrm{E}(Y)$ that the model equates to the systematic component

Modeling the outcome variable

- The outcome variable is binary.
- The objective is to study the effect of explanatory variables.
  - An outcome variable with two possible categorical outcomes (1=success; 0=failure).
  - The events are independent from subject to subject.
  - Explanatory variables can be categorical and/or continuous.
  - A way to estimate the probability p of the success event of the outcome variable.

Measuring the Probability of an outcome

- The probability of the outcome is measured by the odds of occurrence of an event.

- If p is the probability of an event, then $(1 - p)$ is the probability of it not occurring.

- Odds of success $= \frac{p}{(1-p)}$.

- The effects of explanatory variable, say x, on the probability of the diseases can be put on the odds as:

$$\text{Odds} = \frac{p}{1-p} = \exp(\alpha + \beta x)$$

Logistic Regression

- Taking natural logarithm on both sides:

$$\ln(\frac{p}{1-p}) = \log[\exp(\alpha + \beta_1 x_1)].$$

$$\Rightarrow \text{logit}(p) = \alpha + \beta_1 x_1.$$

- Consider an outcome as ‘diseased’ or ‘not diseased’.
- α represents the overall disease risk.
- β_1 represents the fraction by which the disease risk is altered by a unit change in x_1.

Multiple Logistic Regression Model Cont'd

- The joint effects of all explanatory variables, say $x_1, x_2, \ldots, x_p$ put together on the odds is:

$$\text{Odds} = \frac{p}{1-p} = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p).$$

- Taking natural logarithm on both sides:

$$\ln(\frac{p}{1-p}) = \log[\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)].$$

$$\Rightarrow \text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

- Consider an outcome as 'diseased' or 'not diseased'.
- α represents the overall disease risk.
- β_1 represents the fraction by which the disease risk is altered by a unit change in x_1.

Multiple Logistic Regression Model Cont'd

- β_2 represents the fraction by which the disease risk is altered by a unit change in x_2, and so on.
- What changes is the log odds. The odds themselves are changed by $\exp(\beta)$.
- Only one predictor variable.

$$\text{logit}(p) = \alpha + \beta_x.$$

Jimma Infant Data

- Outcome: birth weight status(bwt, $1 = $ ‘*low*).
- Factor: sex ($1 = $ ‘*male*′).

```
Logistic regression                  Number of obs    =       7873
LR chi2(1)      =       30.82
Prob > chi2     =      0.0000
Log likelihood = -2266.5465          Pseudo R2        =     0.0068

---------------------------------------------------------------
bwt |   Coef.     Std. Err.   z    P>|z|   [95% Conf.Interval]
-------------+-------------------------------------------------
1.sex | -.4531187  .0823256  -5.50   0.000  -.6144739  -.2917634
_cons | -2.171871  .0531052  -40.90  0.000  -2.275955  -2.067786
---------------------------------------------------------------
```

- The fitted logistic regression model is:

$$\text{logit}(p) = \hat{\alpha} + \hat{\beta} sex.$$

- Substituting the estimated values yields:

$$\text{logit}(p) = -2.172 - 0.453 sex.$$

STATA CODE:

```
logit bwt i.sex
```

  - 'i' indicates that sex is a categorical variable.
  - The option 'or' can be added to obtain the odds ratio estimates.

```
Logistic regression                Number of obs    =       7873
LR chi2(1)      =      30.82
Prob > chi2     =     0.0000
Log likelihood = -2266.5465      Pseudo R2        =     0.0068

-------------------------------------------------------------
bwt | Odds Ratio  Std. Err.  z    P>|z|  [95% Conf.Interval]
-------------------------------------------------------------
1.sex | .6356427   .0523297  -5.50  0.000  .5409254  .7469452
_cons | .1139642   .0060521 -40.90  0.000  .1026988  .1264654
-------------------------------------------------------------
```

STATA CODE:

```
recode sex (1=0)  (0=1), generate(sex2)
```

- The estimate stands for males, i.e., *sex* $= 1$ and females are considered as a reference group.
- Interpretation of $\exp(\hat{\beta})$?
- Interpretation of $\exp(\hat{\alpha})$?
- Conclusion? Why?

- Take males as a reference group.
- The estimate now stands for females.
- This requires to recode the sex to say, sex2 and fit the model again using the following code:

```
recode sex (1=0)  (0=1), generate(sex2)
logit bwt i.sex2,or
```

```
Logistic regression                 Number of obs   =       7873
LR chi2(1)      =      30.82
Prob > chi2     =     0.0000
Log likelihood = -2266.5465         Pseudo R2       =     0.0068

----------------------------------------------------------------
bwt | Odds Ratio  Std. Err.  z    P>|z|  [95% Conf. Interval]
-------------+--------------------------------------------------
1.sex2 |  1.573211  .1295156   5.50  0.000  1.338786    1.848684
_cons |  .0724405   .004557 -41.73  0.000  .0640375    .0819461
----------------------------------------------------------------
```

- Compare $\exp(\hat{\beta})$ and $\exp(\hat{\alpha})$ in the above table with the previous ones.
- Interpretation and conclusion?

Multiple Logistic Regression Model

- Logistic regression with more than one predictor variables.
- It is also referred to as multivariate logistic regression.
- Control effect of one independent variable on another independent variable and vice-versa.
- Jimma Infant Data
  - Outcome: birth weight status(bwt, $1 = `low'$).
  - Factor1: sex2 ($`1 = female'$).
  - Factor2: place of residence ($1 = `urban', 2 = `semi\text{-}urban', 3 = `rural'$).

```
Logistic regression                Number of obs    =       7873
LR chi2(3)      =     258.69
Prob > chi2     =     0.0000
Log likelihood =  -2152.613        Pseudo R2        =     0.0567

-----------------------------------------------------------------
bwt | Odds Ratio  Std. Err.  z    P>|z|  [95% Conf. Interval]
-----------------------------------------------------------------
1.sex2 |  1.54585   .1289271   5.22  0.000   1.31273    1.820368
|
place  |
2   | 1.935937   .3516486   3.64  0.000  1.356053    2.763794
3   | 5.436301    .835046  11.02  0.000  4.023039    7.346031
|
_cons | .0210955    .003245 -25.09  0.000  .0156048    .0285184
-----------------------------------------------------------------
```

$$\text{logit}(p) = \hat{\alpha} + \hat{\beta}_1 sex2 + \hat{\beta}_2 SU + \hat{\beta}_3 R.$$

SU and R refer to semi-urban and rural, respectively.

STATA CODE:

```
logit bwt i.sex2 i.place, or
```

- Interpretation of $\exp(\hat{\beta}_1) = 1.55$?
- Interpretation of $\exp(\hat{\beta}_2) = 1.94$?
- Interpretation of $\exp(\hat{\beta}_3) = 5.44$?
- Conclusion?

Jimma Infant Data

Add a third variable:

- Outcome: birth weight status(bwt, $1 = `low'$).
- Factor1: sex2 ($1 = `female'$).
- Factor2: place of residence ($1 = `urban', 2 = `semi\text{-}urban', 3 = `rural'$).
- Factor3: maternal age, ‘momage’ measured in years (continuous).

```
Logistic regression                 Number of obs   =       7873
LR chi2(4)      =     271.06
Prob > chi2     =     0.0000
Log likelihood = -2146.4273         Pseudo R2       =     0.0594

-----------------------------------------------------------------
bwt |    Coef.   Std. Err.   z    P>|z|  [95% Conf.Interval]
-----------------------------------------------------------------
1.sex2 |  .4378871  .0834884   5.24  0.000  .2742529   .6015213
|
place |
2   |   .661642     .1817   3.64  0.000   .3055165   1.017767
3   |  1.725587  .1539474  11.21  0.000   1.423855   2.027318
|
momage | -.0231749  .0066597  -3.48  0.001  -.0362277  -.0101222
_cons | -3.274631  .2257413 -14.51  0.000  -3.717076  -2.832186
-----------------------------------------------------------------
```

$$\text{logit}(p) = \hat{\alpha} + \hat{\beta}_1 sex2 + \hat{\beta}_2 SU + \hat{\beta}_3 R + \hat{\beta}_4 A.$$

$$\text{logit}(p) = -3.275 + 0.438 sex2 + 0.662 SU + 1.726 R - 0.023 A.$$

SU, R and A refer to semi-urban, rural and maternal age, respectively.

Odds ratio estimates:

```
Logistic regression                Number of obs    =      7873
LR chi2(4)      =     271.06
Prob > chi2     =     0.0000
Log likelihood = -2146.4273        Pseudo R2        =    0.0594

-----------------------------------------------------------------
bwt | Odds Ratio  Std. Err.  z    P>|z|  [95% Conf.Interval]
-----------------------------------------------------------------
1.sex2 |  1.54943   .1293594   5.24  0.000   1.315547  1.824893
|
place |
2   | 1.937972   .3521295   3.64  0.000   1.357326   2.76701
3   | 5.615815   .8645402  11.21  0.000   4.153101  7.593693
|
momage | .9770915   .0065071  -3.48  0.001   .9644206  .9899289
_cons | .0378308     .00854 -14.51  0.000   .0243049   .058884
-----------------------------------------------------------------
```

- Interpretation of $\exp(\hat{\beta}_4)$?
- Interpretation of, say, $\exp(\hat{\beta}_1)$ with and with out maternal age in the model?
- Conclusion?

Model Comparison

Compare the model fit statistics, say -2Log-likelihood of the three models considered so far, i.e,

- Simple logistic regression: only ‘sex’ as a factor,
- Multiple logistic regression: ‘sex’ and ‘place’ as two factor variables, and
- Multiple logistic regression: ‘sex’, ‘place’ and ‘maternal age’.

Which one fits the data better? Why?

Conditional Logistic Regression Model

- Conditional logistic regression is an extension of logistic regression that allows to take into account stratification and matching.
- Its main field of application is observational studies and in particular epidemiology
- Is appropriate in models of choice behavior, where the explanatory variables may include attributes of the choice alternatives as well as characteristics of the individuals making the choices
- Can be used for both binary or multinomilal response variable

Conditional Logistic Model

- Is used when there is matched studies (matched case-control)
- Logistic regression uses unconditional likelihood estimation
- The probability of X_{i1} belonging to the case, and X_{i2} belonging to the control is

$$P(X_{i1}|Y=1)P(X_{i2}|Y=0)$$

- the probability of observing the unordered values X_{i1}, X_{i2} for the two subjects is

$$P(X_{i1}|Y=1)P(X_{i2}|Y=0)+P(X_{i2}|Y=1)P(X_{i1}|Y=0)$$

Conditional Logistic Model

- The desired conditional probability is

$$\frac{P(X_{i1}|Y=1)P(X_{i2}|Y=0)}{P(X_{i1}|Y=1)P(X_{i2}|Y=0)+P(X_{i2}|Y=1)P(X_{i1}|Y=0)}$$

- Using Bayes' rule, we can rewrite the above conditional probability as

$$\frac{P(Y=1|X_{i1})P(Y=0|X_{i2})}{P(Y=1|X_{i1})P(Y=0|X_{i2})+P(Y=1|X_{i2})P(Y=0|X_{i1})}$$

Conditional Logistic Model

- Applying the logistic regression model for each probability in the last expression, the desired conditional probability is

$$\frac{exp(\beta X_{i1})}{exp(\beta X_{i1}) + exp(\beta X_{i2})}$$

- Appropriate for matched case-control study
- Application in stata

  ```
  clogit outcome covariates, group(pairid)
  ```
- Can be used as an extension for multinomial logistic regression

Ordinal Logistic Regression

- Many categorical dependent variables are ordered
  - Label of agreement (strongly disagree, disagree, agree, strongly agree)
  - Social class
  - Satisfaction
  - Pain Score
- If numerical values assigned to categories do not accurately reflect the true distance
- Thus linear regression may not be appropriate

Strategies to deal with ordered categorical variables

- Use linear regression anyway
  - Commonly done; but can give incorrect results
  - Possibly check robustness by varying coding of interval between outcomes
- Collapse variables to dichotomy, use binary logistic regression model
  - Works fine, but throws away useful information
- If you are not confident about ordering, use multinomial logistic regression
- Ordered logit / ordinal probit

Proportional Odds Assumption

- The fact that you can calculate odds ratios highlights a key assumption of ordered logit
  - Proportional odds assumption
  - parallel regression assumption
- Model assumes that variable effects on the odds of lower vs. higher outcomes are consistent
  - Effect on odds of "too little" vs "about right" is same for "about right" vs "too much"
  - Controlling for all other vars in the model
- If this assumption doesn't seem reasonable, consider multinomial logit
- Interpretation strategies are similar to logit

Test for Proportional odds

- A user-written command omodel can be used after downloading (type search omodel)

- The brant command performs a Brant test, which can be obtained by typing search spost

- The former provide overall test of proportionality odds for all independent variables and the later will provide specific test for each covariate

- Insignificant p-value revealed that the assumption is satisfied

```
omodel logit dependent independent1 independent2, ...
```

```
brant, detail
```

Multinomial Logistic Regression

- What if you want have a dependent variable with more than two outcomes?
    - A polytomous outcome
- Contrast outcomes with a common "reference point"
- Similar to conducting a series of 2-outcome logit models comparing pairs of categories
- The "reference category" is like the reference group when using dummy variables in regression
- It serves as the contrast point for all analyses
- Imagine a dependent variable with M categories
- Conduct binomial logit models for all possible combinations of outcomes
- This will produce results fairly similar to a multinomial output

Multinomial logistic regression

- Choose one category as “reference”
- Output will include two tables if the variable has three categories
- Choice of “reference” category drives interpretation of multinomial logit results
- Choose the contrast(s) that makes most sense
- Be aware of the reference category when interpreting results

Example: Mode of travel

- There are different mode of transport to travel for vacation (car, bus, and plane). We want to model the probability of the choice of transport mode
- The independent variables are income and family size
- If the outcome variable has k number of categories, then we expect $k-1$ models (estimates)
- This example, the outcome variable (mode of transport) has 3 categories, then the *mlogit* command in Stata will give two estimates for each covariate

Parameter estimation:

```
mlogit mode income familysize, base(3) rrr

Multinomial logistic regression                   Number of obs   =        152
LR chi2(4)      =      42.63
Prob > chi2     =     0.0000
Log likelihood = -138.68742                       Pseudo R2       =     0.1332

------------------------------------------------------------------------------
mode |  Odds ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
car         |
income |    .9444061   .0118194    -4.57   0.000     .9215224    .9678581
familysize |    .8204706   .1632009    -0.99   0.320     .5555836    1.211648
-------------+----------------------------------------------------------------
Bus           |
income |    .9743237   .0136232    -1.86   0.063     .9479852    1.001394
familysize |    .4185063   .1370806    -2.66   0.008     .2202385    .7952627
------------------------------------------------------------------------------
(mode==plane is the base outcome)
```

```
mlogit mode income familysize, base(3) rrr
```

Interpretation

- Holding family size constant, a unit increase in income reduces the odds of traveling by car rather than plane by 0.944 (decreases by 5.6%)
- Holding income constant, a unit increase in family size reduces the odds of traveling by bus rather than plane by 0.418 (decreases by 59.2%)

Assignment 3

- Consider the maternal depression data. The outcome variable is depression among pregnant mothers.
- Explore the data using chi-square and odds ratio
- Fit Simple binary logistic regression for each of the independent variables and compute crude odds ratio
- Fit Multiple logistic regression and compute adjusted odds ratio

Assignment 3

- Compare the crude and adjusted odds ratio and explain the difference
- Check the goodness of fit of the model
- Interpret all the results

Count Data Analysis

Poisson Regression Model

The Poisson Regression Model

- Count data are very common in many applications.
- Examples include:
  - Number of patients visiting a certain hospital per day,
  - CD4 counts,
  - Number of live births in a given district per year, etc.
- Count data are commonly analyzed using Poisson regression model.

