Module 2

Generalized Estimating Equations and Mixed-Effects Models for Longitudinal Data Analysis

Anna Plantinga, PhD
Department of Mathematics and Statistics, Williams College

Katie Wilson, PhD
Department of Biostatistics, University of Washington

SISCER 2023
July 10-12, 2022

A Bit About Us

Katie Wilson

Dept. of Biostatistics
Univ. of Washington

Anna Plantinga

Math/Stat Department
Williams College

We also thank prior module instructors including Dr. Colleen Sitlani (UW) and Dr. Benjamin French (Vanderbilt) for sharing materials and prior TA Dr. Yiqun Chen (Stanford) for his work on the R activities.

What To Expect From This Module

- Brief review of some key features of longitudinal studies
- Exploratory analysis and graphical displays for longitudinal data
- Learn about two key families of longitudinal models in some depth:
    - Generalized estimating equations
    - (Generalized) linear mixed models
- General approach:
    - The focus will be on practical application of these methods, with illustrative examples in R
    - Some theoretical background and technical details will be provided
- At the conclusion of this module, you should be able to apply appropriate exploratory and regression techniques to summarize and generate inference from longitudinal data

Resources

- **Module website:** slides, R code, schedule, etc.
- **Slack:** ask us questions, interact with other module participants, access recordings
- **Office hours:**
  - Anna: 1-2pm PT on Monday, July 10
  - Katie: 3:30-4:30 PT on Wednesday, July 12
- We'll recommend **textbooks and articles** for further reading

A Bit About You

Please type in the chat:

- Briefly introduce yourself - what's your name, and what state or country are you currently located in?
- What's your primary role in most of the studies in which you participate?

Today, July 10, 8:30-12:

8:30-9:30	Introduction to longitudinal studies + some key terminology
9:30-9:45	Break
9:45-10:45	Data activity: Exploring longitudinal data
10:45-11:00	Break
11:00-12:00	Generalized least squares, generalized estimating equations

Tuesday, July 11, 8:30-12:

8:30-9:45	Linear mixed models + discussion
9:45-10:00	Break
10:00-11:00	Data activity: Longitudinal analyses with continuous outcomes
11:00-11:15	Break
11:15-12:00	Marginal models for binary/count data (GEE)

Wednesday, July 12, 8:30-12:

8:30-9:30	Conditional models for binary/count data (GLMM) + discussion
9:30-9:45	Break
9:45-10:45	Data activity: Longitudinal analyses with binary/count outcomes
10:45-11:00	Break
11:15-12:00	Special topics + wrap-up

Resources

Introductory

- Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Wiley, 2011.
- Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/ Hierarchical Models*. Cambridge University Press, 2007.
- Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. Wiley, 2006.

Advanced

- Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*, 2nd Edition. Oxford University Press, 2002.
- Molenbergs G, Verbeke G. *Models for Discrete Longitudinal Data*. Springer Series in Statistics, 2006.
- Verbeke G, Molenbergs G. *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, 2000.

Overview

Day 1: Monday, July 10

Introduction

Activity

General Linear Model

Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

Longitudinal Studies

Repeatedly collect information on the same individuals over time

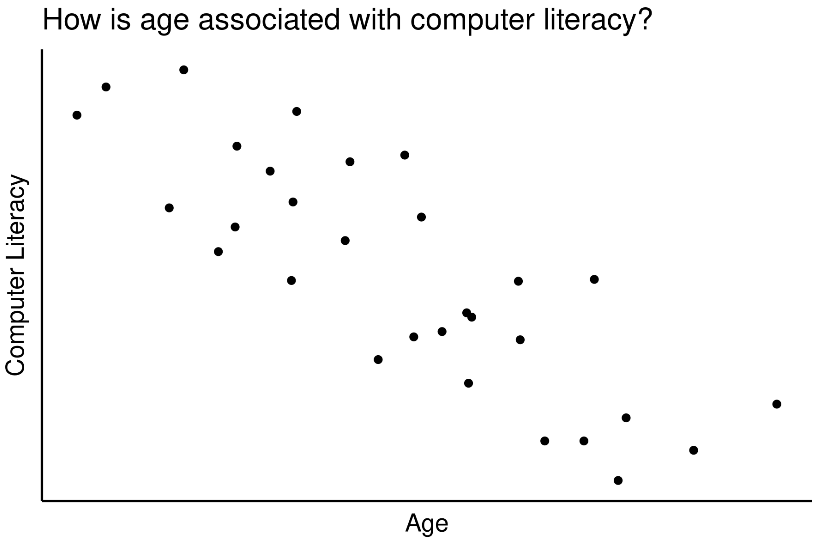
Longitudinal Studies

Repeatedly collect information on the same individuals over time

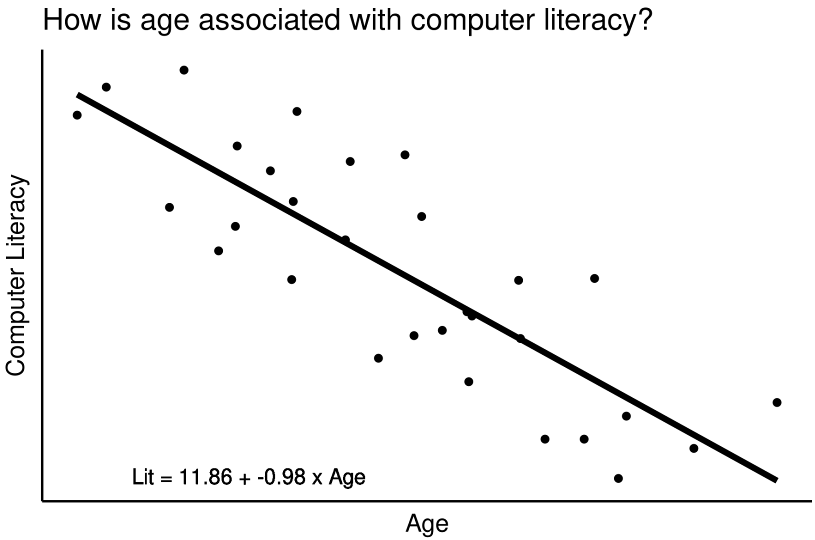
Benefits

- Separate cohort and age effects

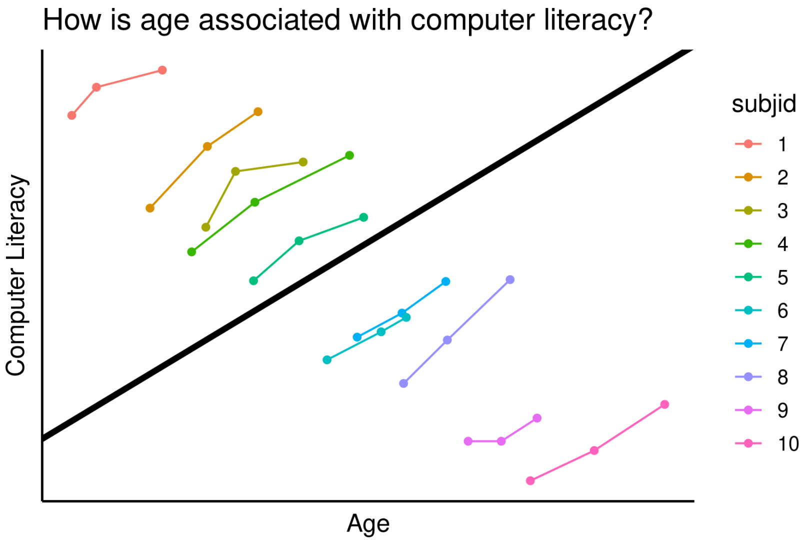
Cohort vs. Age Effects



Cohort vs. Age Effects



Cohort vs. Age Effects

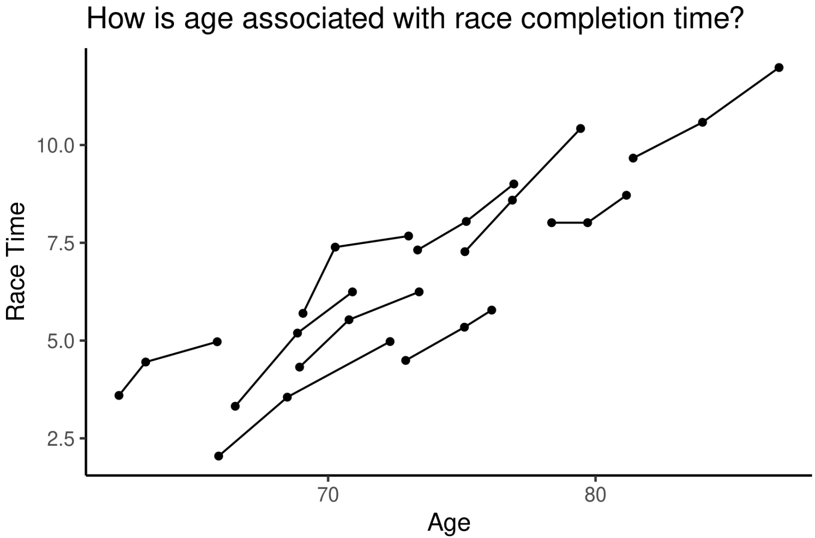


Cohort vs. Age Effects

- On average, older individuals have lower computer literacy
  - Visible from both cross-sectional and longitudinal data
- Longitudinal data shows that:
  - Older individuals began at a lower level *(cohort effect)*
  - Everyone's computer literacy improved as they get older *(age effect)*
- Note: period effects (calendar date) are also sometimes important
  - Any two of age, cohort, period determine the third

Cohort vs. Age Effects

Cohort and age effects can also be similar:



Two-Stage Model

- Formally, cohort and age effects may be represented as a two-stage model (subjects $i = 1, ..., N$; time points $j = 1, ..., n_i$):
  1. Cross-sectional comparison at baseline,
  $$\mathrm{E}[Y_{i1}] = \beta_0 + \beta_C x_{i1}$$
  2. Longitudinal comparison,
  $$\mathrm{E}[Y_{ij} - Y_{i1}] = \beta_L (x_{ij} - x_{i1})$$
- Overall association:
$$\mathrm{E}[Y_{ij}] = \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \epsilon_{ij}$$
- To estimate change over time, cross-sectional studies assume $\beta_C = \beta_L$

Longitudinal Studies

Repeatedly collect information on the same individuals over time

Benefits

- Separate cohort and age effects
- Demonstrate time ordering of exposure and outcome

Timing of Exposure and Outcome

- Cross-sectional study:

Less lonely → Healthier
Healthier → Less lonely

- Longitudinal study (e.g., Harvard Study of Adult Development):

Close relationships at age 50 → Physical health at age 80

- Provides some evidence towards causality
  - One of Hill's Criteria for Causality
  - ⋆ There are several other challenges to generating causal inference from longitudinal data, particularly observational longitudinal data

Longitudinal Studies

Repeatedly collect information on the same individuals over time

Benefits

- Separate cohort and age effects
- Demonstrate time ordering of exposure and outcome
- Statistically:
  - Gains in efficiency: fewer subjects needed to detect the same effect
  - Partition within-subject and between-subject variability
  - “Each subject acts as their own control”

Longitudinal Studies

Repeatedly collect information on the same individuals over time

Challenges

- Analysis must account for longitudinal correlation

Accounting for Correlation

- Individuals are assumed to be independent
- Longitudinal dependence may be a “nuisance” feature (not the primary scientific interest)
- Ignoring dependence may lead to incorrect inference
    - Longitudinal correlation usually positive
    - Estimated standard errors may be too small
    - Confidence intervals are too narrow; too often exclude true value