The Poisson Regression Model Cont'd

- The Poisson model assumes that y_i, given the vector of covariates x_i, is independently Poisson-distributed with

$$f(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}.$$

- The mean is given by $\mu = \lambda$ and the variance, $\mathrm{var}(\mu) = \lambda$.
- Suppose $Y_1, \ldots, Y_N$ is a set of independent count outcomes, and let $x_1, \ldots, x_N$ represent the corresponding p-dimensional vectors of covariate values.
- The Poisson regression model with β a vector of p fixed, unknown regression coefficients is given by $\log(\lambda_i) = x_i{}'\beta$.

Jimma Infant Data

- Outcome: Total child deaths (‘deaths’).
- Factor1: place of residence $(1 = `urban', 2 = `semi\text{-}urban', 3 = `rural')$.
- Factor2: monthly family income,‘faminc’ in birr (continuous).

$$\log(\lambda) = \beta_0 + \beta_1 SU + \beta_2 R + \beta_3 IN.$$

‘IN’ refers to monthly family income, ‘SU’ and ‘R’ are defined as before.

```
Poisson regression                           Number of obs    =      7556
LR chi2(3)      =     203.15
Prob > chi2     =     0.0000
Log likelihood = -6849.1816                  Pseudo R2        =    0.0146

------------------------------------------------------------------------
deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------
place |
2   |     .035657   .0560403     0.64   0.525      -.07418      .145494
3   |    .2299316   .0500235     4.60   0.000     .1318874     .3279758
|
faminc |  -.0013383   .0001642    -8.15   0.000    -.0016601   -.0010164
_cons |  -.7445872   .0510623   -14.58   0.000    -.8446676   -.6445069
------------------------------------------------------------------------
```

$$\log(\lambda) = -0.745 + 0.036SU + 0.230R - 0.001IN.$$

STATA CODE:

```
poisson deaths i.place  faminc
```

The option ‘irr’ displays the incidence rate ratio estimates.

```
Poisson regression                          Number of obs   =       7556
LR chi2(3)      =     203.15
Prob > chi2     =     0.0000
Log likelihood = -6849.1816                 Pseudo R2       =     0.0146

------------------------------------------------------------------------
deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------
place |
2  |     1.0363   .0580746     0.64   0.525     .9285045    1.156611
3  |   1.258514   .0629552     4.60   0.000      1.14098    1.388155
|
faminc |   .9986626    .000164    -8.15   0.000     .9983413    .9989841
_cons |   .4749303    .024251   -14.58   0.000     .4297002    .5249213
------------------------------------------------------------------------
```

STATA CODE:

```
poisson deaths i.place faminc, irr
```

- $\exp(\beta_i)$ measures the change in the expected log counts (incidence rate).

- $\exp(\beta_2) = \exp(0.230) = 1.25$ implies that the expected log counts of child death in rural is higher by 1.25 as compared to that of urban, or the incidence rate of child death in rural is higher by nearly 25% as compared to that of urban.

- $\exp(\beta_3) = \exp(-0.0013) = 0.9987$ implies that as family income increases by one Birr, the incidence rate of child death decreases by nearly 0.13%.

Negative-Binomial Regression I

- Recall: Poisson distribution assumes that its ‘mean’ and ‘variance’ are equal.
- However, count data in practice violate this assumption as a result of unobserved heterogeneity.
- Usually, the sample variance is higher than the sample mean, a phenomenon referred to as ‘overdispersion’.
- The Overdispersion in the data can’t be fully captured by observed covariates.
- Hence, there is a need of models having more parameters than the Poisson distribution.
- Negative-binomial is an important extension to Poisson in this regard.

Negative-Binomial Regression II

- Negtive-binomial distribution is given by

$$P(y) = \frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})y!}\Big(\frac{\alpha\mu}{1+\alpha\mu}\Big)^{y}\Big(\frac{1}{1+\alpha\mu}\Big)^{1/\alpha}.$$

- $E(y) = \mu$, $\text{Var}(y) = \mu + \alpha\mu^2$.
- For regression purpose, we assume

$$y_i \sim \text{Negbin}(\mu_i, \alpha).$$

- Applying a log link,

$$\log\mu_i = {x_i}^T\beta.$$

Jimma Infant Data

- Outcome: Total child deaths (‘deaths’).
- Factor1: place of residence $(1 = `urban', 2 = `semi\text{-}urban', 3 = `rural')$.
- Factor2: monthly family income,‘faminc’ in birr (continuous).

Summary statistics of the outcome (‘deaths’):

```
Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------------------------------------------------------------
deaths   |      8040     .460199    .7191223        0          2
```

- Sample variance is higher than sample mean, i.e., suggesting evidence of overdispersion.
- Negative-binomial model might be a good idea instead of the Poisson model which was fitted earlier to these data.

```
Negative binomial regression                    Number of obs   =       7556
LR chi2(3)      =     176.10
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -6824.794                      Pseudo R2       =     0.0127

--------------------------------------------------------------------------
deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------------------------------------------------------------
place |
2   |   .0352622   .0595191     0.59   0.554    -.0813931    .1519174
3   |   .2328639   .0532109     4.38   0.000     .1285724    .3371554
|
faminc |  -.0013315   .0001712    -7.78   0.000    -.0016671   -.0009959
_cons |  -.7470684   .0540689   -13.82   0.000    -.8530415   -.6410953
--------------------------------------------------------------------------
/lnalpha |  -1.124534   .1619651                      -1.441979    -.807088
--------------------------------------------------------------------------
alpha |   .3248039   .0526069                      .2364592    .4461554
--------------------------------------------------------------------------
Likelihood-ratio test of alpha=0 chibar2(01)   48.78 Prob>=chibar2 = 0.000
```

Observations ...

STATA CODE:

```
nbreg deaths i.place faminc
```

- The dispersion parameter *alpha* now in the Negative-binomial model captures the extra variability in the data.
- The Negative-binomial model is an improvement in model fit over the Possion model.
- Furthermore, likelihood ratio test of the parameter *alpha* with ($p_{value} < 0.001$) confirms the above result.

Zero-Inflated Models I

- Most count data are characterized by excessive zeros beyond what the common count distributions can predict.
- Often because of heterogeneity between subjects.
- Or omission of important covariances
- Not appropriately accounting for this feature leads to biased estimates.
- And hence erroneous conclusion.
- Hence, models for such extension are needed.

Zero-Inflated Models II

Commonly used models in this regard:

- Zero-inflated Poisson model (ZIP).
- Zero-inflated negative-binomial (ZINB).
- ZINB is considered when data are characterized by both overdispersion and zero-inflation.

Zero-Inflated Models III

- Zero-inflated Poisson model (ZIP)is commonly used models in this regard:
- This model allows for overdispersion assuming that there are two different types of individuals in the data:
- (1) Those who have a zero count with a probability of 1 (Always-0 group), and
- (2) those who have counts predicted by the standard Poisson (Not always-0 group)
- Observed zero could be from either group, and if the zero is from the Always-0 group, it indicates that the observation is free from the probability of having a positive outcome

Zero-Inflated Poisson Model

- Two processes:
  - The first process generates only zeros with probability, say π_i for observation i,
  - And the second process generates counts with probability, say $(1 - \pi_i)$
  - In a zero-inflated model, Y_i follows a zero-inflation probability distribution given by

$$p(Y_i = y_i) = \begin{cases} \pi_i + (1 - \pi_i) f_i(0|\lambda_i) & \text{if } y_i = 0, \\ (1 - \pi_i) f_i(y_i|\lambda_i) & \text{if } y_i > 0. \end{cases}$$

Jimma Infant Data: Consider the variable ‘deaths’.

```
total child |
deaths |      Freq.     Percent        Cum.
------------+-----------------------------------
0 |      5,420       67.41       67.41
1 |      1,540       19.15       86.57
2 |      1,080       13.43      100.00
------------+-----------------------------------
Total |      8,040      100.00
```

Nearly 67% are zeros, implying presence of excessive zero observations.

STATA CODE:

```
tab1 deaths
```

Zero-Inflated Poisson for Jimma Data

- Consider ‘deaths’ as outcome.
- ‘Place’ and ‘family income’ as covariates both in positive counts and zero-inflation parts.

```
Zero-inflated Poisson regression                  Number of obs    =      7556
Nonzero obs      =       2500
Zero obs         =       5056

Inflation model = logit                           LR chi2(3)       =     12.95
Log likelihood  =  -6733.22                       Prob > chi2      =    0.0047

------------------------------------------------------------------------------
deaths |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
deaths       |
place |
2  |  -.0466848   .1017143    -0.46   0.646    -.2460411    .1526715
3  |   .0721294   .0830098     0.87   0.385    -.0905668    .2348256
|
faminc |  -.0007839   .0002924    -2.68   0.007     -.001357   -.0002109
_cons |  -.2563935   .0827636    -3.10   0.002    -.4186072   -.0941797
-------------+----------------------------------------------------------------
inflate      |
place |
2  |  -.2032598   .2185156    -0.93   0.352    -.6315425    .2250229
3  |  -.4668808   .1817955    -2.57   0.010    -.8231934   -.1105682
|
faminc |   .0010883   .0005015     2.17   0.030     .0001054    .0020713
_cons |  -.4231007   .1668808    -2.54   0.011    -.7501811   -.0960203
------------------------------------------------------------------------------
Vuong test of zip vs. standard Poisson:           z =     9.43  Pr>z = 0.0000
```

STATA CODE:

```
zip deaths i.place faminc, inflate(i.place faminc) vuong
```

The option ‘vuong’ gives test for zero-inflation.

Observations

- Log-likelihood: Poisson model$= -6849.1816$, ZI-Poisson $=-6733.22$ → considerable improvement in model fit.
- Vuong test suggests significant zero-inflation
- The intercept and family income are significant both in the positive counts and zero-inflation components.
- Interpretation of zero-in ation model coefficients: Being a rural resident reduces the odds of not having infant death by 38% (exp(-.4668808)=0.626954819), and is statistically significant
- Interpretation of count model coefficients: Among those who have died infant, being rural increases the expected rate of death by 7.5 % (exp(.0721294)=1.074794413)), holding other variables constant, and this is not statistically significant.
- Family income affects the two parts oppositely.

Chapter 4

Survival Analysis

Introduction

- Survival Analysis refers to statistical methods for analyzing survival data.
- Survival data could be derived from laboratory, clinical, epidemiological studies, etc.
- Response of interest is the time from an initial observation until occurrence of a subsequent event.

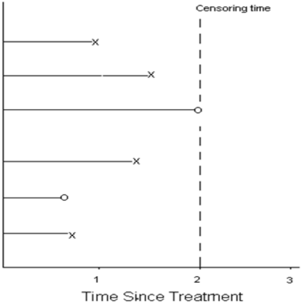
Examples

- In many biomedical studies, the primary endpoint is time until an event occurs
- The event of interest includes
    - Time to death, remission
    - Time to adverse events
    - Time to relapse
    - Time to recovery from diseases
    - Time to new symptom
    - Time to discharge
    - Time to start talk for a child etc
    - Time from transplant surgery until the new organ fails
    - Marriage duration
- This time interval from a starting point and a subsequent event is known as the survival time.

- Which of the following data sets is likely to lend itself to survival analysis?
- A case-control study of caffeine intake and breast cancer
  A randomized controlled trial where the outcome was whether or not women developed breast cancer in the study period
  A cohort study where the outcome was the time it took women to develop breast cancer.
  A cross-sectional study which identified both whether or not women have ever had breast cancer and their date of diagnosis.

Event and Censoring

- We may not observe all the event of interest within the defined follow-up period
- Thus, data are typically subject to censoring when a study ends before the event occurs
- Censoring: Subjects are said to be censored if they are lost to follow up or drop out of the study, or if the study ends before they die or have an outcome of interest



Jimma Infant Data

- Infants in Southwest Ethiopia were studied for one year.
- Measurements were taken approximately every two months.
- Nearly 8000 infants at baseline.
- Data on infant, maternal and household characteristics were collected.
- Death of infants within their first year is an important problem with these data.
- We will therefore analyze the time from birth to death (in days).
- For infants still alive when these data were collected, time is the time from birth to the time of data collection.

Jimma Infant Data II

- The variable event is an indicator for whether time refers to death (1) or end of study (0).
- Possible explanatory variables for time-to-death could be place of residence and sex of infants, among others.
- These data can be described as survival data.
- Duration or survival data can generally not be analyzed by conventional methods such as linear regression.
- The main reason for this is that some durations are usually right-censored.
- That is, the endpoint of interest has not occurred during the period of observation.
- Another reason is that survival times tend to have positively skewed distributions.

Survivor Function

- The survival time T may be regarded as a random variable with a probability distribution $F(t)$ and probability density function $f(t)$.

- An obvious quantity of interest is the probability of surviving to time t or beyond, the survivor function or survival curve $S(t)$, which is given by

$$S(t) = P(T \geq t) = 1 - F(t).$$

Hazard Function

- A further function which is of interest for survival data is the hazard function.
- This represents the instantaneous death rate, that is, the probability that an individual experiences the event of interest at a time point given that the event has not yet occurred.
- The hazard function is given by

$$h(t) = \frac{f(t)}{S(t)}.$$

- The instantaneous probability of death at time t divided by the probability of surviving up to time t.
- Hazard function is just the incidence rate.

Kaplan-Meier Estimator

- The Kaplan-Meier estimator is a nonparametric estimator of the survivor function $S(t)$.
- If all the failure times, or times at which the event occurs in the sample are ordered,
- And labeled $t_{(j)}$ such that $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(n)}$, the estimator is given by

$$\widehat{S(t)} = \prod_{j|t_{(j)} \leq t} (1 - \frac{d_j}{n_j}).$$

Kaplan-Meier Estimator II

- Recall

$$\widehat{S(t)} = \prod_{j|t_{(j)} \leq t} (1 - \frac{d_j}{n_j}).$$

- d_j is the number of individuals who experience the event at time $t_{(j)}$,
- n_j is the number of individuals who have not yet experienced the event at that time,
- And are therefore still ‘at risk’ of experiencing it (including those censored at $t_{(j)}$).

Jimma Infant Data

- Recall: the Jimma Infant Data.
- Before any analysis, we declare the data as being of the form ‘st’ (for survival time) using the ‘stset’ command.

```
stset duration, failure(event)
```

```
failure event:  event != 0 & event < .
obs. time interval:  (0, duration]
exit on or before:  failure

------------------------------------------------------------------
8050  total obs.
57  obs. end on or before enter()
------------------------------------------------------------------
7993  obs. remaining, representing
698  failures in single record/single failure data
2591422  total analysis time at risk, at risk from t =         0
earliest observed entry t =         0
last observed exit t =       464
```

Look at the summary of the data using:

```
stsum
```

The following output results:

```
failure _d:  event
analysis time _t:  duration

|               incidence   no. of    |------ Survival time -----|
| time at risk     rate    subjects        25%       50%       75%
-----------------------------------------------------------------------
total |    2591422    .000269       7993          .         .         .
```

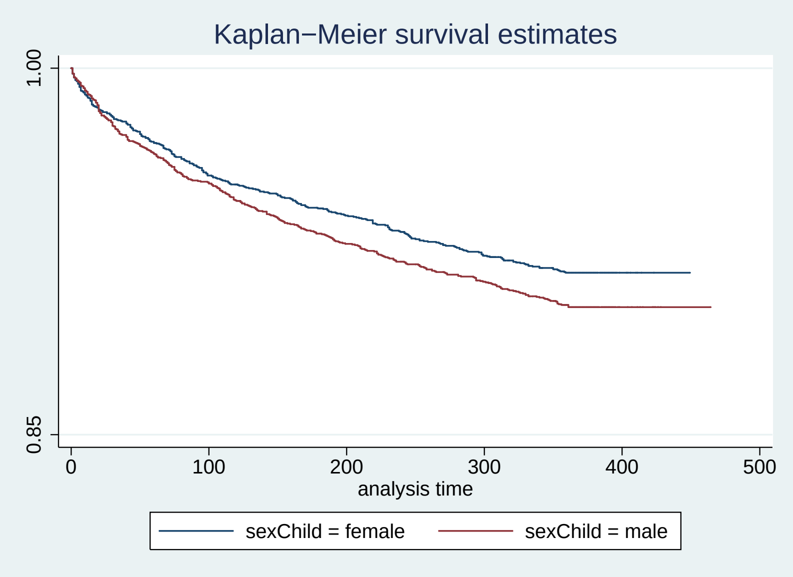
Note: There are 7993 subjects, and if the incidence rate (i.e., the hazard function) could be assumed to be constant, it would be estimated as 0.00027 per day which corresponds to 0.098 per year.

Life Table

STATA CODE:

```
. ltable duration event, survival intervals(30)
Beg.                                   Std.
Interval      Total  Deaths  Lost    Survival    Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
0    30      8050      216   102    0.9730    0.0018     0.9692    0.9763
30    60      7732      91     6     0.9615    0.0022     0.9571    0.9655
60    90      7635      82   116     0.9511    0.0024     0.9462    0.9557
90   120      7437      62     9     0.9432    0.0026     0.9379    0.9481
120   150      7366      42    96     0.9378    0.0027     0.9322    0.9429
150   180      7228      48     7     0.9316    0.0028     0.9258    0.9369
180   210      7173      35    81     0.9270    0.0029     0.9210    0.9325
210   240      7057      45    12     0.9211    0.0030     0.9149    0.9268
240   270      7000      34    72     0.9166    0.0031     0.9102    0.9225
270   300      6894      33    17     0.9122    0.0032     0.9057    0.9182
300   330      6844      30   150     0.9081    0.0033     0.9015    0.9143
330   360      6664      30   912     0.9037    0.0034     0.8970    0.9101
360   390      5722       1  5636     0.9034    0.0034     0.8966    0.9098
390   420        85       0    77     0.9034    0.0034     0.8966    0.9098
420   450         8       0     7     0.9034    0.0034     0.8966    0.9098
450   480         1       0     1     0.9034    0.0034     0.8966    0.9098
-------------------------------------------------------------------------------
```

The Kaplan-Meier estimator of the survivor functions for males and females.



STATA CODE:

```
sts graph, by(sexChild) ylabel(0.85(0.05)1.0)
```

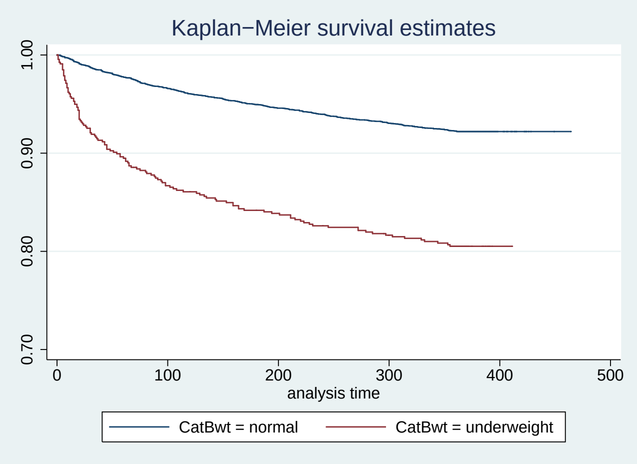
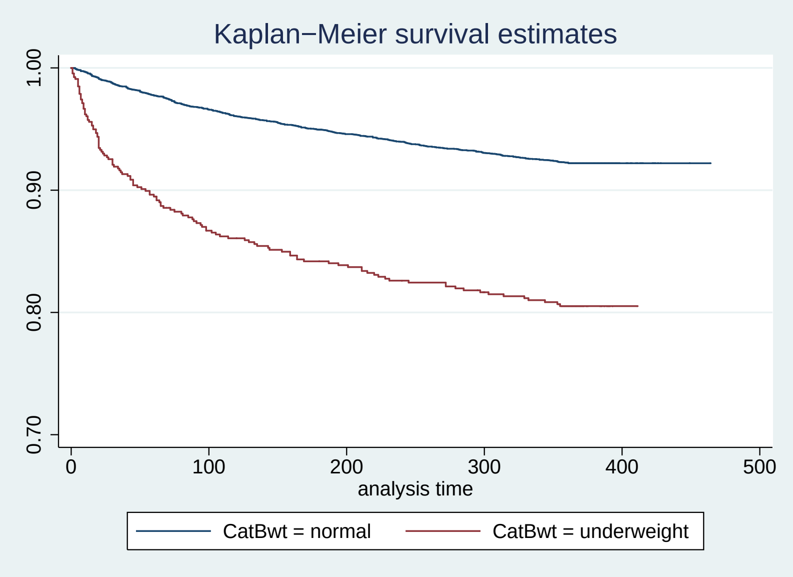


Figure 16: Leverage versus Normalized residual squared.

The Kaplan-Meier estimator of the survivor functions for normal and underweight infants.



STATA CODE:

```
sts graph, by(CatBwt) ylabel(0.7(0.1)1.0)
```

Log-rank Test

- Tests equality of survival functions.
- Stratified log-rank test shows within-stratum tests.

Jimma Infant Data

- Test equality of survival functions of underweight and normal infants.
- Also consider log-rank test with sexChild as stratification variable.

Log-rank test by sex

```
Log-rank test for equality of survivor functions

|   Events         Events
sexChild |  observed       expected
---------+-------------------------
female   |       318         345.01
male     |       380         352.99
---------+-------------------------
Total    |       698         698.00

chi2(1) =       4.18
Pr>chi2 =     0.0408
```

- There seems a marginal significance.

STATA CODE:

```
sts test sexChild, logrank
```

Log-rank test by Catbwt

```
Log-rank test for equality of survivor functions

|   Events         Events
CatBwt      |  observed       expected
------------+-------------------------
normal      |       537         611.52
underweight |       126          51.48
------------+-------------------------
Total       |       663         663.00

chi2(1) =     117.03
Pr>chi2 =     0.0000
```

- Survival functions are unequal.

STATA CODE:

```
sts test CatBwt, logrank
```

Stratified by sexChild...

```
Stratified log-rank test for equality of survivor functions

|   Events         Events
CatBwt      |  observed    expected(*)
------------+-------------------------
normal      |       537        612.58
underweight |       126         50.42
------------+-------------------------
Total       |       663        663.00

(*) sum over calculations within sexChild

chi2(1) =     123.26
Pr>chi2 =     0.0000
```

- Survival functions are unequal.

STATA CODE:

```
sts test CatBwt, logrank strata(sexChild)
```

Cox Proportional Hazards Model

- Kaplan-Meier and significance tests (e.g. Log-rank) can be used to compare survival in different subgroups.
- However, when there are several explanatory variables (when some of these are continuous) a regression method such as Cox regression is preferred.
- The hazard function for individual i is modeled as

$$h_i(t) = h_0(t)\exp(\beta^T x_i).$$

- $h_0(t)$ is the baseline hazard function.
- β are regression coefficients.
- x_i are covariates of interest.

Cox Proportional Hazards Model II

- The baseline hazard is the hazard when all covariates are zero, and is left unspecified.

- And this idea combined with a parametric representation of the effects of covariates gives rise to the term semiparametric.

- The model assumes the hazard of any individual i is a time-constant multiple of the hazard function of any other individual j, the factor being $\exp(\beta^T(x_i - x_j))$.
- This property is called the proportional hazards assumption.
- The exponentiated regression parameters can therefore be interpreted as hazard ratios.
- Stratified Cox regression can be used to relax the assumption of proportional hazards for a categorical predictor.

Jimma Infant Data

- Outcome: Duration.
- Factor: CatBwt (1 = underweight, 0 = normal).
- Stratification variable: sexChild (1 = male, 0 = female).

```
Cox regression -- Breslow method for ties

No. of subjects =         7867                     Number of obs   =      7867
No. of failures =          663
Time at risk    =      2558581
LR chi2(1)      =     86.02
Log likelihood  =   -5849.1468                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
_t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
CatBwt |    2.78779   .2759798    10.36   0.000      2.29612    3.384742
------------------------------------------------------------------------------
```

- Underweight infants are nearly 3 times more likely to die at any given time (given that they remained in the study until that time) as compared to that of normal.

STATA CODE:

```
stcox CatBwt
```

Jimma Infant Data

Two factors...

- Outcome: Duration.
- Factor1: CatBwt ($1 =$ underweight, $0 =$ normal).
- Factor2: sexChild ($1 =$ male, $0 =$ female).
- Stratification variable: site ($1 =$ urban, $2 =$ semi-urban, $3 =$ rural).

```
Cox regression -- Breslow method for ties

No. of subjects =         7867                     Number of obs    =      7867
No. of failures =          663
Time at risk    =      2558581
LR chi2(2)      =     95.39
Log likelihood  =   -5844.4616                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
_t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.CatBwt |   2.873672   .2859198    10.61   0.000     2.364535    3.492437
1.sexChild |   1.270629   .0997042     3.05   0.002     1.089498    1.481874
------------------------------------------------------------------------------
```

- Male infants are 27% more likely to die at any given time (given that they remained in the study until that time) than that of females.

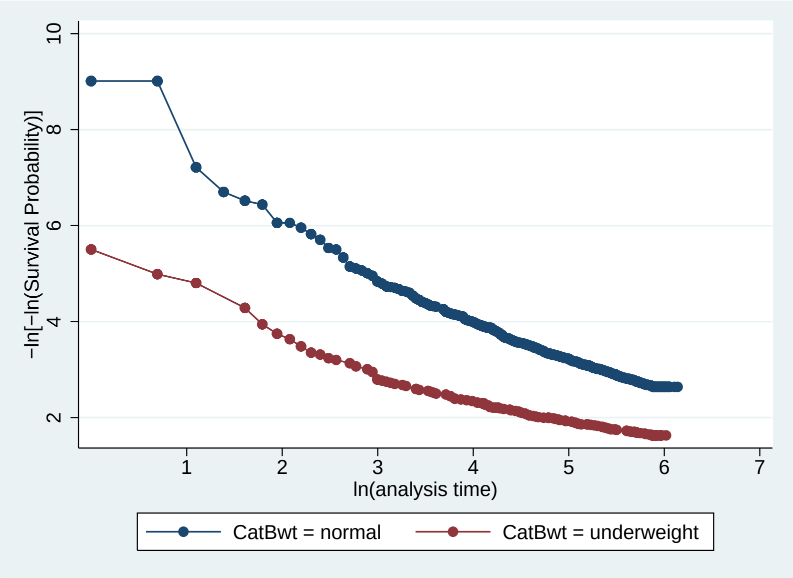
STATA CODE:

```
stcox i.CatBwt i.sexChild
```

Assessing Proportionality Assumption

- A graphical approach is available for categorical predictors if there are sufficient observations for each value of the predictor
- In this case the model is first estimated by stratifying on the categorical predictor of interest
- Thus not making any assumption regarding the relationship between the baseline hazards for different values of the predictor or strata
- The log cumulative baseline hazards for the strata are then derived from the estimated model and plotted against time
- The resulting curves should be parallel if the proportional hazards assumption holds

Jimma Infant Data ...



- What do you conclude?

STATA CODE:

```
stphplot, strata(CatBwt) adjust(sexChild)/*
*/zero xlabel(1/7)
```

Jimma Infant Data

- Consider a continuous variable-‘wt (kg)’
- Sex as a stratification variable
- Fit cox regression

```
Cox regression -- Breslow method for ties
No. of subjects =         7867                    Number of obs   =      7867
No. of failures =          663
Time at risk    =      2558581
LR chi2(5)      =     98.81
Log likelihood  =   -5842.7512                    Prob > chi2     =    0.0000

-----------------------------------------------------------------------------
_t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
1.sexChild |    1.26778   .0994933     3.02   0.002     1.087034     1.47858
|
CatGravida |
2  |   .9121891   .0982528    -0.85   0.394     .7385861    1.126597
3  |   .8681939   .1307178    -0.94   0.348     .6463346    1.166208
|
1.CatBwt |    2.83658   .2839624    10.41   0.000     2.331222    3.451489
momage |   .9961322   .0089392    -0.43   0.666     .9787648    1.013808
-----------------------------------------------------------------------------
```

- The hazard of death increased by 27% in male as compared to female.

Test of PH Assumption

```
 estat phtest
Test of proportional-hazards assumption

Time:  Time
----------------------------------------------------------------
|                      chi2       df       Prob>chi2
------------+---------------------------------------------------
global test |                     28.48        5         0.0000
----------------------------------------------------------------
```

- This indict that PH assumption is violated (p-value$<$0.05)
- We can refit the model by removing one variable at a time

Test of PH Assumption

- When we remove weight from the model, the test becomes

```
. estat phtest

Test of proportional-hazards assumption

Time:  Time
----------------------------------------------------------------
|                      chi2       df       Prob>chi2
------------+---------------------------------------------------
global test |                      8.62        4         0.0712
----------------------------------------------------------------
```

Time-varying Covariates

- A covariate could be time-dependent
- Is one possible route to evaluate proportionality
- It could also be an interaction term

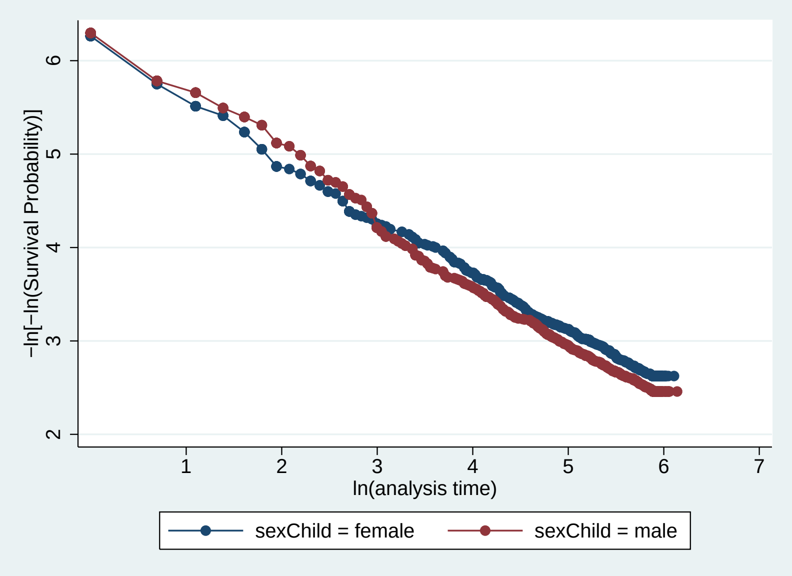
Time Dependent Cox-regression

-

Jimma Infant Data

- Consider birth weight of infant (continuous)
- Consider weight-time interaction
- This can be considered as a time-varying covariate.
- Likelihood need to be evaluated for values of the time-varying covariates at the times of the death
- These values are not available for the denominator since each subject is represented only once
- There is a need to create the required dataset

Recall: Jimma Infant Data ...