Longitudinal Studies

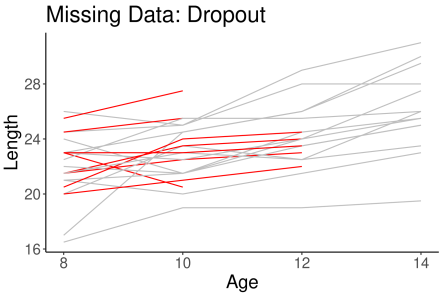
Repeatedly collect information on the same individuals over time

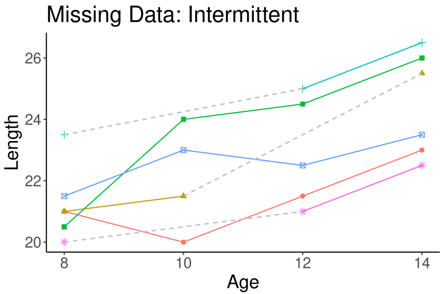
Challenges

- Analysis must account for longitudinal correlation
- Account for incomplete participant follow-up

Missing Data in Longitudinal Studies

- **Balanced** design: all participants measured at the same time points
- **Unbalanced** design: measurement times not intended to match
- **Incomplete** data: observations not available at all intended times





Longitudinal Studies

Repeatedly collect information on the same individuals over time

Challenges

- Analysis must account for longitudinal correlation
- Account for incomplete participant follow-up
- Time-varying covariates
  - Complicates causal argument
  - Requires choosing exposure lag

Longitudinal Studies: Disambiguation

- **Time to clinical outcome:** survival analysis
  - Longitudinal observations *and* time-to-event outcomes: see Module 14
- **Time series:** many time points, one or a few “individuals”
  - Stock market, climate, etc.
  - Different statistical methods apply
- **Panel studies:** social scientists’ name for longitudinal studies
  - Many of the same methods apply
- **Clustered data**: Larger class that includes longitudinal studies
  - Correlation may be due to other shared characteristics (e.g., school, family, neighborhood)
  - Many of the same methods apply

Exploratory Data Analysis

- Summary statistics over time (by group)
- Plots of individual trajectories and/or mean values
- Empirical covariance structure

Goal: Summarize mean and covariance structure

(Easier for quantitative outcomes than other types!)

Exploratory Data Analysis: Best Practices

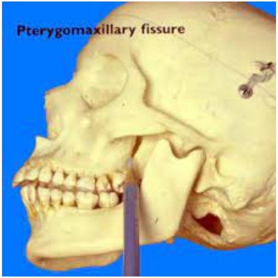
1. Show as much of the data as possible
2. Highlight aggregate patterns of potential scientific interest
3. Identify both cross-sectional (cohort) and longitudinal (age) patterns
4. Facilitate identification of unusual individuals or observations

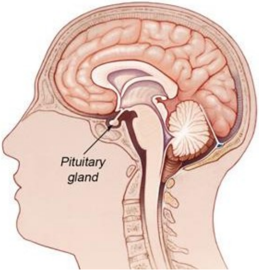
Three Case Studies

1. Dental Growth
2. Air Pollution and Health
3. Amenorrhea with Birth Control

Case Study 1: Dental Growth

- **Study: Dental growth in preteens**
  - Measured distance between the pituitary gland and the pterygomaxillary fissure at ages 8, 10, 12, and 14
  - Easy to identify on x-rays of the side of the head

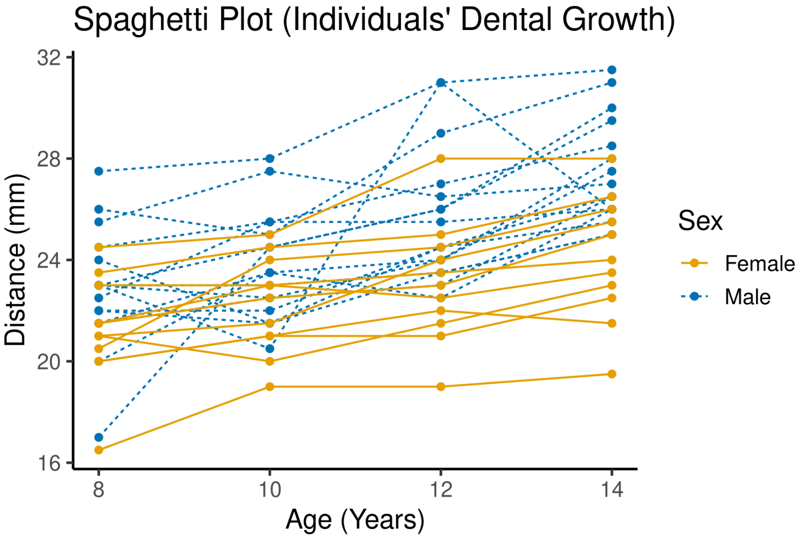




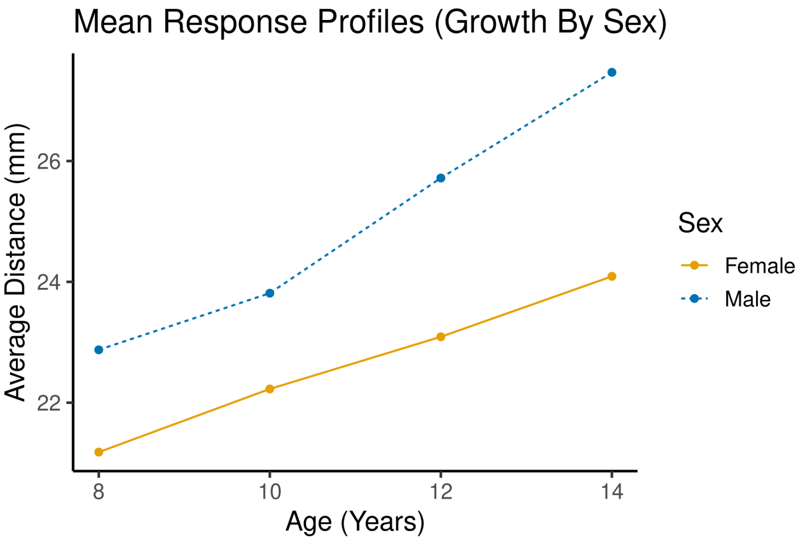
Case Study 1: Dental Growth

- **Study: Dental growth in preteens**
  - 11 girls and 16 boys
  - A measure of growth (see below) taken at ages 8, 10, 12, and 14
  - Balanced and complete data
- **Outcome of interest**
  - Distance (mm) from center of the pituitary gland to the pteryomaxillary fissure
- **Research questions**
  - What is the trajectory of growth in preteens?
  - Does the growth rate differ between boys and girls?
  - How much heterogeneity is there in children's growth rates?

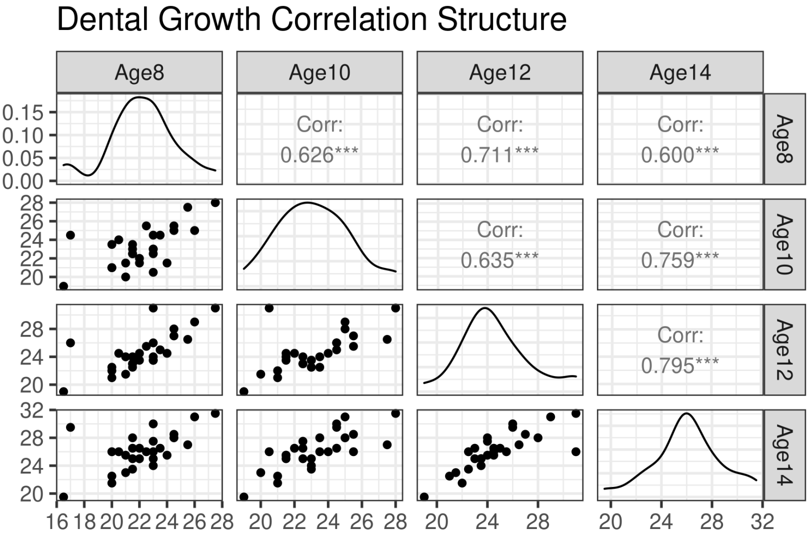
Case Study 1: Dental Growth



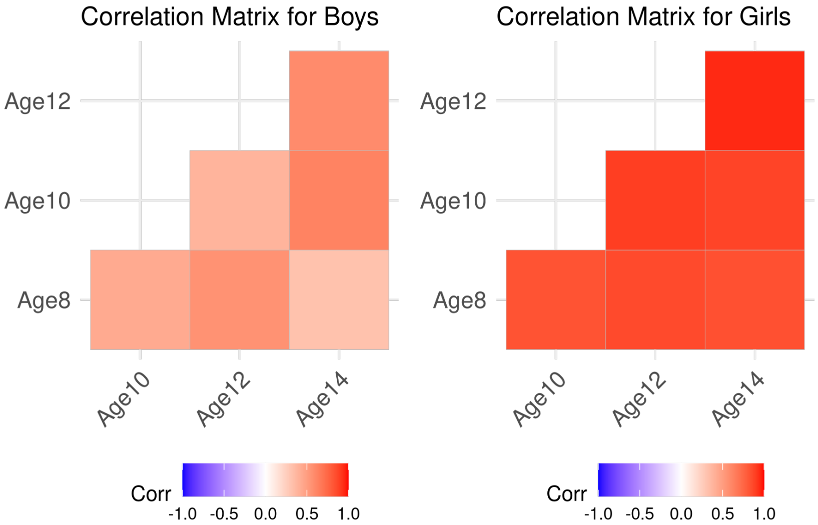
Case Study 1: Dental Growth



Case Study 1: Dental Growth



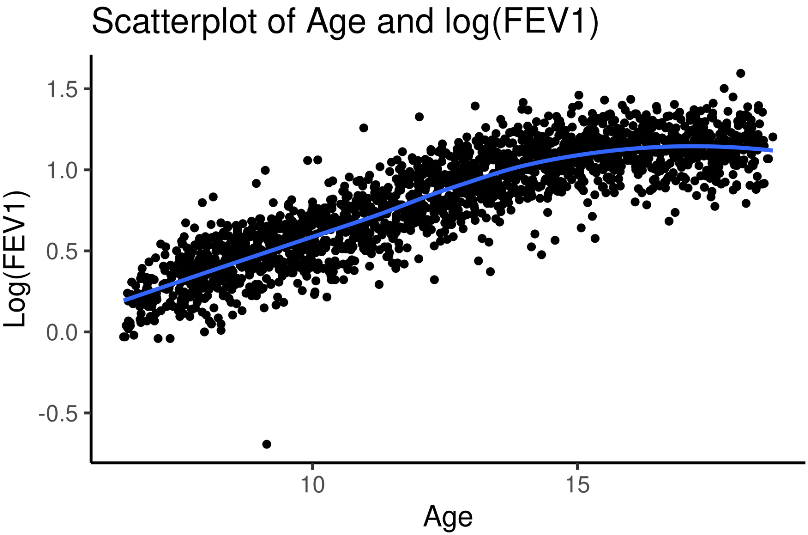
Case Study 1: Dental Growth



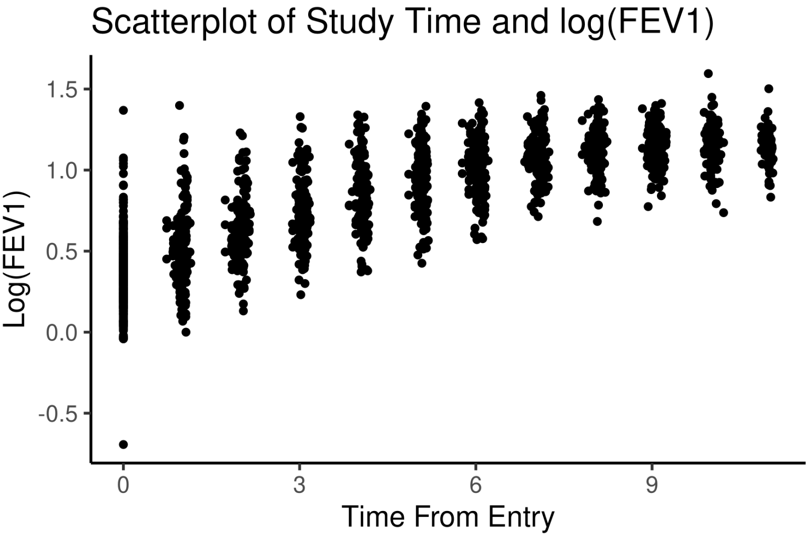
Case Study 2: Air Pollution and Health

- **Six Cities Study of Air Pollution and Health**
  - 300 school-age female children in Topeka, Kansas, most enrolled in 1st or 2nd grade (age 6-7)
  - Height, age, FEV1 (lung function) measured annually until high school graduation or loss to follow-up
  - Incomplete data (severity depends on what is considered "time 0")
- **Outcome of interest**
  - FEV1 (lung function)
- **Research questions**
  - How does lung function change as children age?
  - (Original study also compared more-polluted to less-polluted cities)