Parametric Survival Models

- Recall: Cox Proportional Hazards Model.
- Distributional form of $h_0(t)$ is left unspecified.
- Hence, the Cox Proportional Hazards Model is a semi-parametric approach.
- Interest is mainly in the proportional factors rather than the baseline hazard.
- However, the underlying proportionality assumption may not be valid in practice.
- A fully parametric modeling approach may be needed.

Parametric Models...

- Exponential Survival Model
- Weibull Survival Model
- Weibull-gamma,
- etc

Computing Risk Model

-

Part B

Models for Longitudinal/Cluster Data

Chapter 5

Longitudinal Data Analysis

Longitudinal Data

- A longitudinal study refers to an investigation where participant outcomes and possibly treatments or exposures are collected at multiple follow-up times
- Repeated Measures data / Clustered data
- 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

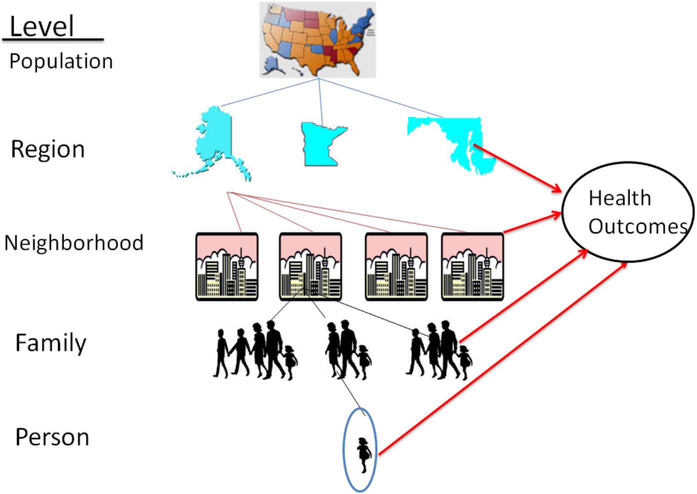
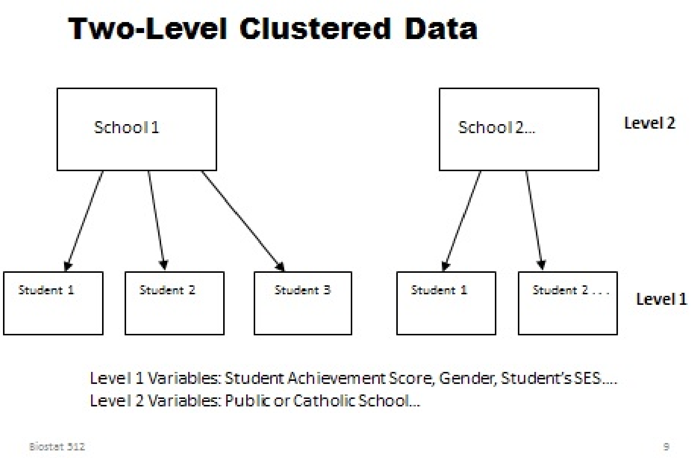


Figure 17: Multilevel Data Structure.

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

School 1

School 2...

Level 2

Student 1

Student 2

Student 3

Student 1

Student 2 . . .

Level 1

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

Level 2 Variables: Public or Catholic School...

Biostat 512

9

Figure 18: Two Level cluster data.

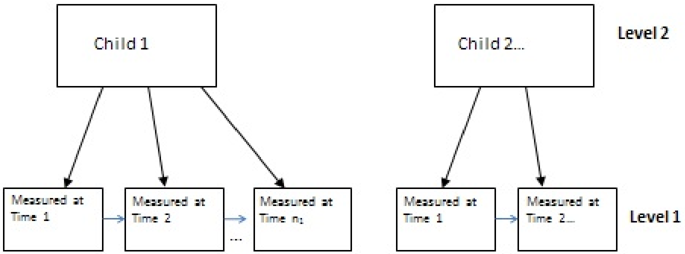
Longitudinal data

- 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)



Child 1

Child 2...

Level 2

Measured at Time 1

Measured at Time 2

...

Measured at Time n_i

Measured at Time 1

Measured at Time 2...

Level 1

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

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

Figure 19: Two Level longitudinal 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

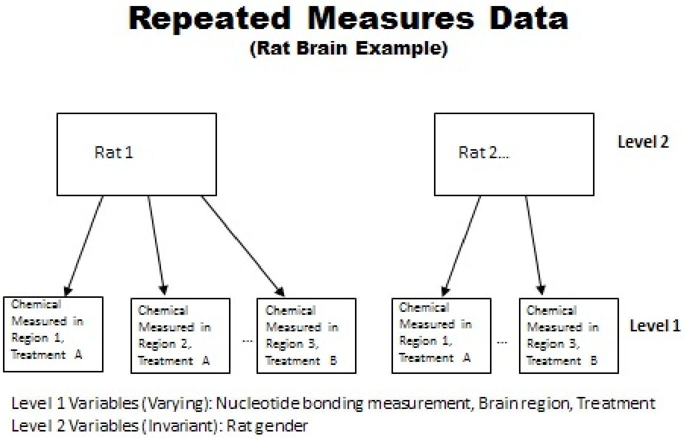

Repeated Measures Data
(Rat Brain Example)
Rat 1
Rat 2...
Level 2
Chemical Measured in Region 1, Treatment A
Chemical Measured in Region 2, Treatment A
...
Chemical Measured in Region 3, Treatment B
Chemical Measured in Region 1, Treatment A
...
Chemical Measured in Region 3, Treatment B
Level 1
Level 1 Variables (Varying): Nucleotide bonding measurement, Brain region, Treatment
Level 2 Variables (Invariant): Rat gender

Figure 20: Two Level longitudinal data.

Examples

- 1. A research study in education aims to assess the impact of school type (public vs. catholic) as well as student gender on student-level math achievement scores. Scores are measured once for the students in the school.
- 2. Researchers are studying the effect of two different treatments on nucleotide bonding in three regions of the brain in rats. Measurements are taken from the same three regions of the brain of each rat, after each of the two different treatments

Examples

- In all the examples, data are often hierarchical in nature, and we should not ignore this
- Using single-level (OLS, GLM) analysis leads to: Unit of analysis problem, aggregation bias
- Incorrectly estimated precision / standard errors
- Results in incorrect p-values and incorrect conclusions
- Need Multilevel Models

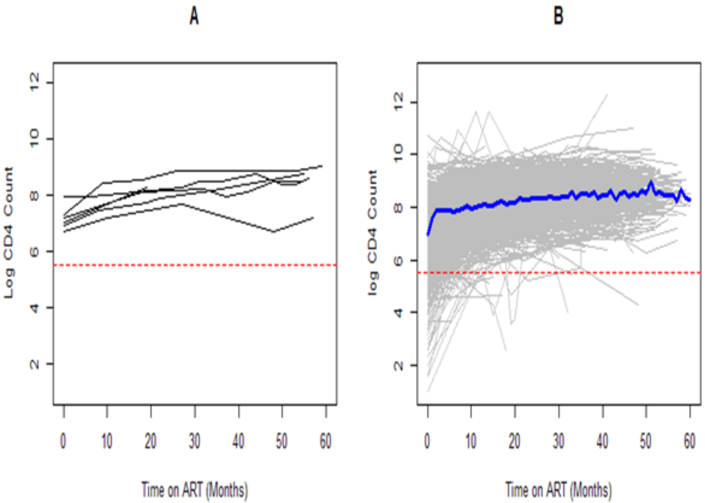


Figure 21: Subject specific profiles of CD4 cell counts.

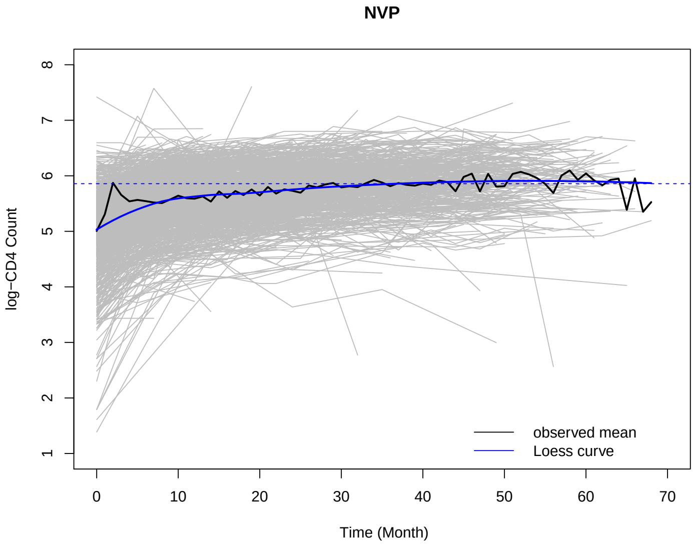


Figure 22: Subject specific profiles of CD4 cell counts.

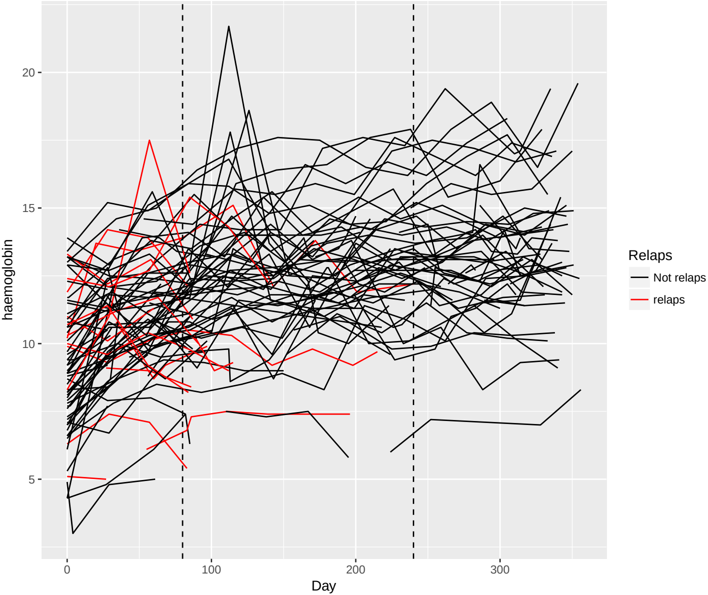


Figure 23: Subject specific profiles of Hemoglobine.

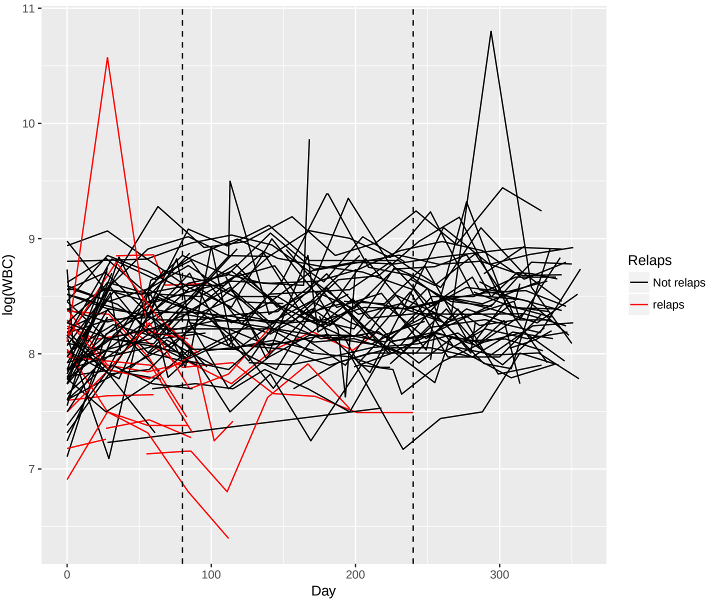


Figure 24: Subject specific profiles of WBC.

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

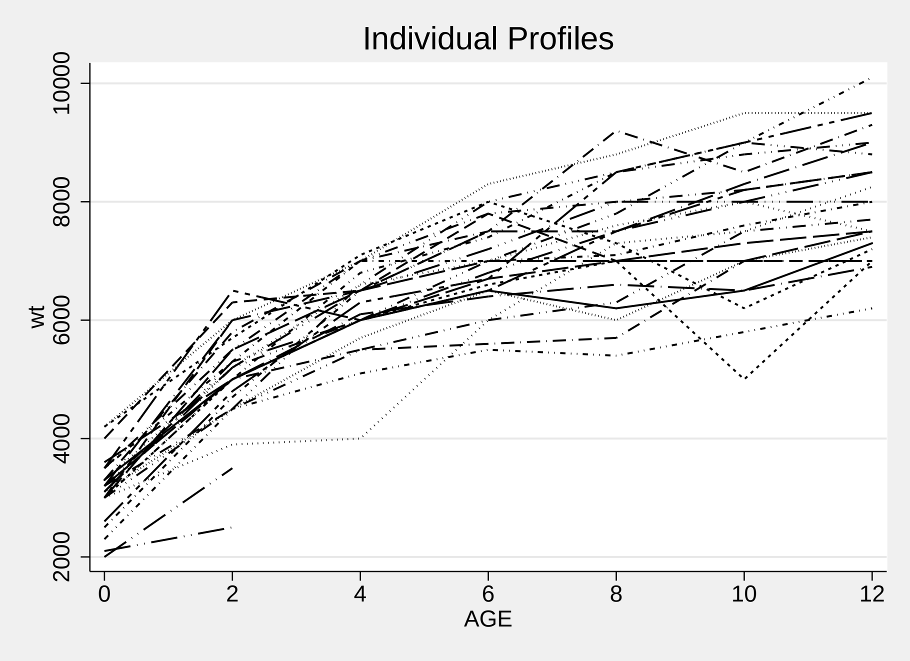
- Random effects are used to accommodate the heterogeneity in the data, which may arise from subject or clustering effects or from spatial correlation

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?

Part of the data ...

Obs	child	age(months)	weight(grams)	sex
1	1	0	2900	1
2	1	2	3100	1
3	1	4	3180	1
.	.	.	.	.
7	1	12	8000	1
8	2	0	3200	0
9	2	2	3340	0
.	.	.	.	.
25	3	0	3015	1
26	3	2	3200	1
.	.	.	.	.


Individual Profiles
wt
2000
4000
6000
8000
10000
0
2
4
6
8
10
12
AGE

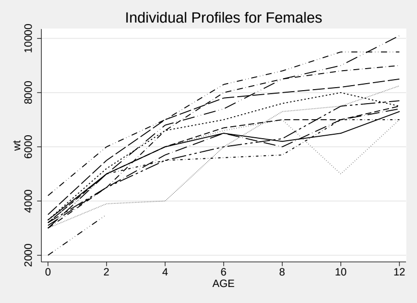
Individual profiles for the first 30 subjects ...

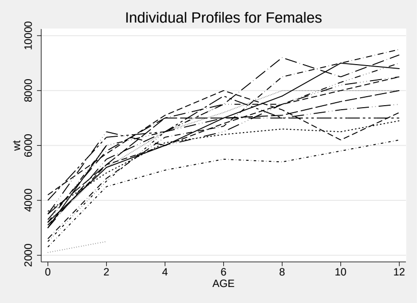
STATA CODE:

```
gen wt=weight if ind<=30
xtline wt, overlay t(age) tlabel(#6) i(ind) legend(off) /*
*/ title("Individual Profiles") scheme(s2mono)
```

Remarks:

- Subjects high (low) at baseline seem to remain high (low) over time
- Much variability within subjects
- Much variability between subjects
- The variability between subjects at higher ages is relatively larger than that of the baseline


Individual Profiles for Females
wt
2000
4000
6000
8000
10000
0
2
4
6
8
10
12
AGE


Individual Profiles for Females
wt
2000
4000
6000
8000
10000
0
2
4
6
8
10
12
AGE

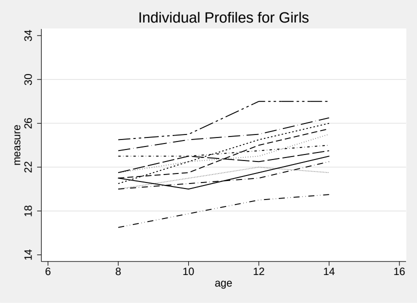
Individual profiles for the first 30 subjects ...

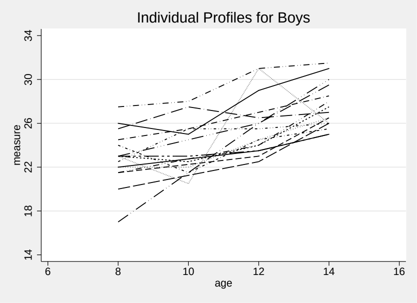
STATA CODE:

```
xtline wt if sex==0, overlay t(age) tlabel(#6) i(ind) legend(off)/*
/* title("Individual Profiles for Females") scheme(s2mono)
xtline wt if sex==1, overlay t(age) tlabel(#6) i(ind) legend(off)/*
*/ title("Individual Profiles for Males") scheme(s2mono)
```

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?


Individual Profiles for Girls
measure
34
30
26
22
18
14
6
8
10
12
14
16
age


Individual Profiles for Boys
measure
34
30
26
22
18
14
6
8
10
12
14
16
age

Individual profiles ...

STATA CODE:

```
xtline measure if sex==2, overlay t(age) tlabel(6(2)16) i(ind) legend(off)/*
/* title("Individual Profiles for Girls") scheme(s2mono)  ylabel(14(4)34)

xtline measure  if sex==1, overlay t(age) tlabel(6(2)16) i(ind) legend(off)/*
*/ title("Individual Profiles for Boys") scheme(s2mono)   ylabel(14(4)34)
```

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

Longitudinal versus Cross-sectional Data

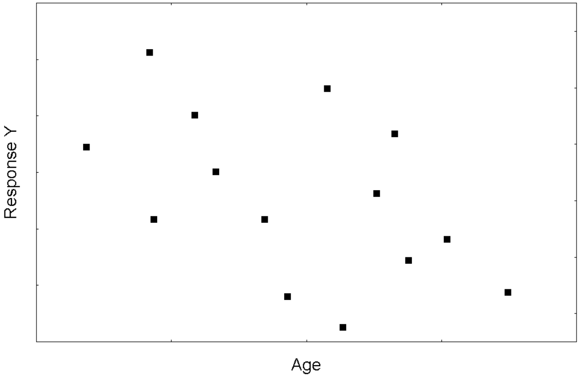
- Recall: Longitudinal data refers to measurements made repeatedly over time to study how the subjects evolve over time

- 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

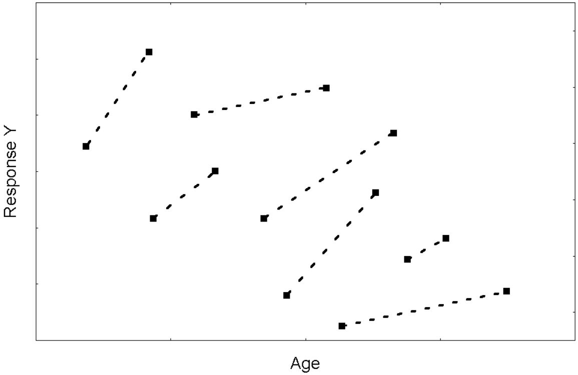
- Suppose it is of interest to study the relation between some response Y and age
- A cross-sectional study yields the following data:



- The graph suggests a negative relation between Y and age

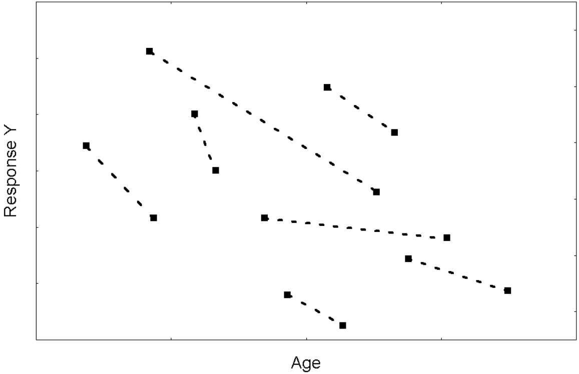
- Exactly the same observations could also have been obtained in a longitudinal study, with 2 measurements per subject as shown below:

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

Simple Methods

- Analysis at each time point separately
- Analysis of endpoints
- Analysis of increments
- The above methods have limitations such as:
  - Does not consider ‘overall’ differences
  - Does not allow to study evolution differences
  - Problem of multiple testing

A model for Longitudinal Data

- Linear Mixed Model (LMM) is used to analyze repeated continuous data
- LMM contains both fixed and random effects
- Fixed effects: the only levels under consideration are contained in the coding of those effects
- Random effects: the levels contained in the coding of those factors are a random sample of the total number of levels
- LMM account for the correlation in the data by including subject specific random effects
- These random effects are usually of a Gaussian type

General linear mixed-effects model

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

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

A linear Mixed Model makes assumptions about:

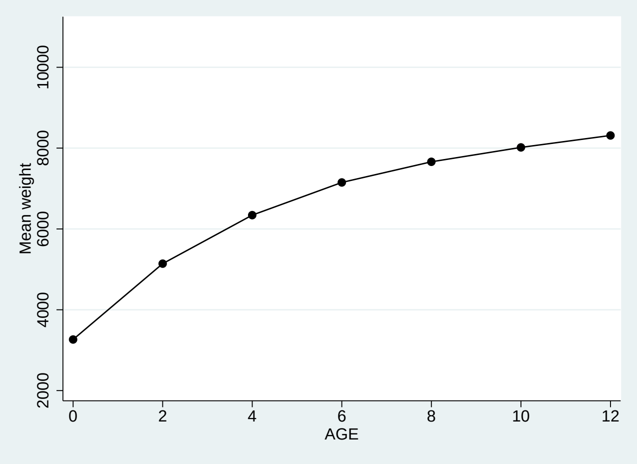
- mean structure: (non-)linear, covariates,. . .
- variance function: constant, quadratic, . . .
- correlation structure: constant, serial, . . .
- subject-specific profiles: linear, quadratic,

Exploratory Analysis

- It comprises techniques to visualize patterns in the data: usually graphically
- The following aspects of the data will be looked:
  - individual profiles
  - average evolution
  - correlation structure

Jimma Infant Data

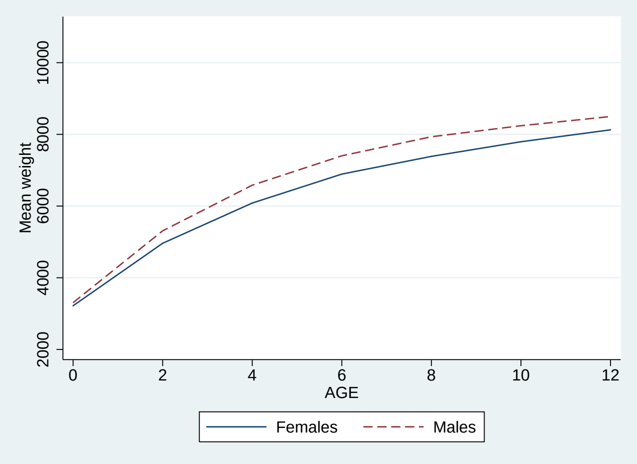
- Average profile for both:



STATA CODE:

```
label define group 0 "Female" 1 "Male"
label values sex group
collapse (mean) weight (sd) sdweight=weight /*
/* (count) n=weight, by(age)
twoway (connected weight age, mcolor(black) /*
/* clcolor(black)), ytitle("Mean weight")/*
*/ ylabel(2000(2000)11000) xlabel(0(2)12)
```

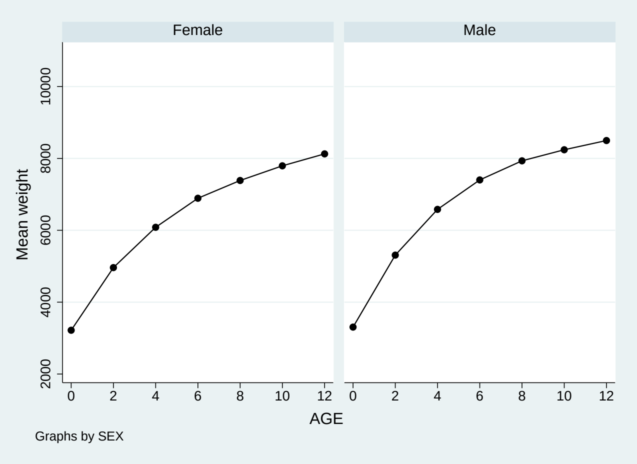
- Average profile separately (one plot):



STATA CODE:

```
label define group 0 "Female" 1 "Male"
label values sex group
collapse (mean) weight (sd) sdweight=weight /*
/* (count) n=weight, by(age sex)
twoway (line weight age if sex==0)(line weight age if sex==1,clpat(dash)),/*
*/legend(order(1 "Females" 2 "Males")) ytitle("Mean weight")/*
*/ ylabel(2000(2000)11000) xlabel(0(2)12)
```

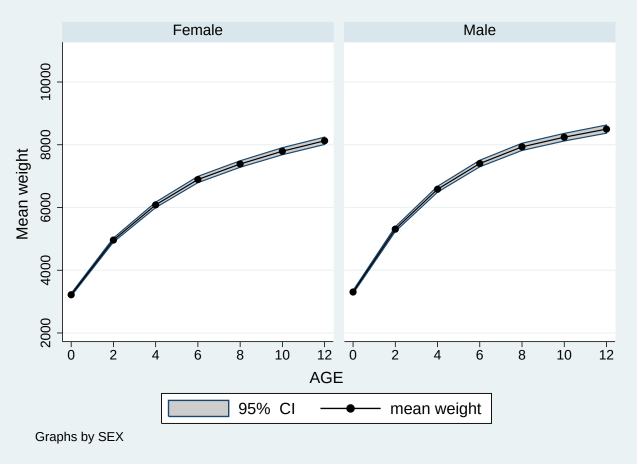
- Average profile separately (separate plots):



STATA CODE:

```
label define group 0 "Female" 1 "Male"
label values sex group
collapse (mean) weight (sd) sdweight=weight /*
/* (count) n=weight, by(age sex)
twoway (connected weight age, mcolor(black) /*
/* clcolor(black)), by(sex) ytitle("Mean weight")/*
*/ ylabel(2000(2000)11000) xlabel(0(2)12)
```

- With confidence regions:



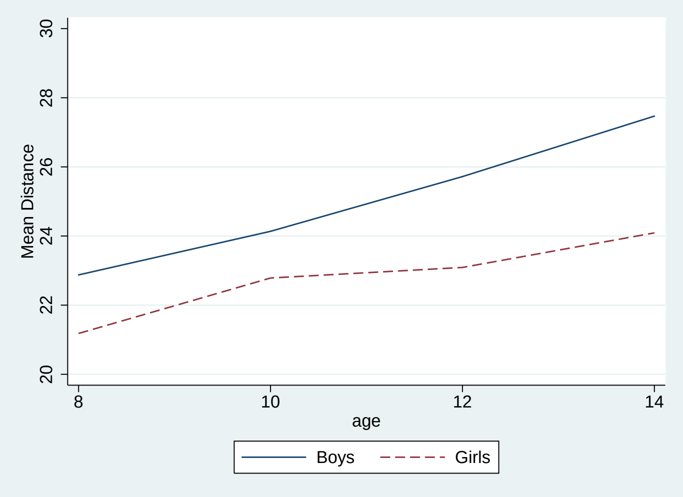
STATA CODE:

```
label define group 0 "Female" 1 "Male"
label values sex group
collapse (mean) weight (sd) sdweight=weight /*
/*(count) n=weight, by(age sex)
gen high =weight + 2*sdweight/sqrt(n)
gen low = weight - 2*sdweight/sqrt(n)
twoway (rarea low high age, bfcolor(gs12)) /*
/* (connected weight age, mcolor(black) /*
/* clcolor(black)), by(sex) legend(order(1 "95% /*
/* CI" 2 "mean weight")) ytitle("Mean weight")/*
*/ ylabel(2000(2000)11000) xlabel(0(2 )12)
```

- 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.

Growth Data

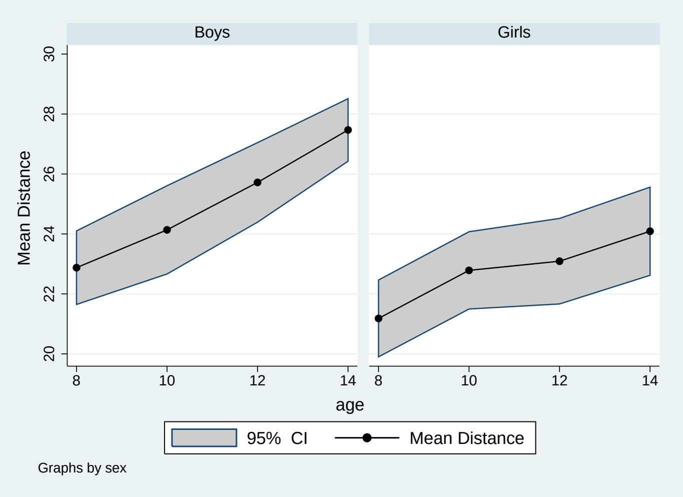
- Average profiles by sex:



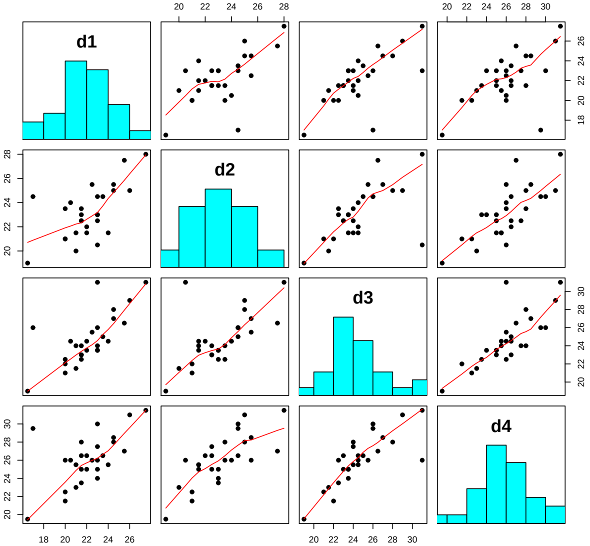
STATA CODE:

```
label define group 1 "Boys" 2 "Girls"
label values sex group
collapse (mean) measure (sd) sdmeasure=measure /*
/*(count) n=measure, by(age sex)
twoway (line measure age if sex==1)(line measure age/*
*/ if sex==2,clpat(dash)),legend(order(1 "Boys" 2 "Girls"))/*
*/ ytitle("Mean Distance") ylabel(20(2)30) xlabel(8(2)14)
```

- Average profiles by sex and CI:



- Correlation: Scatter Plot Matrix



Estimation

- Restricted Maximum Likelihood (REML)
- Maximum Likelihood (ML)
- ML is the default in STATA

Recall:

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

Marginally:

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

- In REML, we transform Y so that the mean vanishes from the likelihood
- Note that Likelihood at convergence of REML is NOT the likelihood for the original data Y
- And hence can not be considered for comparison of models.