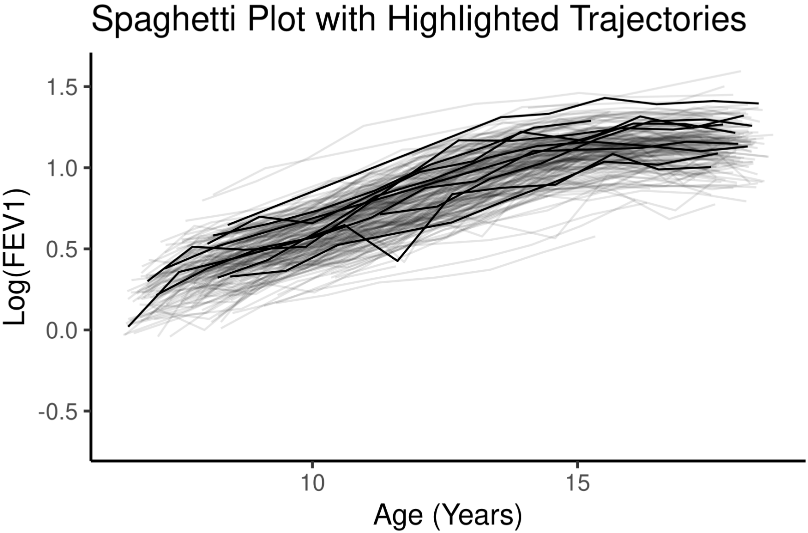
Case Study 2: Air Pollution and Health



Case Study 2: Air Pollution and Health



Case Study 2: Air Pollution and Health



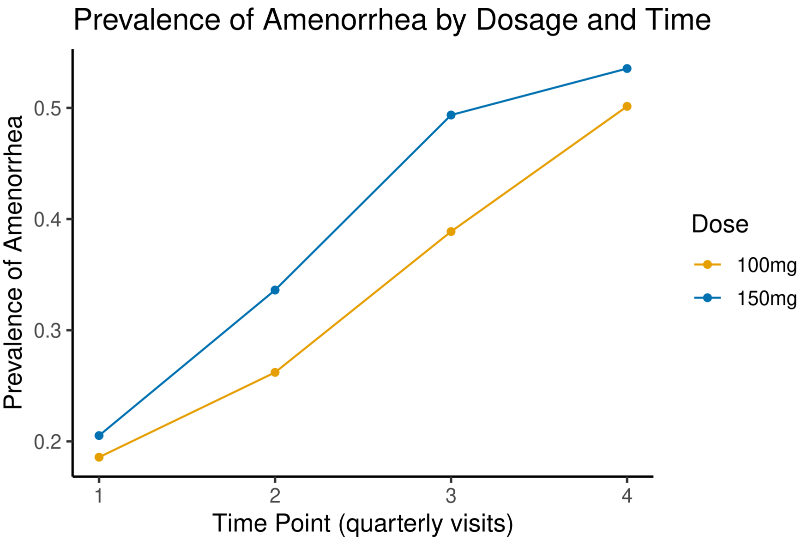
Case Study 3: Amenorrhea with Birth Control

- **Clinical Trial of Contracepting Women**
  - 1151 women randomized to either 100mg or 150mg of DMPA
  - Four injections given at 90-day intervals; final follow-up 90 days after last injection
  - Menstrual diary to record vaginal bleeding pattern disturbances
  - Substantial dropout: over 1/3 of participants dropped out before the study ended
- **Outcome of interest**
  - Amenorrhea (absence of menstrual bleeding) during each 3-month interval after an injection (note: this is binary!)
- **Research questions**
  - How do subject-specific risks of amenorrhea change over the course of the study?
  - What is the influence of dosage on amenorrhea risk?

Case Study 3: Amenorrhea with Birth Control

Visit 1	Visit 2	Visit 3	Visit 4	% among 100mg	% among 150mg
0	0	0	0	24.7	20.7
0	0	0	1	8.5	6.3
0	0	1	1	7.1	7.7
⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	□	3.5	1.9
0	0	1	□	2.3	1.7
⋮	⋮	⋮	⋮	⋮	⋮
0	□	□	□	13.2	11.8
1	□	□	□	4.0	5.4

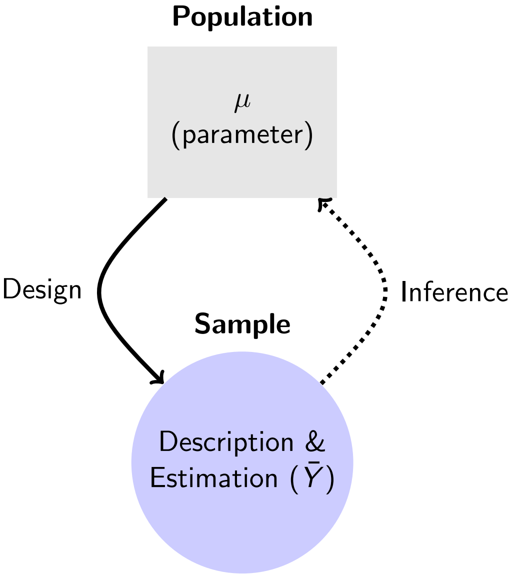
Case Study 3: Amenorrhea with Birth Control



Review of Key Stats/Regression Concepts

1. Big Picture of Statistics
2. Linear regression - interpretation and inference
3. Linear regression - effect modification

Big Picture of Statistics



Population

μ
(parameter)

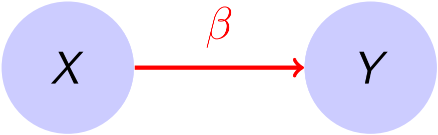
Design

Inference

Sample

Description &
Estimation ($\bar{Y}$)

Linear Regression - Estimation and Inference



$$E[Y \mid X = x] = \beta_0 + \beta_1 x$$

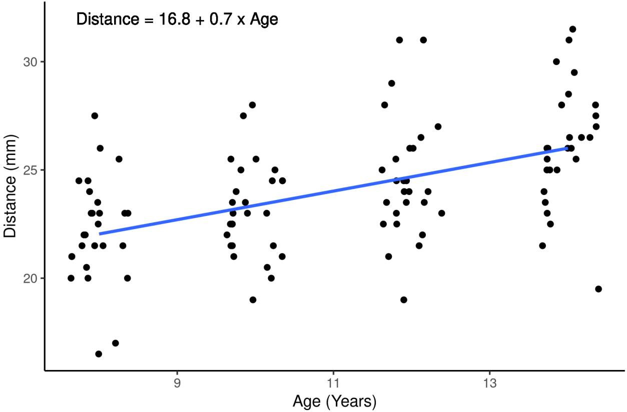
Estimation

- Coefficient estimates $\hat{\beta}$
- Standard errors for $\hat{\beta}$

Inference

- Confidence intervals for β
- Hypothesis tests for $\beta = 0$

Linear Regression - Interpretation (Dental Growth)



We estimate that, comparing two children one year apart in age, the average orthodontic distance is 0.7 mm longer for the older child.

Multiple Linear Regression

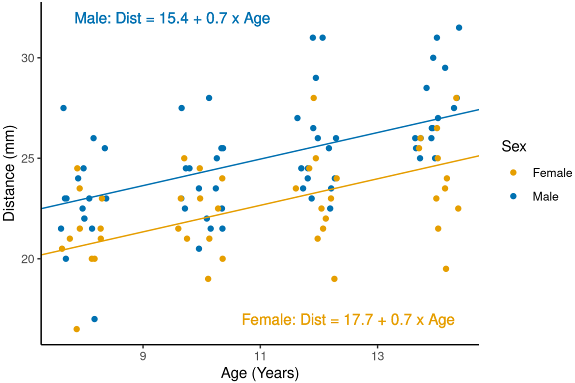
- Another variable affects the level of the outcome
- Mechanistically: "adjust" for additional variables

$$E[Y \mid x, t] = \beta_0 + \beta_1 x + \beta_2 t$$

- Results in parallel lines for groups based on x.
- E.g., if x is sex (0 = female, 1 = male):

$$\begin{aligned} \text{Female: } E[Y \mid x = 0, t] &= \beta_0 + \beta_2 t \\ \text{Male: } E[Y \mid x = 1, t] &= \beta_0 + \beta_1 + \beta_2 t \\ &= (\beta_0 + \beta_1) + \beta_2 t \end{aligned}$$

Multiple Linear Regression - Interpretation (Dental Growth)



We estimate that, comparing two children of the same sex, but one year apart in age, the average distance is 0.7 mm longer for the older child.

Linear Regression - Effect Modification

- Association of interest depends on value of another variable
- Mechanistically: interaction terms

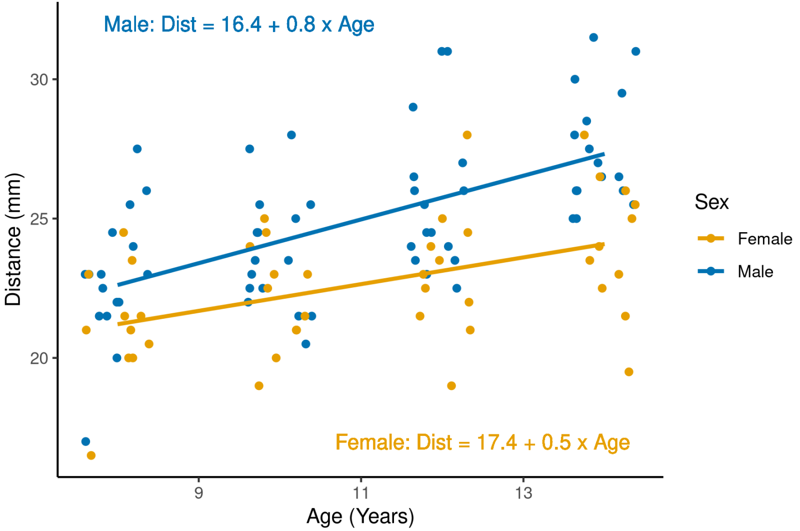
$$\mathrm{E}[Y \mid x, t] = \beta_0 + \beta_1 x + \beta_2 t + \beta_3 x \times t$$

- Results in separate models for groups based on x.
- E.g., if x is sex ($0 =$ female, $1 =$ male):

$$\begin{aligned} \text{Female: } \mathrm{E}[Y \mid x = 0, t] &= \beta_0 + \beta_2 t \\ \text{Male: } \mathrm{E}[Y \mid x = 1, t] &= \beta_0 + \beta_1 + \beta_2 t + \beta_3 t \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) t \end{aligned}$$

- Does association differ between females and males? $H_0 : \beta_3 = 0$

Linear Regression - Effect Modification (Dental Growth)



Effect Modification

Full equation:

$$\text{Distance} = 17.4 + 0.5 \times \text{Age} - 1.0 \times \text{Male} + 0.3 \times \text{Age} \times \text{Male}$$

Can be broken down into sex-specific equations:

- If child i is female,

$$E[\text{Dist}_{ij} | \text{Male}_i = 0, \text{Age}_{ij}] = 17.4 + 0.5 \times \text{Age}_{ij}$$

- If child i is male,

$$\begin{aligned} E[\text{Dist}_{ij} | \text{Male}_i = 1, \text{Age}_{ij}] &= (17.4 - 1.0) + (0.5 + 0.3) \times \text{Age}_{ij} \\ &= 16.4 + 0.8 \times \text{Age}_{ij} \end{aligned}$$

Effect Modification: Why Do We Care?

- Many longitudinal studies are interested in whether associations differ over time
- E.g., do placebo and treatment group have different disease progression trajectories?
- Interaction term (between treatment and time) tests this hypothesis

A Note About Notation

- n_i = number of observations for subject $i = 1, ..., N$
- $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^\top$ = outcome for subject i at times $j = 1, ..., n_i$
- $\mathbf{X}_i = \begin{bmatrix} X_{i11} & \cdots & X_{i1p} \\ X_{i21} & \ddots & X_{i2p} \\ \vdots & \ddots & \vdots \\ X_{in_i1} & \cdots & X_{in_ip} \end{bmatrix}$ = exposure/covariate matrix for subject i

A Note About Notation

- Mean of Y_{ij} is $\mu_{ij} = \mathrm{E}[Y_{ij}]$
- Variance of Y_{ij} is $\sigma_j^2 = \mathrm{E}[(Y_{ij} - \mu_{ij})^2]$
- Covariance between responses at time j and time k is $\sigma_{jk} = \mathrm{E}[(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})]$
- So for subject i (and n observations), the full variance-covariance matrix is

$$\boldsymbol{\Sigma}_i = \mathrm{Cov}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$$

First half of course: Methods for continuous outcomes

- General linear models
- Linear mixed models
- Data analysis: Multicenter AIDS Cohort Study (MACS)
  - Model CD4+ data over time; various levels of R scaffolding
  - Individual or group
  - Stay on Zoom (breakout rooms) if you want to ask questions in person
  - We'll also be monitoring Slack as time permits

Overview

Day 1: Monday, July 10

Introduction

Activity

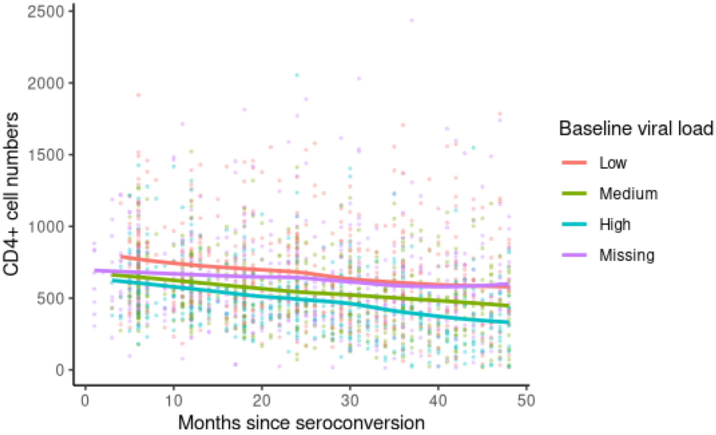
General Linear Model

Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

Activity: MACS Data Exploration

- Multicenter Aids Cohort Study
- Goal: Characterize time course of CD4+ T-cell depletion



Activity Guidelines

- Prefer to work alone? Feel free to stay here (in case you have questions) or ask on Slack
- Prefer to work with others? Join a breakout room and call us in if your group has questions

Overview

Day 1: Monday, July 10

- Introduction
- Activity
- General Linear Model

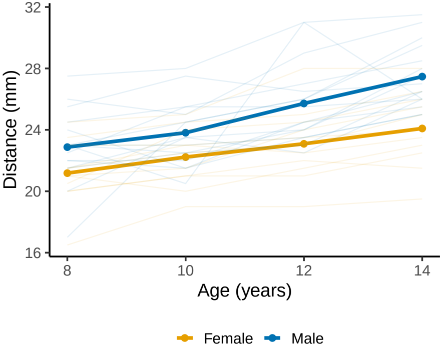
Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

Case Study: Dental Growth

Goals:

1. Estimate average growth curve for all children
2. Estimate growth curves for individual children
3. Characterize heterogeneity in children's growth rates
4. Assess whether the growth rate differs between boys and girls



Case Study: Dental Growth

Goals:

1. Estimate average growth curve for all children
   - ✓ We will focus on this
2. Estimate growth curves for individual children
   - X We will look into this when we talk about **linear mixed models**
3. Characterize heterogeneity in children's growth rates
   - X We will look into this when we talk about **linear mixed models**
4. Assess whether the growth rate differs between boys and girls
   - ✓ We will focus on this

Case Study: Dental Growth

Mean model:

$$E[\text{Dist}_{ij}|\text{Male}_i, \text{Age}_{ij}] = \beta_0 + \beta_1(\text{Age}_{ij} - 8) + \beta_2\text{Male}_i + \beta_3(\text{Age}_{ij} - 8) \times \text{Male}_i$$

So the **sex-specific mean models** are:

- If child i is a girl:

$$E[\text{Dist}_{ij}|\text{Male}_i = 0, \text{Age}_{ij}] = \beta_0 + \beta_1(\text{Age}_{ij} - 8)$$

- If child i is a boy:

$$E[\text{Dist}_{ij}|\text{Male}_i = 1, \text{Age}_{ij}] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)(\text{Age}_{ij} - 8)$$

Case Study: Dental Growth

Mean model:

$$E[\text{Dist}_{ij}|\text{Male}_i, \text{Age}_{ij}] = \beta_0 + \beta_1(\text{Age}_{ij} - 8) + \beta_2\text{Male}_i + \beta_3(\text{Age}_{ij} - 8) \times \text{Male}_i$$

Covariance: $\text{Cov}[\mathbf{Dist}_i \mid \text{Male}_i,\ \mathbf{Age}_i] = \boldsymbol{\Sigma}_i$

Working covariance model:

$$\text{Cov}[\mathbf{Dist}_i \mid \text{Male}_i,\ \mathbf{Age}_i] = \mathbf{V}_i = \sigma^2\mathbf{R}_i$$

where $\mathbf{R}_i$ is the working correlation model

Correlation Models

Independence: $\text{Corr}[Y_{ij}, Y_{ij'} \mid X_i] = 0$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Exchangeable: $\text{Corr}[Y_{ij}, Y_{ij'} \mid X_i] = \alpha$

$$\begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix}$$

Correlation Models

Auto-regressive: $\text{Corr}[Y_{ij}, Y_{ij'} \mid X_i] = \alpha^{|j-j'|}$

$$\begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n-1} & \alpha^{n-2} & \alpha^{n-3} & \cdots & 1 \end{bmatrix}$$

Unstructured: $\text{Corr}[Y_{ij}, Y_{ij'} \mid X_i] = \alpha_{jj'}$

$$\begin{bmatrix} 1 & \alpha_{21} & \alpha_{31} & \cdots & \alpha_{n1} \\ \alpha_{12} & 1 & \alpha_{32} & \cdots & \alpha_{n2} \\ \alpha_{13} & \alpha_{23} & 1 & \cdots & \alpha_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{1n} & \alpha_{2n} & \alpha_{3n} & \cdots & 1 \end{bmatrix}$$

Correlation models

Correlation between any two observations on the same subject. . .

- **Independence**: . . . is assumed to be zero
  - Appropriate with use of robust variance estimator (large N)
- **Exchangeable**: . . . is assumed to be constant
  - More appropriate for clustered data
- **Auto-regressive**: . . . is assumed to depend on time or distance
  - More appropriate for equally-spaced longitudinal data
- **Unstructured**: . . . is assumed to be distinct for each pair
  - Only appropriate for short series (small n) on many subjects (large N)

General Linear Model

Goal: How can we estimate and make inference on parameters, $\boldsymbol{\beta}$, in a linear model in the presence of correlation?

For individual i, $i = 1, \ldots, N$:

$$E[\mathbf{Y}_i|\mathbf{X}_i] = \mathbf{X}_i\boldsymbol{\beta}$$
$$\text{Cov}[\mathbf{Y}_i|\mathbf{X}_i] = \boldsymbol{\Sigma}_i$$

- Review the setting of independent responses
  - Approach: ordinary least squares
  - Considerations: robust standard errors
- Extend to the setting of correlated responses
  - Approach: multivariate weighted least squares
  - Considerations: robust standard errors

Notation Reminder

Notation

Independent responses:

$$E[Y_i|\mathbf{X}_i] = \mathbf{X}_i\boldsymbol{\beta}$$
$$\text{Var}[Y_i|\mathbf{X}_i] = \sigma^2$$

- Suppose we have one measurement on subject i, $i = 1, \dots, N$
- Y_i = outcome for subject i
- $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$= exposure/covariate vector for subject i

Correlated responses:

$$E[\mathbf{Y}_i|\mathbf{X}_i] = \mathbf{X}_i\boldsymbol{\beta}$$
$$\text{Cov}[\mathbf{Y}_i|\mathbf{X}_i] = \boldsymbol{\Sigma}_i$$

- Suppose we have n_i measurements on subject i, $i = 1, \dots, N$
- $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^\top =$ outcome for subject i at times $j = 1, ..., n_i$
- $\mathbf{X}_i = \begin{bmatrix} X_{i11} & \cdots & X_{i1p} \\ X_{i21} & \ddots & X_{i2p} \\ \vdots & \ddots & \vdots \\ X_{in_i1} & \cdots & X_{in_ip} \end{bmatrix} =$ exposure/covariate matrix for subject i

Estimation and Inference

	Independent responses:	**Correlated responses**:
Setup	$E[Y_i \mid \mathbf{X}_i] = \mathbf{X}_i\beta$ $\text{Var}[Y_i \mid \mathbf{X}_i] = \sigma^2$	$E[\mathbf{Y}_i \mid \mathbf{X}_i] = \mathbf{X}_i\beta$ $\text{Cov}[\mathbf{Y}_i \mid \mathbf{X}_i] = \mathbf{\Sigma}_i$
Estimation Procedure	We can use the method of **ordinary least squares** to find estimates for β, which involves minimizing: $\sum_{i=1}^{N}(Y_i - \mathbf{X}_i\beta)^2$	We can use the method of multivariate **weighted least squares** to find estimates for β, which involves minimizing: $\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{X}_i\beta)^\top \mathbf{W}_i(\mathbf{Y}_i - \mathbf{X}_i\beta)$ where $\mathbf{W}_i$ is a weight matrix

Estimation and Inference

Estimating Equations

To find $\hat{\boldsymbol{\beta}}$ we take the derivatives (with respect to $\beta_0,\ \beta_1, \ldots, \beta_p$), set the functions equal to 0 to give the following **estimating equations** for $\boldsymbol{\beta}$:

Independent responses:

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{X}_i^\top (Y_i - \mathbf{X}_i \hat{\beta})$$

Therefore,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i^\top Y_i$$

Correlated responses:

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta})$$

Therefore,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{Y}_i$$

$\hat{\beta}$

Estimation and Inference

Variance of $\hat{\beta}$

Independent responses:

$$\mathrm{Cov}(\hat{\beta}) = \underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{X}_i\right)^{-1}}_{\text{bread}} \times$$

$$\underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathrm{Var}(Y_i) \mathbf{X}_i\right)}_{\text{cheese}} \times$$

$$\underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{X}\right)^{-1}}_{\text{bread}}$$

Correlated responses:

$$\mathrm{Cov}(\hat{\beta}) = \underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i\right)^{-1}}_{\text{bread}} \times$$

$$\underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathrm{Cov}(\mathbf{Y}_i) \mathbf{W}_i \mathbf{X}_i\right)}_{\text{cheese}} \times$$

$$\underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}\right)^{-1}}_{\text{bread}}$$

Estimation and Inference

Summary

Independent responses:

- Assumes that observations are **independent**
- Do not need to have an assumption that the errors have a normal distribution
- We can use a **robust variance estimate** if we do not want to assume constant σ^2 (does require large enough sample size to work)

Correlated responses:

- Assumes that observations are **independent** across individuals but there may be **correlation** between observations on the same individual
- Do not need to have an assumption that the errors have a (multivariate) normal distribution
- We can use a **robust variance estimate** if we do not want to assume correctly specified $\boldsymbol{\Sigma}$ (more on this...)