Jimma Infant Data

- From the exploratory analysis
  - mean structure seems quadratic over time
  - variability between subjects at baseline
  - variability between subjects in the way they evolve
- Hence a quadratic mean, with random intercept and slope is a good idea...

Model:

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

- 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

Results (ML):

```
Mixed-effects ML regression      Number of obs      = 6113
Group variable: ind             Number of groups    = 1000

Obs per group: min = 1
avg = 6.1
max = 7


Wald chi2(2) =  21610.03
Log likelihood = -5623.9512      Prob > chi2  =    0.0000
```

Fixed Effects

```
---------------------------------------------------------------------------
wt |  Coef.    Std. Err.     z    P>|z|   [95% Conf. Interval]
----------------+----------------------------------------------------------
1.sex | .1032682  .0413105    2.50   0.012   .0223011    .1842354
age |   .79505  .0086158   92.28   0.000   .7781633    .8119367
|
c.age#c.age |-.0346971  .0005926  -58.55   0.000  -.0358585   -.0335356
|
sex#c.age |
1  | .1185881  .0120292    9.86   0.000   .0950113     .142165
|
sex#c.age#c.age |
1  |-.0084046  .0008312  -10.11   0.000  -.0100337   -.0067755
|
_cons | 3.352453  .0297906  112.53   0.000   3.294064    3.410841
---------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
Random-effects Parameters  | Estimate   Std. Err.  [95% Conf. Interval]
-----------------------------+------------------------------------------------
ind: Unstructured            |
sd(age) | .0973595   .0030093   .0916365    .1034399
sd(_cons) | .5215715   .0160432   .4910565    .5539828
corr(age,_cons) | .3098716   .0433613   .2225995    .3922169
-----------------------------+------------------------------------------------
sd(Residual) | .4390212   .0048254   .4296646    .4485815
------------------------------------------------------------------------------
LR test vs. linear regression:  chi2(3) =  6048.29   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

- sd(b_{0i})=0.5216, sd(b_{1i})=0.0974, Corr(b_{0i}, b_{1i})=0.3099;
- sd(ε_{ij})=0.4390.

STATA CODE:

```
gen wt=weight/1000
xtmixed wt i.sex age c.age#c.age i.sex#c.age /*
*/c.age#c.age#i.sex  ||ind: age, cov(un)
```

For REML estimation,

STATA CODE:

```
xtmixed wt i.sex age c.age#c.age i.sex#c.age /*
*/c.age#c.age#i.sex  ||ind: age, cov(un) reml
```

- Ignore the correlation in the data and fit linear regression

```
Mixed-effects ML regression  Number of obs      = 6113
Group variable: ind




Wald chi2(2) =  18271.22
Log likelihood = -8648.0948   Prob > chi2  =    0.0000
```

- Observe the major impact in model fit
- Formally, one can consider likelihood ratio test for model comparison
- All fixed effects are statistically significant
- The random effects capture the correlation in the data
- Males tend to be higher at baseline ($\beta_1 = .103$), as well as in evolution over time ($\beta_4 = .119$)

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...

Model:

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

- 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

Results (ML):

```
Mixed-effects ML regression           Number of obs      = 99
Group variable: ind                   Number of groups   = 27

Obs per group: min = 3
avg = 3.7
max = 4


Wald chi2(2) =  110.96
Log likelihood = -200.22607            Prob > chi2  =  0.0000
```

```
----------------------------------------------------------------------
measure |  Coef.     Std. Err.   z    P>|z|    [95% Conf. Interval]
----------------------------------------------------------------------
2.sex   .9352545   1.682763   0.56   0.578   -2.362901    4.233409
age | .7884227   .0880488   8.95   0.000    .6158502    .9609952
|
sex#c.age |
2  |-.2987766    .138049  -2.16   0.030   -.5693477   -.0282054
|
_cons | 16.27586     1.0727  15.17   0.000    14.17341    18.37831
----------------------------------------------------------------------
```

```
------------------------------------------------------------------------
Random-effects Parameters  | Estimate  Std. Err.  [95% Conf. Interval]
----------------------------+-------------------------------------------
idn: Unstructured           |
sd(age) |  .183553  .1083691   .0577052    .5838595
sd(_cons)   2.605069  1.111511   1.128847    6.011784
corr(age,_cons) |-.7317685  .2582298  -.9655557    .1557669
----------------------------+-------------------------------------------
sd(Residual) | 1.330411   .142086   1.079143    1.640186
------------------------------------------------------------------------
LR test vs. linear regression:   chi2(3) =  41.13   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

- sd(b_{0i})=2.605, sd(b_{1i})=0.184, corr(b_{0i}, b_{1i})=-0.732;
- sd(ε_{ij})=1.330

For ML estimation,

STATA CODE:

```
xtmixed  measure i.sex age i.sex#c.age ||idn: age, cov(un)
```

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 D:

Longitudinal Categorical Data Analysis

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

Introduction

- 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 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%).

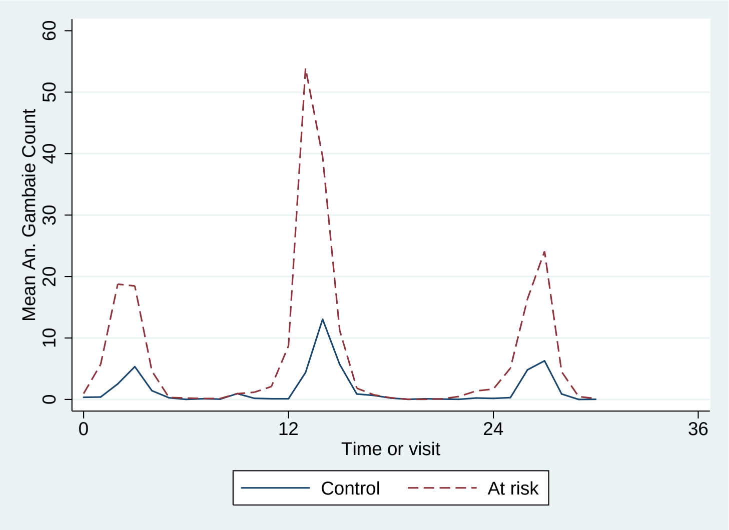


Figure 26: Average evolution of An. gambaie

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.

Marginal models vs 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
- 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 $(Y_1, \ldots, Y_T)$, this method only links each marginal mean to a linear predictor and provides a guess for the variance–covariance structure of $(Y_1, \ldots, Y_T)$
- 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

- Once we have specified a marginal model for each Yt , for the GEE method we must:
- Assume a particular distribution for each Y_t. This determines how $Var(Y_t)$ 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 $\rho = Corr(Y_s, Y_t)$ as identical (but unknown) for all pairs s and t

- 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 $Var(Y)$ 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).

Correlation Structures

- The independence working correlation structure assumes $Corr(Y_s, Y_t) = 0$ for each pair. This treats the observations in a cluster as uncorrelated

- Autoregressive structure: This has the form $Corr(Y_s, Y_t) = \rho^{t-s}$

- Unstructured working correlation matrix permits $Corr(Y_s, Y_t)$ to differ for each pair

- 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

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 jth outcome measured for subject $i = 1, \ldots, N$, $j = 1, \ldots, n_i$ and group the n_i measurements into a vector Y_i
- Conditionally upon q-dimensional random effects $\sim N(0, D)$, the outcomes Y_{ij} are independent with densities of the form

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

with

$$\eta[\psi'(\lambda_{ij})] = \eta(\mu_{ij}) = \eta[E(Y_{ij}|bi, xi)] = x'_{ij}xi + z'_{ij}$$

- For a known link function $\eta(\cdot)$, with x_{ij} and z_{ij} p-dimensional and q-dimensional vectors of known covariate values
- with ξ a p-dimensional vector of unknown fixed regression coefficients, and with ϕ a scale (overdispersion) parameter
- With ξ a p-dimensional vector of unknown fixed regression coefficients, and with ϕ a scale (overdispersion) parameter
- Finally, let $f(bi|D)$ be the density of the $N(0, D)$ distribution for the random effects bi

GEE versus GLMM

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 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)

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

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

GEE Model

```
GEE population-averaged model           Number of obs      =  6113
Group variable:                ind      Number of groups   =  1000
Link:                         logit     Obs per group: min =     1
Family:                     binomial                   avg =   6.1
Correlation:               exchangeable                max =     7
Wald chi2(3)     =  2.13
Scale parameter:               1          Prob > chi2      = 0.5462
```

GEE Model

```
------------------------------------------------------------------
BMIBIN |  Coef.    Std. Err.   z    P>|z|  [95% Conf. Interval]
------------------------------------------------------------------
1.sex | .148298  .1376528   1.08  0.281  -.1214966    .4180926
age | .0016801 .0124828   0.13  0.893  -.0227856    .0261459

sex#age |
1  |-.0180339 .0173981  -1.04  0.300  -.0521336    .0160658
|
_cons |-1.872127 .1011225 -18.51  0.000  -2.070323    -1.67393
```

STATA CODE:

```
xtgee BMIBIN i.sex  c.age i.sex#c.age, i(ind) t(age)/*
*/ corr(exc) link(logit) family(bin)
```

- The option ‘robust’ can be used to obtain the empirically corrected standard error estimates.
- The correlation matrix can be requested by ‘xtcorr’.
- The odds ratios can be requested by ‘eform’.

```
GEE population-averaged model            Number of obs      =    6113
Group variable:                ind       Number of groups   =    1000
Link:                         logit       Obs per group:min =       1
Family:                     binomial                      avg =     6.1
Correlation:                exchangeable                  max =       7
Wald chi2(3)      =    1.45
Scale parameter:               1         Prob > chi2        =  0.6930

(Std. Err. adjusted for clustering on ind)
--------------------------------------------------------------------
|           Semirobust
BMIBIN |   Coef.   Std. Err.    z    P>|z|  [95% Conf. Interval]
--------------------------------------------------------------------
1.sex |  .148298   .1531917   0.97  0.333  -.1519523  .4485483
age |  .0016801  .0146615   0.11  0.909  -.0270558  .0304161
|
sex#c.age |
1   | -.0180339  .0212268  -0.85  0.396  -.0596377  .0235699
|
_cons | -1.872127  .1118498 -16.74  0.000  -2.091348 -1.652905
-------------------------------------------------------------------
```

STATA CODE:

```
xtgee BMIBIN i.sex  c.age i.sex#c.age, i(ind) t(age)/*
*/ corr(exc) link(logit) family(bin) robust
```

xtcorr

```
.xtcorr

Estimated within-ind correlation matrix R:

c1       c2       c3       c4       c5       c6       c7
r1  1.0000
r2  0.1458  1.0000
r3  0.1458  0.1458  1.0000
r4  0.1458  0.1458  0.1458  1.0000
r5  0.1458  0.1458  0.1458  0.1458  1.0000
r6  0.1458  0.1458  0.1458  0.1458  0.1458  1.0000
r7  0.1458  0.1458  0.1458  0.1458  0.1458  0.1458  1.0000
```

Odds ratio estimates ...

```
GEE population-averaged model           Number of obs        =   6113
Group variable:                ind       Number of groups     =   1000
Link:                         logit        Obs per group: min         1
Family:                       binom                         avg =    6.1
Correlation:               exchangeable                     max =      7
Wald chi2(3)        =      1.45
Scale parameter:                1        Prob > chi2          =   0.6930

(Std. Err. adjusted for clustering on ind)
----------------------------------------------------------------------
|             Semirobust
BMIBIN | Odds Ratio   Std. Err.    z    P>|z|    [95% Conf. Interval]
-------------+--------------------------------------------------------
1.sex |  1.159858   .1776807   0.97   0.333   .8590292   1.566037
age |  1.001682   .0146861   0.11   0.909   .9733069   1.030883
|
sex#c.age |
1  |  .9821277   .0208474  -0.85   0.396   .9421058    1.02385
|
_cons |  .1537962   .0172021 -16.74   0.000   .1235205   .1914927
```

STATA CODE:

```
xtgee BMIBIN i.sex  c.age i.sex#c.age, i(ind) t(age)/*
*/ corr(exc) link(logit) family(bin) robust eform
```

- Compare the standard errors of the ‘model based’ and the ‘robust’ versions.
- Interpretation of parameter estimates?

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, $t_{ij} = 1, 2, \ldots$ 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 λ_{ij}

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

```
GEE population-averaged model             Number of obs        =      1419
Group variable:                     id    Number of groups     =        89
Link:                              log    Obs per group: min =           2
Family:                        Poisson                     avg =      15.9
Correlation:              exchangeable                     max =        27
Wald chi2(3)        =       1.79
Scale parameter:                     1    Prob > chi2          =    0.6177

(Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------
|             Semirobust
nseizw |   Coef.     Std. Err.   z    P>|z|  [95% Conf.Interval]
---------------+--------------------------------------------------------
1.trt |  .0155622  .2947376   0.05  0.958  -.5621129  .5932374
studywee | -.0147014    .01688  -0.87  0.384  -.0477856  .0183827
|
trt#c.studywee |
1  |  .0034596  .0202146   0.17  0.864  -.0361603  .0430795
|
_cons |  1.316504  .1809591   7.28  0.000   .9618308  1.671178
------------------------------------------------------------------------
```

STATA CODE:

```
xtgee nseizw i.trt c.studywee i.trt#c.studywee, i(id)\*
*/ t(studywee) corr(exc) link(log) family(poisson) robust
```

The incidence rate ratio estimates can be obtained by including the option ‘eform’.

```
----------------------------------------------------------------------
|           Semirobust
nseizw |   IRR     Std. Err.   z    P>|z|  [95% Conf.Interval]
----------------------------------------------------------------------
1.trt | 1.015684  .2993603   0.05  0.958  .5700034   1.809838
studywee | .9854061  .0166336  -0.87  0.384  .9533382   1.018553
|
trt#c.studywee |
1  | 1.003466  .0202847   0.17  0.864  .9644857   1.044021
|
_cons | 3.730358  .6750423   7.28  0.000  2.616482   5.318427
----------------------------------------------------------------------
```

STATA CODE:

```
xtgee nseizw i.trt c.studywee i.trt#c.studywee, i(id) t(studywee)\*
*/ corr(exc) link(log) family(poisson) robust eform
```

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.