Properties of $\hat{\boldsymbol{\beta}}$

- Unbiased given $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ and $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_N$

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i E[\mathbf{Y}_i]\right) \\ &= \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i \boldsymbol{\beta}\right) \\ &= \boldsymbol{\beta} \end{aligned}$$

- The variance of $\hat{\boldsymbol{\beta}}$ depends on the weights:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathrm{Cov}(\mathbf{Y}_i) \mathbf{W}_i \mathbf{X}_i\right) \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{X}\right)^{-1}$$

Summary

- Any set of reasonable ('positive definite') weights provides a valid estimator
- $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$, which loosely means it 'hones' in on the truth as the sample size N gets larger no matter what the weights are!
- When $\mathbf{W}_i = \boldsymbol{\Sigma}_i^{-1}$, we get an estimator for $\boldsymbol{\beta}$ that is most efficient (among linear estimators)
- In the GEE approach, we will specify a form for the weights which may depend on unknown parameters ($\boldsymbol{\alpha}$) that need to be estimated

GEE approach

Setting: For linear regression of Y_{ij} on covariates $X_{ij1}, \ldots, X_{ijp}$, i.e. with mean model:

$$\begin{aligned} E[Y_{ij}|X_{ij}] &= \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} \\ &= \mathbf{X}_{ij}\boldsymbol{\beta} \quad j = 1, \ldots, n_i;\; i = 1, \ldots, N \end{aligned}$$

and a **working covariance matrix V**.

Algorithm:

1. Fit the weighted linear regression where the weights of each cluster are the inverse of their current working covariances ($\mathbf{W}_i = \hat{\mathbf{V}}_i^{-1}$)

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^\top \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i^\top \hat{\mathbf{V}}_i^{-1} \mathbf{Y}_i$$

2. Update the working covariances using the residuals obtained from fitting the model in (1), $Y_{ij} - \mathbf{X}_{ij}\hat{\boldsymbol{\beta}}$
3. Iterate steps (1) and (2) until the result converges

GEE approach

The resulting estimator $\hat{\beta}$, the solution to the generalized estimating equations, is robust:

- The regression coefficient estimates will be correct (in large samples) even if the working covariance model does not match the true covariance model
- However, to obtain valid standard errors, we must capture the correlation in the data, either through choosing the correct covariance model, or using an alternative variance estimate
- Correctly specified (or close to the truth) covariance model will yield regression estimate $\hat{\beta}$ that is most (or more) efficient (smaller variance)

Robust Variance Estimate

- GEE computes a sandwich variance estimator
  - Aka empirical variance, robust variance, Huber-White correction
  - This is the default standard error output when using `geeglm()`
- Empirical variance gives valid standard errors for the estimated regression coefficients even if the working covariance model was wrong
  - Valid in large samples (this means it can be used with data sets that contain at least 40 subjects)

$$\widehat{\text{Cov}}(\hat{\beta}) = \underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i\right)^{-1}}_{\text{bread}} \underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \hat{\mathbf{V}}_i^{-1} \widetilde{\text{Cov}}(\mathbf{Y}_i) \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i\right)}_{\text{cheese}} \underbrace{\left(\sum_{i=1}^{N} \mathbf{X}_i^\top \hat{\mathbf{V}}_i^{-1} \mathbf{X}\right)^{-1}}_{\text{bread}}$$

where we can use the residuals, i.e., $\widetilde{\text{Cov}}(\mathbf{Y}_i)$ has the entries $e_{ij}e_{ik}$ where $e_{ij} = Y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{ij1} - \hat{\beta}_2 X_{ij2} - \cdots - \hat{\beta}_p X_{ijp}$

GEE: Inference

With the GEE approach, we can perform Wald tests and construct Wald confidence intervals:

- $\hat{\beta}_k/\text{s.e.}$ - valid test
- $\hat{\beta}_k \pm 1.96 \times \text{s.e.}$ - valid 95% confidence interval

Cannot perform a likelihood ratio test (no fully specified probability model) so no AIC or BIC either.

Case Study: Dental Growth

Mean model:

$$E[\text{Dist}_{ij}|\text{Male}_i, \text{Age}_{ij}] = \beta_0 + \beta_1(\text{Age}_{ij} - 8) + \beta_2\text{Male}_i + \beta_3(\text{Age}_{ij} - 8) \times \text{Male}_i$$

Working covariance model:

$$\text{Cov}[\mathbf{Dist}_i \mid \text{Male}_i,\ \mathbf{Age}_i] = \sigma^2\mathbf{R}_i$$

where we will show results for $\mathbf{R}_i$:

- Independent: [`corstr = "independence"`]
- Exchangeable: [`corstr = "exchangeable"`]
- Auto-regressive: [`corstr = "ar1"`]
- Unstructured: [`corstr = "unstructured"`]

Case Study: Dental Growth

R Code:

```
library(geepack)
m_ind <- geeglm(distance ~ I(age-8)*Sex, id = Subject,
                data = Orthodont, corstr = "independence")
m_exc <- geeglm(distance ~ I(age-8)*Sex, id = Subject,
                data = Orthodont, corstr = "exchangeable")
m_ar1 <- geeglm(distance ~ I(age-8)*Sex, id = Subject,
                data = Orthodont, corstr = "ar1")
m_uns <- geeglm(distance ~ I(age-8)*Sex, id = Subject,
                data = Orthodont, corstr = "unstructured")

m_ols <- lm(distance ~ I(age-8)*Sex, data=Orthodont)
```

Case Study: Dental Growth

```
geeglm(formula = distance ~ I(age - 8) * Sex, data = Orthodont,
id = Subject, corstr = "independence")

Coefficients:
Estimate Std.err   Wald Pr(>|W|)
(Intercept)          21.2091  0.5604 1432.2  < 2e-16 ***
I(age - 8)            0.4795  0.0631   57.7  3.1e-14 ***
SexMale               1.4065  0.7738    3.3   0.0691 .
I(age - 8):SexMale   0.3048  0.1169    6.8   0.0091 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

Estimate Std.err
(Intercept)     4.91    1.01
Number of clusters:   27  Maximum cluster size: 4
```

Case Study: Dental Growth

```
geeglm(formula = distance ~ I(age - 8) * Sex, data = Orthodont,
id = Subject, corstr = "exchangeable")

Coefficients:
Estimate Std.err   Wald Pr(>|W|)
(Intercept)         21.2091  0.5604 1432.2  < 2e-16 ***
I(age - 8)           0.4795  0.0631   57.7  3.1e-14 ***
SexMale              1.4065  0.7738    3.3   0.0691 .
I(age - 8):SexMale   0.3048  0.1169    6.8   0.0091 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

Estimate Std.err
(Intercept)     4.91    1.01
Link = identity

Estimated Correlation Parameters:
Estimate Std.err
alpha    0.618   0.131
Number of clusters:   27  Maximum cluster size: 4
```

Case Study: Dental Growth

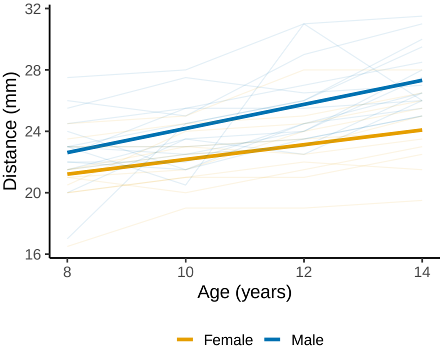
	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)
Ind	21.2 (0.56)	0.5 (0.06)	1.4 (0.77)	0.3 (0.12)
Exch	21.2 (0.56)	0.5 (0.06)	1.4 (0.77)	0.3 (0.12)
AR1	21.2 (0.59)	0.5 (0.06)	1.6 (0.83)	0.3 (0.12)
Unst	21.2 (0.55)	0.5 (0.06)	1.4 (0.76)	0.3 (0.12)
OLS	21.2 (0.57)	0.5 (0.15)	1.4 (0.74)	0.3 (0.20)

- Working independence and OLS give exactly the same point estimates
  - The estimating equations and thus estimator are exactly the same
- OLS standard errors too large for $\hat{\beta}_1$ and $\hat{\beta}_3$
  - This is because age varies within-subject
  - Inference using OLS would be wrong
- Working independence and exchangeable provide exactly the same results
  - Data are balanced and complete
- Unstructured results are similar
- Autoregressive results are slightly (but not importantly) different

Case Study: Dental Growth

Results from working independence:

- $\hat{\beta}_1$: The estimated rate of change in mean dental length per year for female children is 0.48 mm/year (95% CI: 0.36, 0.60 mm/year)
- $\hat{\beta}_1 + \hat{\beta}_3$: The estimated rate of change in mean dental length per year for male children is 0.78 mm/year (95% CI: 0.59, 0.98 mm/year)

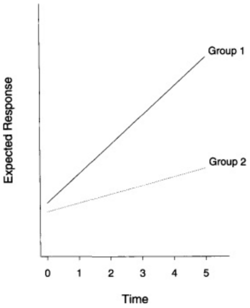


Dental Growth: Did We Answer Our Questions?

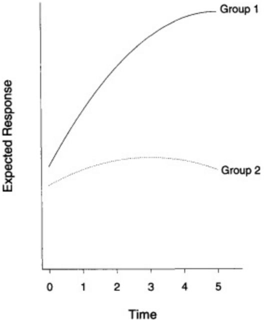
1. Estimate average growth curve for all children
   - Coefficients $\hat{\beta}$ from our mean model
2. Estimate growth curves for individual children
   - See next section
3. Characterize heterogeneity in children's growth rates
   - See next section
4. Assess whether the growth rate differs between boys and girls
   - It does! Interaction term is significantly nonzero (p-value < 0.01)
   - (and scientifically interesting - growth rate 0.8 mm/yr for boys, 0.5 mm/yr for girls)

Modeling the Mean Response over Time

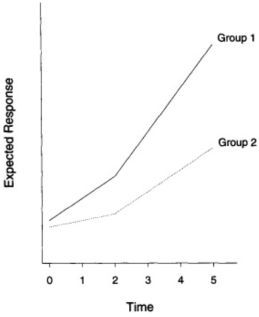
Linear



Quadratic



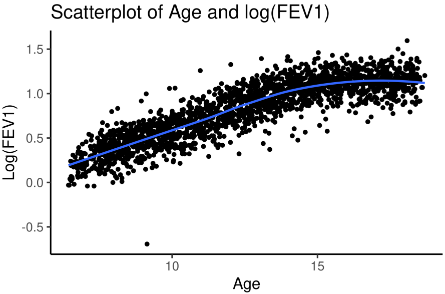
Linear Splines



(Images from Fitzmaurice, Garrett M., Nan M. Laird, and James H. Ware. *Applied longitudinal analysis.*)

Case Study: Six Cities

- 300 school-age female children, most enrolled around ages 6-7
- Height, age, FEV1 (lung function) measured approximately annually until high school graduation or loss to follow-up
- **Goal**: Explore various ways to model mean FEV1 as children age



Linear Trend

Notation:

- logFEV1_{ij}: log FEV1 for subject i at measurement occasion j
- age_{ij}: age of subject i at measurement occasion j

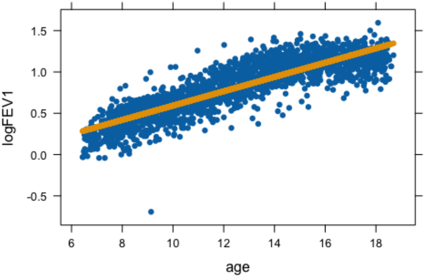
Mean Model:

$$E[\text{logFEV1}_{ij}|\text{age}_{ij}] = \beta_0 + \beta_1\text{age}_{ij}$$

In this model, the **rate of change** is constant (β_1)

R Code:

```
geeglm(logFEV1 ~ age,
        id = id, data = dat)
```



Quadratic Trend

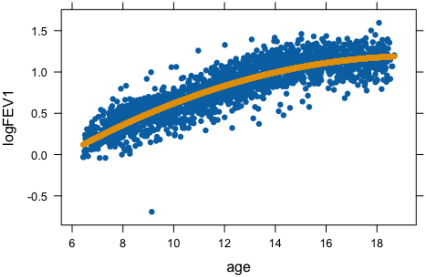
Mean Model:

$$E[\text{logFEV1}_{ij} | \text{age}_{ij}] = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{age}_{ij}^2$$

In this model, the **rate of change** is non-constant $(\beta_1 + 2\beta_2 \text{age}_{ij})$

R Code:

```
geeglm(logFEV1 ~ age + I(age^2),
        id = id, data = dat)
```



Linear Splines

Notation:

- $(\text{age}_{ij} - \text{age}_k)_+ = \max(\text{age}_{ij} - \text{age}_k, 0)$: linear spline based on knot age_k

Mean Model:

$$E[\text{logFEV1}_{ij}|\text{age}_{ij}] = \beta_0 + \beta_1\text{age}_{ij} + \beta_2(\text{age}_{ij} - 10)_+ + \beta_3(\text{age}_{ij} - 14)_+$$

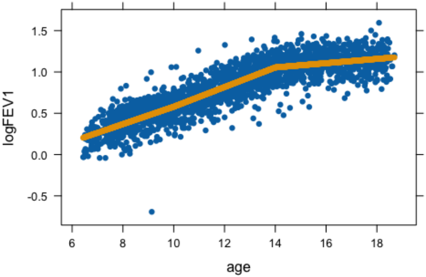
In this model, the **rate of change** is:

- β_1: for $\text{age}_{ij} < 10$
- $\beta_1 + \beta_2$: for $10 \le \text{age}_{ij} < 14$
- $\beta_1 + \beta_2 + \beta_3$: for $\text{age}_{ij} \ge 14$

R Code:

```
dat$ageSpline10 <- pmax(dat$age - 10, 0)
dat$ageSpline14 <- pmax(dat$age - 14, 0)

geeglm(logFEV1 ~ age + ageSpline10 +
        ageSpline14,
        id = id, data = dat)
```



Modeling the Mean Response over Time

- Mean models as regression with time, and perhaps additional functions of time
- Polynomial models:
  - Quadratic: $\beta_1 t_{ij} + \beta_2 t_{ij}^2$
  - Others: higher order polynomials
  - Allow for non-linear mean curve
  - Allow for non-constant rate of change
- Regression splines:
  - Linear splines: $\beta_1 t_{ij} + \beta_2 (t_{ij} - 14)_+$
  - Others: cubic splines
  - Allow for non-linear mean curve
  - Allow for non-constant rate of change

Assumptions

Valid inference from a general linear model relies on

- **Mean model**: As with any regression model for an average outcome, need to correctly specify the functional form of $\mathbf{X}_{ij}\beta$
  - Included important covariates in the model
  - Correctly specified any transformations or interactions
- **Covariance model**: Correct covariance model is required for correct standard error estimates for $\hat{\beta}$ if using **model-based variance estimate** otherwise we can use **robust/empirical variance estimate**
- N sufficiently large

Summary

- Primary focus is on the mean model
- Longitudinal correlation is secondary to mean model of interest and is treated as a nuisance
- Requires selection of a 'working' correlation model
- Semi-parametric model: mean + correlation
- Working correlation model does not need to be correctly specified to obtain consistent estimator for $\boldsymbol{\beta}$ or valid standard errors for $\hat{\boldsymbol{\beta}}$ but efficiency gains possible if correlation model is correct
- Wald testing

Issues:

- Accommodates only one source of correlation: longitudinal or cluster
- Requires that any missing data are missing completely at random
- Issues arise with time-dependent exposures and covariance weighting

End of Day 1

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

- Recap
- Linear Mixed Model
- Wrap-Up: Methods for Continuous Outcomes
- Activity
- Review of Generalized Linear Models
- Generalized Estimating Equations

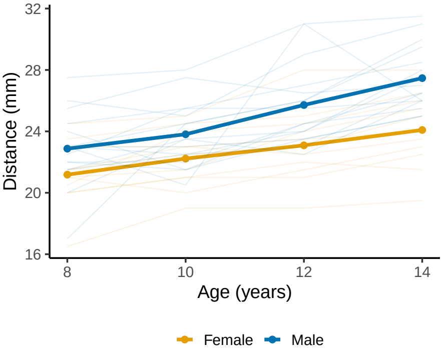
Day 3: Wednesday, July 12

Recap: Notation

- n_i = number of observations for subject $i = 1, ..., N$
- $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^\top$ = outcome for subject i at times $j = 1, ..., n_i$
- $\mathbf{X}_i = \begin{bmatrix} X_{i11} & \cdots & X_{i1p} \\ X_{i21} & \ddots & X_{i2p} \\ \vdots & \ddots & \vdots \\ X_{in_i1} & \cdots & X_{in_ip} \end{bmatrix}$ = exposure/covariate matrix for subject i

Recap: Dental Growth

- $N = 27$ children (11 girls and 16 boys)
- Each was measured $n_i = 4$ times (ages 8, 10, 12, and 14)
- Dental distance was measured at each age
- Questions of interest we have so far focused on:
  - What is the growth trajectory in girls and boys?
  - Does the growth rate differ between boys and girls?



Recap: Linear regression using the GEE approach

- Primary focus of the analysis is a marginal mean regression model (*so far we have focused on linear models, we'll see later today how to extend this to generalized linear models*)

$$\begin{aligned} E[Y_{ij}|X_{ij}] &= \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} \\ &= \mathbf{X}_{ij}\boldsymbol{\beta} \quad j = 1, \ldots, n_i; \; i = 1, \ldots, N \end{aligned}$$

  - Dental growth case study:

    $$E[\text{Dist}_{ij}|\text{Male}_i, \text{Age}_{ij}] = \beta_0 + \beta_1(\text{Age}_{ij} - 8) + \beta_2 \text{Male}_i + \beta_3(\text{Age}_{ij} - 8) \times \text{Male}_i$$

- Longitudinal correlation is secondary to the mean model of interest and is treated as a nuisance feature of the data
- Requires selection of a 'working' correlation model
  - E.g. working independence, working unstructured

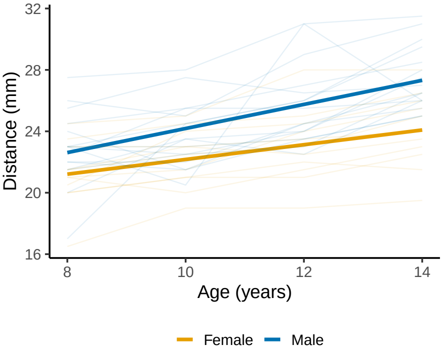
Recap: Linear regression using the GEE approach

- In the GEE approach there is no likelihood (e.g., do not assume normality), so how do we obtain estimates for $\boldsymbol{\beta}$?
  - By constructing an unbiased estimating function (*we'll see this more today*)
- Properties of our estimator, $\hat{\boldsymbol{\beta}}$?
  - Consistent estimator for $\boldsymbol{\beta}$ even if working correlation model incorrect
  - Can obtain valid standard errors for $\hat{\boldsymbol{\beta}}$ using the sandwich variance estimator
  - Efficiency gains possible if correlation model is correct
- Note: Wald statistics for hypothesis testing

Recap: Dental Growth

Results from working independence:

- $\hat{\beta}_1$: The estimated rate of change in mean dental length per year for female children is 0.48 mm/year (95% CI: 0.36, 0.60 mm/year)
- $\hat{\beta}_1 + \hat{\beta}_3$: The estimated rate of change in mean dental length per year for male children is 0.78 mm/year (95% CI: 0.59, 0.98 mm/year)



Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

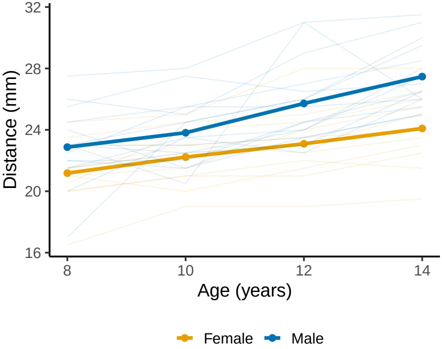
- Recap
- Linear Mixed Model
- Wrap-Up: Methods for Continuous Outcomes
- Activity
- Review of Generalized Linear Models
- Generalized Estimating Equations

Day 3: Wednesday, July 12

Case Study: Dental Growth

Goals:

1. Estimate average growth curve for all children
2. Estimate growth curves for individual children
3. Characterize heterogeneity in children's growth rates
4. Assess whether the growth rate differs between boys and girls



Case Study: Dental Growth

Goals:

1. Estimate average growth curve for all children
   - ✓ GEE and LMM can do this
2. Estimate growth curves for individual children
   - X GEE isn't meant to do this - will address with LMM
3. Characterize heterogeneity in children's growth rates
   - X GEE is not great at this - will address with LMM
4. Assess whether the growth rate differs between boys and girls
   - ✓ GEE and LMM can do this

Population-Averaged vs. Individual-Specific Models

- GEE coefficients have **population-averaged** interpretations
  - E.g., average dental length for male children increases by 0.3mm more per year than average dental length for female children
- What if we want **individual-specific** trajectories?
  - Not enough data to estimate separate regression lines for everyone
  - Instead, assume that each subject has a regression model that includes **fixed effect** parameters common to everyone, and subject-specific parameters (**random effects**) that follow some distribution
- Subject-specific random effects also induce a correlation structure

Linear Mixed Effects Model (Simplest Case)

The simplest format for a mixed effects model is:

$$
\begin{aligned}
Y_{ij} = \mu_j & \qquad \text{Shared mean model} \\
+ b_i & \qquad \text{Random intercept for subject } i \\
+ \epsilon_{ij} & \qquad \text{Measurement error}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathrm{Var}(b_i) = \tau^2 & \qquad \text{Between-person variation} \\
\mathrm{Var}(\epsilon_{ij}) = \sigma^2 & \qquad \text{Within-person variation} \\
b_i \perp \epsilon_{ij} &
\end{aligned}
$$

Induced Correlation Structure

Random effects implicitly specify covariance structure:

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{ik}) &= \text{Cov}(\mu_j + b_i + \epsilon_{ij}, \mu_k + b_i + \epsilon_{ik}) \\ &= \text{Cov}(b_i + \epsilon_{ij}, b_i + \epsilon_{ik}) \\ &= \text{Cov}(b_i, b_i) + \text{Cov}(b_i, \epsilon_{ik}) + \text{Cov}(\epsilon_{ij}, b_i) + \text{Cov}(\epsilon_{ij}, \epsilon_{ik}) \\ &= \text{Var}(b_i) + 0 + 0 + 0 = \tau^2\end{aligned}$$

and similarly

$$\begin{aligned}\text{Var}(Y_{ij}) &= \text{Var}(\mu_j + b_i + \epsilon_{ij}) \\ &= \text{Var}(b_i + \epsilon_{ij}) \\ &= \text{Var}(b_i) + \text{Var}(\epsilon_{ij}) + \text{Cov}(b_i, \epsilon_{ij}) \\ &= \sigma^2 + \tau^2 + 0\end{aligned}$$