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

```
Mixed-effects logistic regression       Number of obs     =      6113
Group variable: ind                     Number of groups  =      1000
Obs per group: min =        1
avg =       6.1
max =         7
Integration points =   7                Wald chi2(3)      =      2.14
Log likelihood = -2325.1686             Prob > chi2       =    0.5444
-------------------------------------------------------------------
BMIBIN | Coef.     Std. Err.    z    P>|z|    [95% Conf. Interval]
-------------------------------------------------------------------
1.sex | .1774603  .163003    1.09   0.276  -.1420197    .4969403
age | .0022733  .0147053   0.15   0.877  -.0265485    .0310952
|
sex#c.age |
1  |-.021603   .0204887  -1.05   0.292  -.0617601    .0185542
|
_cons |-2.336672  .1266153 -18.45   0.000  -2.584834   -2.088511
-------------------------------------------------------------------
-------------------------------------------------------------------
Random-effects Parameters | Estimate  Std. Err. [95% Conf. Interval]
-------------------------------------------------------------------
ind: Identity             |
sd(_cons) | 1.20833  .0752947    1.06941    1.365295
-------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =   222.16
Prob>=chibar2 = 0.0000
```

STATA CODE:

```
xtmelogit BMIBIN i.sex c.age i.sex#c.age|| ind:
```

The odds ratio estimates can be obtained by including the option ‘or’.

```
---------------------------------------------------------------------
BMIBIN | Odds Ratio  Std. Err.  z    P>|z|   [95% Conf. Interval]
---------------------------------------------------------------------
1.sex |  1.194181   .194655   1.09  0.276   .8676042    1.643684
age |  1.002276  .0147388   0.15  0.877   .9738008    1.031584
|
sex#c.age |
1  |  .9786287  .0200509  -1.05  0.292   .9401084    1.018727
|
_cons |  .0966487  .0122372 -18.45  0.000   .0754086    .1238715
---------------------------------------------------------------------


---------------------------------------------------------------------
Random-effects Parameters | Estimate  Std. Err. [95% Conf. Interval]
---------------------------------------------------------------------
ind: Identity             |
sd(_cons)| 1.20833  .075294     1.06941    1.365295
---------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =   222.16
Prob>=chibar2 = 0.0000
```

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, $t_{ij} = 1, 2, \ldots$ 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 λ_{ij}

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

```
Mixed-effects Poisson regression      Number of obs      =      1419
Group variable: id                    Number of groups   =        89
Obs per group: min =        2
avg =      15.9
max =        27
Integration points =   7              Wald chi2(3)       =     18.70
Log likelihood = -3135.9507           Prob > chi2        =    0.0003
---------------------------------------------------------------------------
nseizw |   Coef.    Std. Err.    z   P>|z|   [95% Conf. Interval]
---------------------------------------------------------------------------
1.trt |-.1704619   .2387161  -0.71  0.475   -.6383369   .2974131
studywee |-.0142876    .004404  -3.24  0.001   -.0229193  -.0056559
|
trt#c.studywee |
1  | .0022903    .006167   0.37  0.710   -.0097968   .0143775
|
_cons |  .817696   .1676814   4.88  0.000    .4890465   1.146346
---------------------------------------------------------------------------
---------------------------------------------------------------------------
Random-effects Parameters | Estimate  Std. Err.  [95% Conf. Interval]
---------------------------------------------------------------------------
id: Identity              |
sd(_cons) | 1.075478  .0857167   .919941    1.257312
---------------------------------------------------------------------------
LR test vs.Poisson regression: chibar2(01)=5317.84 Prob>=chibar2 = 0.0000
```

STATA CODE:

```
xtmepoisson nseizw i.trt c.studywee i.trt#c.studywee  || id:
```

The option ‘irr’ can be used to obtain incidence rate ratios.

STATA CODE:

```
xtmepoisson nseizw i.trt c.studywee i.trt#c.studywee || id:, irr
```

```
------------------------------------------------------------------
nseizw |   IRR    Std. Err.  z    P>|z|  [95% Conf.Interval]
------------------------------------------------------------------
1.trt | .8432752 .2013034 -0.71  0.475  .5281701  1.346371
studywee |  .985814 .0043416 -3.24  0.001  .9773413  .9943601

trt#c.studywee |
1  | 1.002293 .0061812  0.37  0.710   .990251  1.014481
|
_cons | 2.265275 .3798444  4.88  0.000   1.63076  3.146672
-----------------------------------------------------------------
```

- 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 varance-covarance matrix can then be ‘unstructured’.

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

```
Mixed-effects Poisson regression          Number of obs      =       1419
Group variable: id                        Number of groups   =         89

Obs per group: min =         2
avg =      15.9
max =        27

Integration points =   7                  Wald chi2(3)       =      10.30
Log likelihood = -3033.2134               Prob > chi2        =     0.0161
```

```
------------------------------------------------------------------------
nseizw |    Coef.   Std. Err.    z    P>|z|  [95% Conf. Interval]
------------------------------------------------------------------------
1.trt | -.2445031  .2547595  -0.96  0.337  -.7438225   .2548162
studywee | -.0271461  .0099359  -2.73  0.006  -.0466201  -.0076721
|
trt#c.studywee |
1  |  .0106855  .0139551   0.77  0.444   -.016666    .038037
|
_cons |   .894286  .1788611   5.00  0.000   .5437247   1.244847
------------------------------------------------------------------------


------------------------------------------------------------------------
Random-effects Parameters |  Estimate  Std. Err.  [95% Conf.Interval]
------------------------------------------------------------------------
id: Unstructured             |
sd(studywee) |   .048832  .0057226   .0388108   .0614408
sd(_cons) |   1.12926  .0977606   .9530268   1.338083
corr(studywee,_cons) | -.3340864  .1312477  -.5628401   -.057797
------------------------------------------------------------------------
LR test vs. Poisson regression: chi2(3)= 5523.31 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

- Compare the log-likelihoods of the two models? which one do you prefer?
- Parameter interpretation?
- What does the correlation parameter of the random effects variances imply?

The ‘irr’ estimates are:

```
nseizw |  IRR       Std. Err.   z    P>|z|  [95% Conf. Interval]
---------------+------------------------------------------------
1.trt |  .7830935  .1995005  -0.96  0.337   .4752936  1.290224
studywee |   .973219  .0096698  -2.73  0.006   .9544499  .9923572
|
trt#c.studywee |
1  |  1.010743  .014105    0.77  0.444   .9834721   1.03877
|
_cons |  2.445589 .4374208    5.00  0.000    1.72241  3.472405
----------------------------------------------------------------
```

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, $t_{ij} = 1, 2, \ldots$ 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 λ_{ij}

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

GEE Parameter Estimates...

```
GEE population-averaged model                   Number of obs      =      4768
Group variable:                         ID      Number of groups   =       160
Link:                                  log      Obs per group: min =        26
Family:                            Poisson                     avg =      29.8
Correlation:                  exchangeable                     max =        32
Wald chi2(4)       =    838.74
Scale parameter:                         1      Prob > chi2        =    0.0000

(Std. Err. adjusted for clustering on ID)
--------------------------------------------------------------------------------
|             Semirobust
Gamb |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
1.Village |   1.368464    .168422     8.13   0.000     1.038363    1.698565
1.Season |   3.065426   .1189174    25.78   0.000     2.832352      3.2985
Time |   .0059768   .0048513     1.23   0.218    -.0035316    .0154852
|
Village#c.Time |
1  |  -.0032723   .0060887    -0.54   0.591    -.0152059    .0086612
|
_cons |  -1.968008     .17534   -11.22   0.000    -2.311668   -1.624348
--------------------------------------------------------------------------------
```

GEE Incidence Rate Estimates...

```
GEE population-averaged model                   Number of obs      =      4768
Group variable:                         ID      Number of groups   =       160
Link:                                  log      Obs per group: min =        26
Family:                            Poisson                     avg =      29.8
Correlation:                  exchangeable                     max =        32
Wald chi2(4)       =    838.74
Scale parameter:                         1      Prob > chi2        =    0.0000

(Std. Err. adjusted for clustering on ID)
--------------------------------------------------------------------------------
|             Semirobust
Gamb |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
1.Village |    3.92931   .6617821     8.13   0.000     2.824589    5.466097
1.Season |   21.44359   2.550016    25.78   0.000     16.98536    27.07199
Time |   1.005995   .0048804     1.23   0.218     .9964746    1.015606
|
Village#c.Time |
1  |    .996733   .0060688    -0.54   0.591     .9849091    1.008699
|
_cons |    .139735   .0245011   -11.22   0.000     .0990958    .1970402
--------------------------------------------------------------------------------
```

GLMM Parameter Estimates...

```
Mixed-effects Poisson regression                Number of obs      =      4768
Group variable: ID                              Number of groups   =       160

Obs per group: min =        26
avg =      29.8
max =        32

Integration points =   7                        Wald chi2(4)       =   8353.14
Log likelihood = -22193.523                     Prob > chi2        =    0.0000

--------------------------------------------------------------------------------
Gamb |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
1.Village |   1.740094   .2167599     8.03   0.000      1.315252    2.164935
1.Season |   3.062087   .0336403    91.02   0.000      2.996153    3.128021
Time |   .0054538   .0015797     3.45   0.001      .0023576      .00855
|
Village#c.Time |
1  |  -.0026978   .0017825    -1.51   0.130    -.0061913    .0007958
|
_cons |  -2.696573    .160075   -16.85   0.000    -3.010314   -2.382831
--------------------------------------------------------------------------------


------------------------------------------------------------------------------
Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
ID: Identity                 |
sd(_cons) |   1.330066   .0886494      1.167186    1.515675
------------------------------------------------------------------------------
LR test vs. Poisson regression:  chibar2(01) =  8909.30 Prob>=chibar2 = 0.0000
```

GLMM Incidence Rate Estimates...

```
Mixed-effects Poisson regression                Number of obs      =      4768
Group variable: ID                              Number of groups   =       160

Obs per group: min =        26
avg =      29.8
max =        32

Integration points =   7                        Wald chi2(4)       =   8353.14
Log likelihood = -22193.523                     Prob > chi2        =    0.0000

--------------------------------------------------------------------------------
Gamb |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
1.Village |   5.697877   1.235071     8.03   0.000      3.72569    8.714038
1.Season |   21.37212   .7189646    91.02   0.000     20.00842    22.82876
Time |   1.005469   .0015884     3.45   0.001      1.00236    1.008587
|
Village#c.Time |
1  |   .9973059   .0017777    -1.51   0.130     .9938278    1.000796
|
_cons |   .0674362   .0107949   -16.85   0.000     .0492762    .0922889
--------------------------------------------------------------------------------

------------------------------------------------------------------------------
Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
ID: Identity                |
sd(_cons) |   1.330066   .0886494      1.167186    1.515675
------------------------------------------------------------------------------
LR test vs. Poisson regression:  chibar2(01) =  8909.30 Prob>=chibar2 = 0.0000
```

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)

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

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

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

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

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 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

Assignment

- Consider EDHS2016 data and outcome variable
  - Outcome variable 1: Age at first marriage
  - Outcome variable 2:Contraceptive utilization
  - Outcome variable 3: ANC visit
  - Outcome variable 4: Outcome of pregnancy
- Independent variable: age, residence, income, religion, ...
- Fit GEE for the data
- Interpret the result

Frailty Model

- Ordinary survival models deal with the simplest case of independent and identically distributed data
- A frailty model is an heterogeneity model where the frailties are assumed to be individual or spell specific
- Parametric specification plus covariates can only go so far in explaining the variability in observed time to failure
- Excess unexplained variability is known as overdispersion
- Overdispersion is caused either by misspecification or omitted covariates
- Frailty models can help explain the unaccounted for heterogeneity
- Current model cannot as such adequately account for why subjects with shorter times to failures are more frail than others
- A frailty model attempts to measure this overdispersion by modeling it as resulting from a latent multiplicative effect on the hazard function

Model Specification

- The model is defined as

$$h_i(t/z) = zh_0(t)exp(\beta_0 + \beta_1 x)$$

  is the hazard function of the *i*th individual
- From a PH perspective, it is easy to see how z may correspond to an omitted covariate
- The distribution of z is specified to be, say, Gamma
- In this situation, the shared frailty model is appropriate, that is multiple observations of the same individual always has the same value of z.

Shared frailty model

- For the jth observation in the ith group, a frailty model treats

$$htij/z_{ij}) = z_{ij}h(t_{ij})$$

- While a shared frailty model has

$$htij/z_{ij}) = z_i h(t_{ij})$$

- The frailty is shared among the group
- Group may represent a family, for example, or simply a single subject for which multiple episodes are observed

- In stata, we can fit frailty model using the command

  ```
  streg  momage, dist(exp)  frailty(gamma) nolog
  ```
- We can specify different distribution for both the parameter and for the frailty term
- Shard frailty model can be fitted

  ```
  streg  momage, dist(exp)  frailty(gamma) shared(idno) nolo
  ```

Conclusion

- The shape parameter is fit as ln p, but streg then reports p and 1/p
- We find that p is greater than 1, which means that the hazard of failure increases with time
- Those individuals who possess $\alpha > 1$ are said to be more frail for reasons left unexplained by the covariates and will have an increased risk of failure
- Conversely, those individuals with $\alpha < 1$ are less frail and will tend to survive longer all else being equal
- As α approaches 0, the proportional hazards property returns

Chapter 6

Non-parametric Statistical Tests

Introduction

- Statistical methods which depend on the assumptions about the distribution of parameters in the population are referred to as parametric methods
- Parametric tests include t-test, ANOVA, Regression, Correlation and so on
- In order to use a parametric test, we must assume a normal distribution for the dependent variable, equality of variance where population are compared and large sample size
- However in real research situations things do not come with labels detailing the characteristics of the population of origin

Introduction

- Non-parametric statistics (we call sometimes distribution free statistics) were designed to be used when we know nothing about the distribution of the variable of interest in the population
- It requires fewer assumptions about the population probability distribution
- It also handles data collected in the form of ranking

Introduction

- More generally, a nonparametric method has the following advantages;
  - Methods quick and easy to apply.
  - Accommodate unusual or irregular sample distribution.
  - Basic data need not be actual measurement.
  - Can be used with small sample size.
  - Not affected by the presence of outliers.
  - Less sensitive for measurement error as it uses ranks.
  - Inherently robust due to lack of stringent assumption.

The One-sample Location

- Interest is on the location (median) of a population.
- Two types of data for which such analysis is important:
- Paired replicates data:-pretreatment and post treatment observations;
- One-sample data:-observations from a single population.

Parametric Versus Non-parametric tests

Parametric	Non-parametric
Unpaired t-test	Wilcoxon rank sum test Mann-Whitney U test
Paired t-test	Sign test Wilcoxon signed rank test
ANOVA	Kruskal-Wallis test
Repeated measures ANOVA	Freedman test

Rank Test

- Many nonparametric procedures are based on ranked data in which the magnitude of the observation will not be taken in to account
- Data are ranked by ordering them from lowest to highest and assigning number to them, in order, the integer values from 1 to the sample size if there are no ties
- In the presence of ties, the tied observation will be given the mean of the ranks they would have received if there were no ties
- Let us denote the rank by w_i: the data on mental health with the corresponding ranks are given below:

Example: Mental health data

- In a hospital there were six patients having diseases in the same degree. To test the effcetiveness of the new drug, three patients were selected at random to receive the new drug and the remaining three were given the control (placebo pill for psychological effect). The outcome variable is global assessment of function (GAF).
- High score means better functioning and the observations vary from 0 to 100.

Control	25	10	35
Treatment	36	26	40

Rank Test

Control	**2**	**1**	**4**
	(25)	(10)	(35)
Treatment	**5**	**3**	**6**
	(36)	(26)	(40)

Significance of ranking

- Rank is solely determined by patient's health status.
- Rank is independent of receiving treatment or placebo.
- Rank is assigned to the patient before the assignments to treatment and control are made.
- The p-value of the test is computed under the assumption that the null hypothesis is true.

Significance of ranking

- All possible divisions of ranks in to two groups

Treated	4	5	6	3	5	6	3	4	6	3	4	5	2	5	6
Control	1	2	3	1	2	4	1	2	5	1	2	6	1	3	4
Treated	2	4	6	2	4	5	2	3	6	2	3	5	2	3	4
Control	1	3	5	1	3	6	1	4	5	1	4	6	1	5	6
Treated	1	5	6	1	4	6	1	4	5	1	3	6	1	3	5
Control	2	3	4	2	3	5	2	3	6	2	4	5	2	4	6
Treated	1	3	4	1	2	6	1	2	5	1	2	4	1	2	3
Control	2	5	6	3	4	5	3	4	6	3	5	6	4	5	6

- Number of possibilities is obtained by:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

- This combination can be handled using R function, choose(), as follows:

```
> choose(6, 3)
[1] 20
```

- Because of random selection of patients to receive the treatment, these 20 possibilities are equally likely in which each of the 20 possibilities has probability $1/20$.

- Hypothesis
  - H_0: treatment has no effect on the GAF
  - H_1: treatment has the effect to increase the GAF
- In terms of ranks:
  - high treatment ranks are in favor of the alternative hypothesis
  - low treatment ranks are in favor of the null hypothesis

- The basic result derived above states that the probability of observing any particular n-tuple $(\mathrm{s}_1, ..., \mathrm{s}_\mathrm{n})$ is

$$P_{H_0}(S_1 = s_1, ..., S_n = s_n) = \frac{1}{\binom{N}{n}}$$

- The hypothesis H of no treatment effect is rejected, and the superiority of the new treatment effect acknowledge if in this ranking the n treated subjects rank sufficiently high.

Test Statistic

- To complete the procedure, it is necessary to decide when the treatment ranks $\mathrm{S}_1 < \mathrm{S}_2 < ... < \mathrm{S}_\mathrm{n}$, are sufficiently large.
  - The relationship between Test statistic and Treatment ranks is

Test statistic	Treatment Ranks
high	high
low	low

- A simple and effective test statistic is the sum of the treatment ranks which is denoted by $\mathrm{W_S}$:

$$\mathrm{W_S = S_1 + S_2 + ... + S_n}:$$

- This statistic is known as **WILCOXON RANK SUM TEST**.

- Reject the null hypothesis when $\mathrm{W_S}$ is sufficiently large:

$$\boxed{\mathrm{W_S} \geq \mathrm{c}}$$

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

treat |      obs    rank sum    expected
-------------+----------------------------------
c |        3           7        10.5
t |        3          14        10.5
-------------+----------------------------------
combined |        6          21          21

unadjusted variance        5.25
adjustment for ties        0.00
----------
adjusted variance          5.25

Ho: GAF(treat==c) = GAF(treat==t)
P{GAF(treat==c) > GAF(treat==t)} = 0.111
```

The Wilcoxon-Mann-Whitney U-test

- Two independent populations (n_1, n_2).
- With $X_1, X_2, ..., X_{n_1} and Y_1, Y_2, ..., Y_{n_2}$.
- The populations are non-normal.
- Treatment effect is denoted by Δ (difference in location parameter, $E(Y) - E(X)$).
- Investigate if $\Delta = 0$.

Steps

- Combine the two samples.
- Arrange the observations in the order of increasing size.
- Write the ranks of the observations.
- Add all the ranks from the first sample R_1 and the second sample R_2 .

Steps Cont'd

- Test Statistics:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_{1.}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_{2.}$$

One-sided upper tail test

- Hypotheses:

$$H0 : \Delta \leq 0$$

$$H1 : \Delta > 0$$

- At α level of significance, reject H_0 if $U_1 < U_{n_1,n_2}, \alpha$ p $< \alpha$.
- Where U_{n_1,n_2}, α is to be obtained from table.

Two-sided test

- Hypotheses:

$$H0 : \Delta = 0$$

$$H1 : \Delta \neq 0$$

- At α level of significance, reject H_0 if $min(U_1, U_2) < U_{n_1,n_2}, \alpha/2$ or p $< \alpha$.

Large sample approximation

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \approx N(0, 1)$$

Example

- Consider 43 patients from whom WBC and BMI (2 categories) were collected.
- The research question is to check if WBC is difference among the BMI category
- Dependent variable: WBC
- Factor variable: BMI

Stata output

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

BMI2 |      obs   rank sum    expected
-------------+-----------------------------------
1 |       24         486          504
2 |       17         375          357
-------------+-----------------------------------
combined |      41         861          861

unadjusted variance     1428.00
adjustment for ties       -4.10
----------
adjusted variance       1423.90

Ho: WBC(BMI2==1) = WBC(BMI2==2)
z =  -0.477
Prob > |z| =   0.6334
```

```
ranksum WBC, by(BMI2)
```

Signed Rank Sum Test

- Suppose we obtain 2n observations.
- Two observations on each of n subjects (blocks, patients, etc).
- Symmetric about a common median θ
- θ is referred to as the treatment effect.

- Hypothesis of interest here is that of zero shifts in location,

$$H_0 : \theta = 0$$

- The distributions are symmetrical around 0.
- Corresponds to no shift in location due to the treatment.

- For $i = 1, 2, ..., n$, the differences,

$$Z_i = Y_i - X_i, i = 1, 2, ..., n$$

- are mutually independent;
- come from a continuous population.

- Form the absolute values:
  $|Z1|, |Z2|, ..., |Zn|$
- Order them from least to greatest.
- Obtain the positive rank sum T^+ of Z_i.
- Obtain the negative rank sum T^- of Z_i.

One-sided upper tail test

- Hypotheses:

$$H_0 : \theta \leq 0$$

$$H_1 : \theta > 0$$

- At *alpha* level of significance, reject H_0 if $T^- < t_\alpha$ or $p < \alpha$.
- Where t_α is to be obtained from table.

One-sided lower tail test

- Hypotheses:

$$H_0 : \theta \geq 0$$

$$H_1 : \theta < 0$$

- At *alpha* level of significance, reject H_0 if $T^+ < t_\alpha$ or $p < \alpha$.
- Where t_α is to be obtained from table.

Two-sided test

- Hypotheses:

$$H_0 : \theta = 0$$

$$H_1 : \theta \neq 0$$

- At *alpha* level of significance, reject H_0 if the smallest of $T^+ or T^- < t_{\alpha/2}$ or $p < \alpha$.

- In 5 pairs of matched patients suffering from chronic schizophrenia, one member of each pair was assigned at random to treatment with a new drug (Stelazine); and the other received a placebo. The behavior ratings of these patients are shown below

Pair	1	2	3	4	5
Control	1.8	2.4	2.0	1.5	1.5
Treatment	1.9	2.5	1.81	1.45	1.54

- Test the hypothesis that the treatment increases the behavior rating.
- Use signed rank test and perform approximate calculation

```
Wilcoxon Signed Ranks Test
Ranks
                      N Mean Rank Sum of Ranks
treated - control Negative Ranks 2a    3.50      7.00
                    Positive Ranks 3b    2.67     8.00
Ties 0c
Total 5
Test Statistics
treated - control
Wilcoxon Signed Ranks Test
Asymp. Sig. (2-tailed) .892
```

The Kruskal-Wallis Test

- For a Gaussian outcome the means of three or more independent groups are compared by one-way ANOVA.
- When the assumption of one-way ANOVA are not met, i.e.:
- Populations are not normally distributed with equal variance, data consist of only ranks.
- The alternative is the Kruskal-Wallis one-way analysis.
- To test the hypothesis of equal location parameter.

Steps

Suppose $n_1, n_2, ..., n_k$ observations from k samples.

- Combine into a single series of size n.
- Arrange in order of magnitude from smallest to largest.
- Assign the rank 1 for the smallest observation, and the rank n for the largest, in the joint ordering. If observations have the same value use the mean rank.
- The ranks assigned to observations in each of the k groups are added separately to give k rank sums.

Test Statistic

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

- k: the number of samples
- n_j : the number of observations in the j^{th} sample
- n: the number of observations in all samples combined
- R_j : the sum of the ranks in the j^{th} sample

Example

- Consider 43 patients from whom WBC and BMI (3 categories) were collected.
- The research question is to check if WBC is difference among the BMI category
- Dependent variable: WBC
- Factor variable: BMI

```
. kwallis WBC, by(BMI3)

Kruskal-Wallis equality-of-populations rank test

+------------------------+
| BMI3 | Obs | Rank Sum |
|------+-----+----------|
|    1 |  14 |   229.00 |
|    2 |  10 |   257.00 |
|    3 |  17 |   375.00 |
+------------------------+

chi-squared =     3.775 with 2 d.f.
probability =     0.1514

chi-squared with ties =     3.786 with 2 d.f.
probability =     0.1506
```

Chapter 7

Time Series Analysis

Time series Data analysis

- Time Series is a sequence of numerical data obtained at regular time intervals
- It is a collection of data y_t $(t = 1, 2, \ldots, T)$, with the interval between y_t and y_{t+1} being fixed and constant.
- Occurs in many areas: economics, finance, environment, medicine
- Components of time series data:
  - Trend
  - Cyclical
  - Seasonal
  - Irregular

- Trend: the long-term patterns or movements in the data. Long-term upward or downward pattern of movement

- Seasonal variation: Regular periodic fluctuations that occur within year

- Cyclical variation: Cyclical variations are similar to seasonal variations. Cycles are often irregular both in height of peak and duration

- Irregular: Unpredictable, random, residual fluctuations

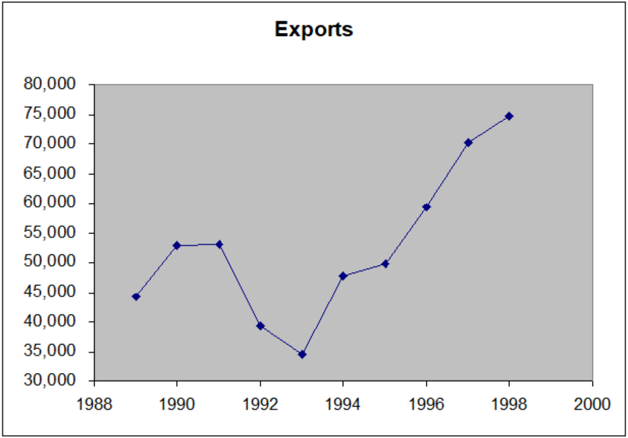
- Observed value in time series is the product of components $Y_i = T_i \times S_i \times C_i \times I_i$

- where
  - T_i = Trend value at year i
  - S_i = Seasonal value at time i
  - C_i = Cyclical value at year i
  - I_i = Irregular (random) value at year i

Example

- Example of time series data
  - GDP of the country, measured each year
  - Number of babies born in each hour
  - Yearly export by country (Ethiopia)
  - Gasoline consumption, which is high in summer when most people go on vacation
- The aims of time series analysis are to describe and summarize time series data
- Understand the past
- fit models, and make forecasts

- Graphical presentation for the time series data



Time series versus Longitudinal data

- Univariate time series data typically arise from the collection of many data points over time from a single source, such as from a person, country, financial instrument, etc
- Longitudinal data typically arise from collecting a few observations over time from many sources, such as a few blood pressure measurements from many people.
- longitudinal data is often used in causal analyses, to understand the impact of interventions or treatments, whereas time series are often used in forecasting

Goals of time series analysis

- Data compression: provide compact description of the data.
- Explanatory: seasonal factors(temperature,humidity, pollution, etc)
- Signal processing: extracting a signal in the presence of noise
- Prediction: use the model to predict future values of the time series

Goals of time series analysis:

- Descriptive: Identify patterns in correlated data—trends and seasonal variation
- Explanation: understanding and modeling the data
- Forecasting: prediction of short-term trends from previous patterns
- Intervention analysis: how does a single event change the time series?
- Quality control: deviations of a specified size indicate a problem

- It is assumed that a time series data set has at least one systematic pattern
- The most common patterns are trends and seasonality
- Trends are generally linear or quadratic
- To find trends, moving averages or regression analysis is often used.
- Seasonality is a trend that repeats itself systematically over time

- Time series are very complex because each observation is somewhat dependent upon the previous observation, and often is influenced by more than one previous observation
- Random error is also influential from one observation to another
- These influences are called autocorrelation—dependent relationships between successive observations of the same variable
- The challenge of time series analysis is to extract the autocorrelation elements of the data, either to understand the trend itself or to model the underlying mechanisms

- Time series reflect the stochastic nature of most measurements over time
- Thus, data may be skewed, with mean and variation not constant, non-normally distributed, and not randomly sampled or independent
- There are two main approaches used to analyze time series (1) in the time domain or (2) in the frequency domain

Model

- The general model introduced by Box and Jenkins (1976) includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model

- The three types of parameters in the model are: the autoregressive parameters (p), the number of differencing passes (d), and moving average parameters (q).

Autoregressive Integrated Moving Average(ARIMA)

- ARIMA (2,0,1)

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \epsilon_t + b_1 \epsilon_{t-1}$$

- Kinds of processes:
  - Random (stochastic) process
  - Deterministic process
  - Mixed

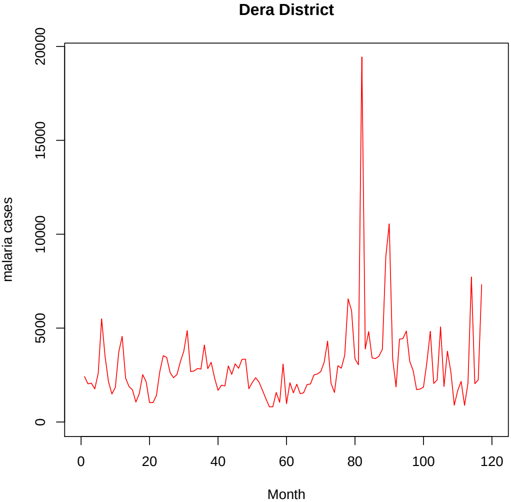
Two Common Processes

- Autoregressive process: Most time series consist of elements that are serially dependent in the sense that you can estimate a coefficient or a set of coefficients that describe consecutive elements of the series from specific, time-lagged

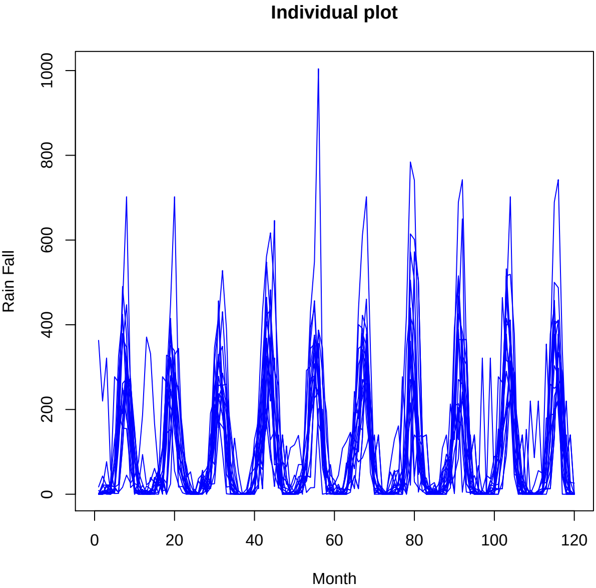
- Moving average process: Independent from the autoregressive process, each element in the series can also be affected by the past error (or random shock) that cannot be accounted for by the autoregressive component

Plotting Time Series Data

- We have data on malaria cases of Dera district in Amhara region for 10 years



- We have data on rainfall of 18 districts in Amhara region for 10 years



Forecasting

- The major purpose of forecasting with time series is to extrapolate beyond the range of the explanatory variables
- Quantitative forecasting methods include;
- Moving Average model
- Exponential smoothing
- Trend modeling

Introduction to Bayesian statistics

-

Conclusion

- Reading Assignment
- Reference Material
- Instruction about the exam
  - The questions will be from from what I taught in the class
  - The content will be multiple choice, short answer, and output interpretation
  - Allowed time will be 2+ hours
  - Exam date:June, 24, 2010

The End