Induced Correlation Structure

From last slide, we have:

$$\begin{aligned}\mathrm{Cov}(Y_{ij}, Y_{ik}) &= \mathrm{Var}(b_i) + 0 + 0 + 0 = \tau^2 \\ \mathrm{Var}(Y_{ij}) &= \sigma^2 + \tau^2 + 0\end{aligned}$$

And since subjects are independent, each subject's covariance matrix is:

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \cdots & \tau^2 & \sigma^2 + \tau^2 \end{pmatrix}$$

Induced Correlation Structure

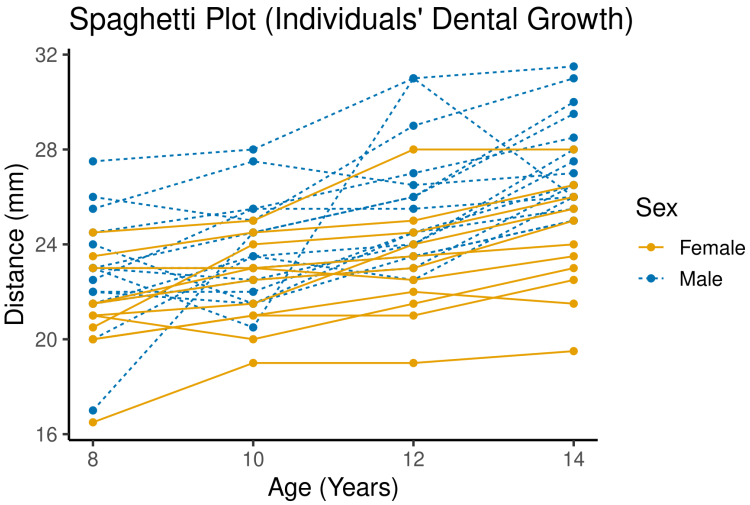
This looks awfully similar to exchangeable correlation – and in fact, it is:

$$\mathbf{\Sigma}_i = (\sigma^2 + \tau^2) \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}$$

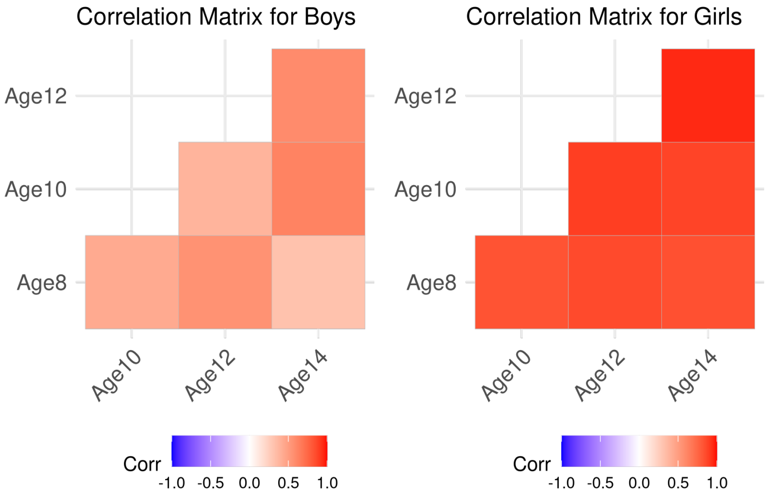
$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

- ρ is called the "intraclass correlation coefficient"
- Ratio of between-person variation to total variation

Example: Dental Growth



Example: Dental Growth



Example: Dental Growth

Random intercept model with interaction between sex and age:

$$\begin{aligned} \text{Dist}_{ij} &= \beta_0 + \beta_1 \times (\text{Age}_{ij} - 8) + \beta_2 \times \text{Male}_i \\ &\quad + \beta_3 \times (\text{Age}_{ij} - 8) \times \text{Male}_i + b_i \end{aligned}$$

So the sex- and subject-specific models are:

- If child i is a girl:

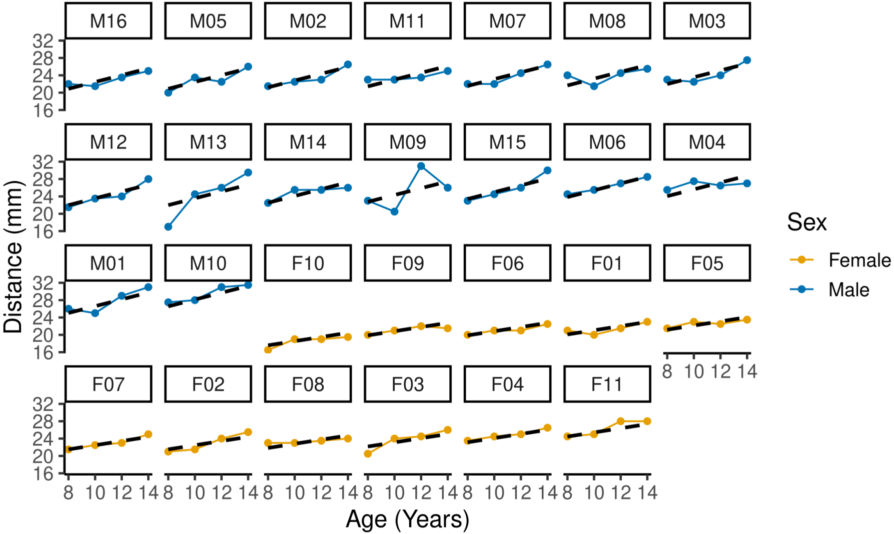
  $$E[\text{Dist}_{ij} | \text{Age}_{ij}, \text{Male}_i = 0] = (\beta_0 + b_i) + \beta_1 \times (\text{Age}_{ij} - 8)$$

- If child i is a boy:

  $$E[\text{Dist}_{ij} | \text{Age}_{ij}, \text{Male}_i = 1] = (\beta_0 + \beta_2 + b_i) + (\beta_1 + \beta_3) \times (\text{Age}_{ij} - 8)$$

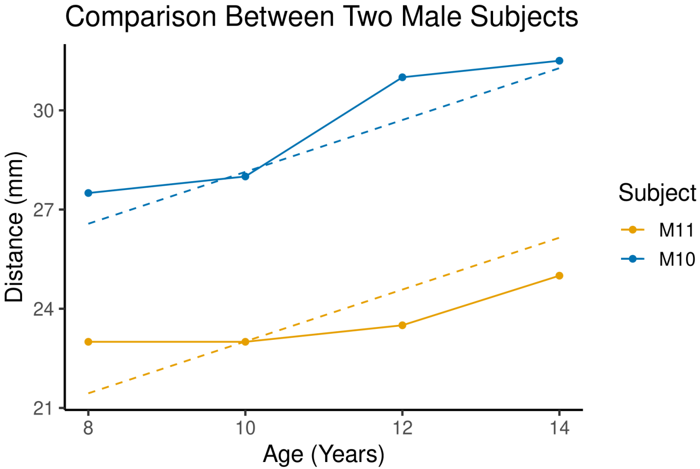
Example: Dental Growth

Visualization of results:



Example: Dental Growth

Zooming in on just two subjects:



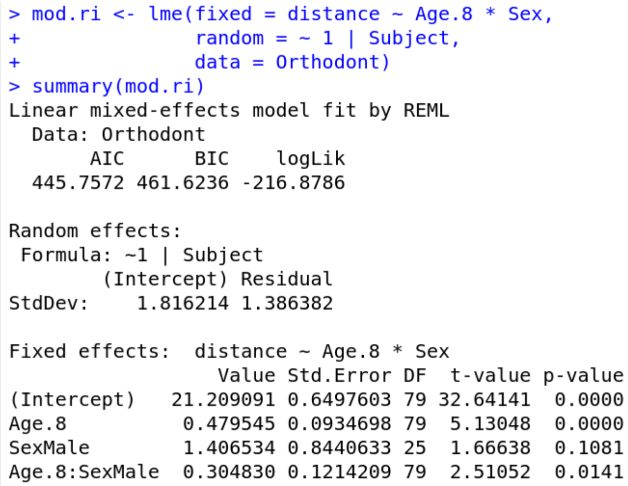
Example: Dental Growth

Observations:

- The two lines are parallel (slope can't vary by subject)
- M11 has a shorter length at every time than M10
- Variability within subject (around their line) looks smaller than variability between subjects
  - Intraclass correlation coefficient will help quantify this

Example: Dental Growth

R output:



```
> mod.ri <- lme(fixed = distance ~ Age.8 * Sex,
+               random = ~ 1 | Subject,
+               data = Orthodont)
> summary(mod.ri)
Linear mixed-effects model fit by REML
  Data: Orthodont
       AIC      BIC    logLik
  445.7572 461.6236 -216.8786

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.816214 1.386382

Fixed effects:  distance ~ Age.8 * Sex
                  Value Std.Error DF  t-value p-value
(Intercept)   21.209091 0.6497603 79 32.64141  0.0000
Age.8          0.479545 0.0934698 79  5.13048  0.0000
SexMale        1.406534 0.8440633 25  1.66638  0.1081
Age.8:SexMale  0.304830 0.1214209 79  2.51052  0.0141
```

Example: Dental Growth

Systematic part of model:

$$\widehat{\text{Dist}}_{ij} = 21.2 + 0.5 \times (\text{Age}_{ij} - 8) + 1.4 \times \text{M}_i + 0.3 \times (\text{Age}_{ij} - 8) \times \text{M}_i$$

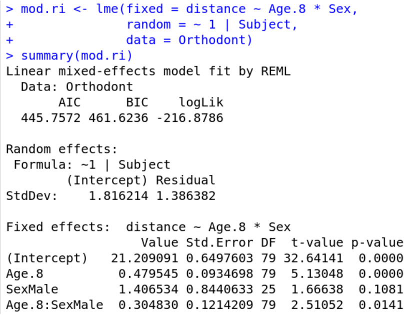
Interpretations can be the same as before:

- $\hat{\beta}_1$: The estimated rate of change in dental length per year for female children is 0.48 mm/year (95% CI: 0.29, 0.67 mm/year)
- $\hat{\beta}_1 + \hat{\beta}_3$: The estimated rate of change in dental length per year for male children is 0.78 mm/year (95% CI: 0.63, 0.94 mm/year)

Note: May be interpreted marginally or conditionally (mathematical property for linear models; more tomorrow about this)

Example: Dental Growth

R output:



```
> mod.ri <- lme(fixed = distance ~ Age.8 * Sex,
+               random = ~ 1 | Subject,
+               data = Orthodont)
> summary(mod.ri)
Linear mixed-effects model fit by REML
  Data: Orthodont
       AIC      BIC    logLik
  445.7572 461.6236 -216.8786

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.816214 1.386382

Fixed effects:  distance ~ Age.8 * Sex
                 Value Std.Error DF  t-value p-value
(Intercept)   21.209091 0.6497603 79 32.64141  0.0000
Age.8          0.479545 0.0934698 79  5.13048  0.0000
SexMale        1.406534 0.8440633 25  1.66638  0.1081
Age.8:SexMale  0.304830 0.1214209 79  2.51052  0.0141
```

Partitioning variability:

$$\widehat{\mathrm{Var}}(\epsilon) = \hat{\sigma}^2 = 1.39^2 = 1.93$$

$$\widehat{\mathrm{Var}}(b_i) = \hat{\tau}^2 = 1.82^2 = 3.31$$

$$\begin{aligned}\widehat{\mathrm{ICC}} &= \frac{\hat{\tau}^2}{\hat{\sigma}^2 + \hat{\tau}^2} \\ &= \frac{3.31}{1.93 + 3.31} \\ &= 0.632\end{aligned}$$

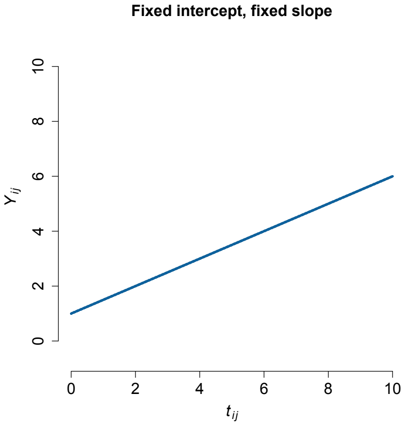
Between-subject variability is 63% of total variability

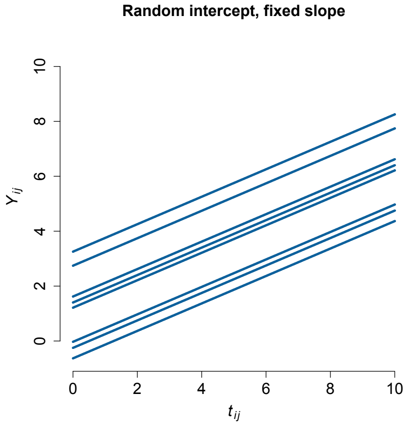
Example: Dental Growth

Observations:

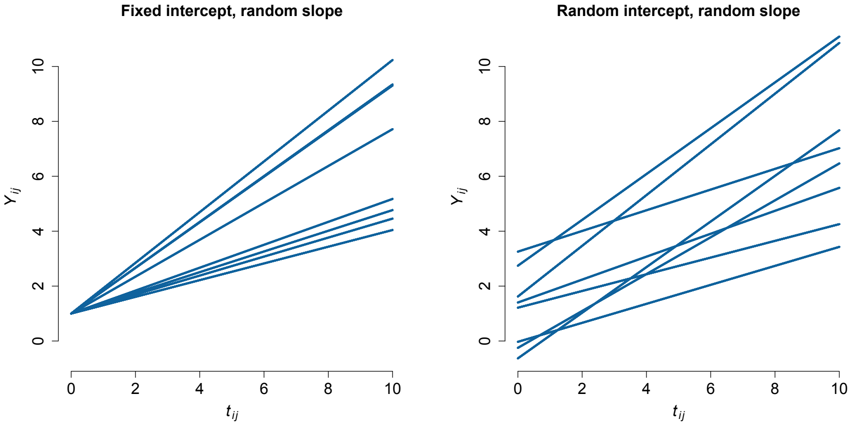
- **The two lines are parallel (slope can't vary by subject)**
  - What if we don't want them to be parallel?
- M11 has a shorter length at every time than M10
- Variability within subject (around their line) looks smaller than variability between subjects
  - Intraclass correlation coefficient will help quantify this

Choices For Random Effects


Fixed intercept, fixed slope


Random intercept, fixed slope

Choices For Random Effects



Linear Mixed Effects Model (General Framework)

(Laird and Ware, 1982)

The model includes the following components ($i = 1, ..., N$; $j = 1, ..., n_i$):

$Y_i = (Y_{i1}, ..., Y_{in_i})^\top$ Outcomes

$\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ Fixed effects

$\boldsymbol{X}_{ij} = (X_{ij1}, ..., X_{ijp})^\top$

$\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{in_i})$ Covariate matrix for fixed effects

$\boldsymbol{b}_i = (b_{i1}, ..., b_{iq})$ Random effects

$\boldsymbol{Z}_{ij} = (Z_{ij1}, ..., Z_{ijq})^\top$

$\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, ..., \boldsymbol{Z}_{in_i})$ Covariate matrix for random effects (typically a subset of X)

Linear Mixed Effects Model (General Framework)

1. Model for response given random effects:

$$Y_{ij} = \boldsymbol{X}_{ij}\beta + \boldsymbol{Z}_{ij}\boldsymbol{b}_i + \epsilon_{ij}$$

2. Model for random effects

$$\begin{aligned} \boldsymbol{b}_i &\sim N(0, \boldsymbol{D}) \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

with $\boldsymbol{b}_i$ and ϵ_{ij} assumed to be independent

Choices for Random Effects

Common linear mixed effects models include:

- **Random intercepts**

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{i0} + \epsilon_{ij} \\ &= (\beta_0 + b_{i0}) + \beta_1 t_{ij} + \epsilon_{ij} \end{aligned}$$

- **Random intercepts and slopes**

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij} \\ &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{ij} + \epsilon_{ij} \end{aligned}$$

Choices for Random Effects: $\boldsymbol{D}$

$\boldsymbol{D}$ quantifies random variation in trajectories across subjects

$$\boldsymbol{D} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$$

- $\sqrt{D_{11}}$ is the typical deviation in the level of the response
- $\sqrt{D_{22}}$ is the typical deviation in the change in the response
- D_{12} is the covariance between subject-specific intercepts and slopes
  - $D_{12} = 0$ indicates subject-specific intercepts and slopes are uncorrelated
  - $D_{12} > 0$ indicates subjects with high level have high rate of change
  - $D_{12} < 0$ indicates subjects with high level have low rate of change

$(D_{12} = D_{21})$

Induced Correlation Structure

Correlation induced by a random intercepts and slopes model:
(derivation is similar to before)

$$\text{Cov}(Y_{ij}, Y_{ik}) = D_{11} + (t_{ij} + t_{ik})D_{12} + t_{ij}t_{ik}D_{22}$$
$$\text{Var}(Y_{ij}) = D_{11} + t_{ij}^2 D_{22} + 2t_{ij}D_{12} + \sigma^2$$

Observations:

1. Allows heteroskedasticity across time as a function of t^2
2. Variance can possibly decrease over time if $\text{Cov}(b_{0i}, b_{1i}) < 0$, but will otherwise increase over time.
3. This is a special case of the general form:

   $$\boldsymbol{\Sigma}_i = \text{Cov}(Y_i) = Z_i \boldsymbol{D} Z_i^\top + \boldsymbol{R}_i$$

   where $\boldsymbol{R}_i$ is the covariance for the errors ϵ_i

Example: Dental Growth

Random intercept and slope model with interaction between sex and age:

$$\begin{aligned} \mathrm{E}[\mathrm{Dist}_{ij}|\mathrm{Age}_{ij}, \mathrm{Male}_i, b_i] = {} & \beta_0 + \beta_1 \times (\mathrm{Age}_{ij} - 8) + \beta_2 \times \mathrm{Male}_i \\ & + \beta_3 \times (\mathrm{Age}_{ij} - 8) \times \mathrm{Male}_i + b_{i0} + b_{i1} \times (\mathrm{Age}_{ij} - 8) \end{aligned}$$

Now the sex- and subject-specific models are:

- If child *i* is a girl:

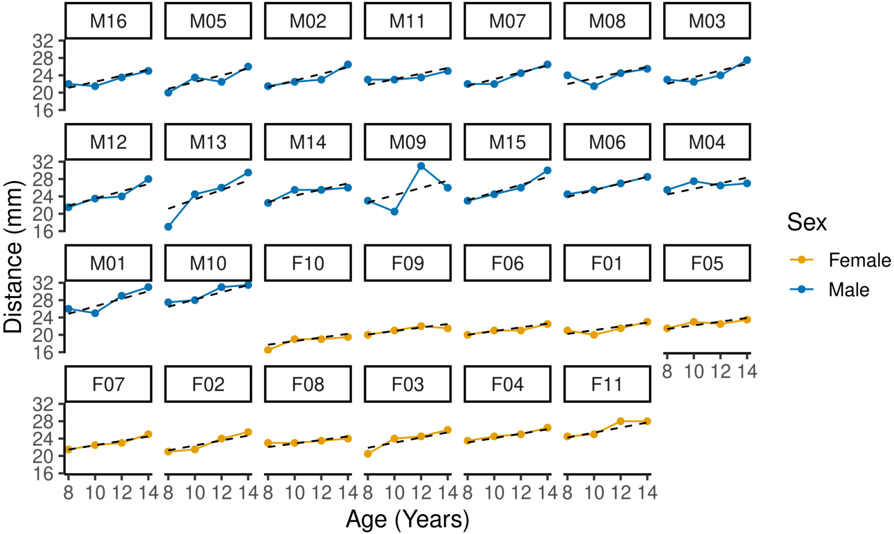
  $\mathrm{E}[\mathrm{Dist}_{ij}|\mathrm{Age}_{ij}, \mathrm{Male}_i = 0, b_i] = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) \times (\mathrm{Age}_{ij} - 8)$

- If child *i* is a boy:

$\mathrm{E}[\mathrm{Dist}_{ij}|\mathrm{Age}_{ij}, \mathrm{Male}_i = 1, b_i] = (\beta_0 + \beta_2 + b_{i0}) + (\beta_1 + \beta_3 + b_{i1}) \times (\mathrm{Age}_{ij} - 8)$

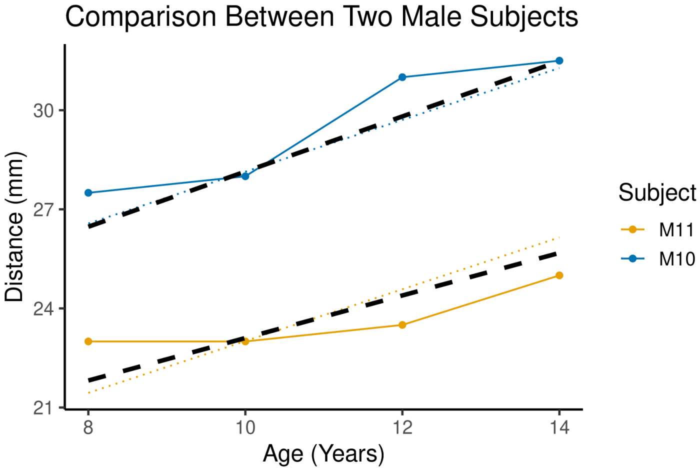
Example: Dental Growth

Taking a look at the results:



Example: Dental Growth

Zooming in on just two subjects again:



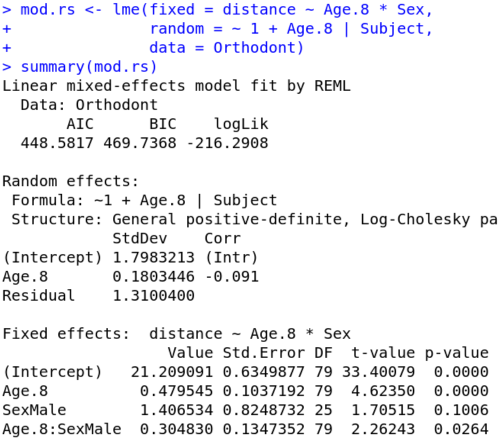
Example: Dental Growth

Observations:

- The two lines are *not* parallel (subject-specific slopes)
    - M10 has a slightly steeper slope, M11 has a less steep slope
- The differences are not large! (Do we really need random slopes?)
- Think: what do you expect to see in the covariance of the random effects?

Example: Dental Growth

R output:



```
> mod.rs <- lme(fixed = distance ~ Age.8 * Sex,
+               random = ~ 1 + Age.8 | Subject,
+               data = Orthodont)
> summary(mod.rs)
Linear mixed-effects model fit by REML
  Data: Orthodont
       AIC      BIC    logLik
  448.5817 469.7368 -216.2908

Random effects:
 Formula: ~1 + Age.8 | Subject
 Structure: General positive-definite, Log-Cholesky pa
            StdDev    Corr
(Intercept) 1.7983213 (Intr)
Age.8       0.1803446 -0.091
Residual    1.3100400

Fixed effects:  distance ~ Age.8 * Sex
                  Value Std.Error DF  t-value p-value
(Intercept)   21.209091 0.6349877 79 33.40079  0.0000
Age.8          0.479545 0.1037192 79  4.62350  0.0000
SexMale        1.406534 0.8248732 25  1.70515  0.1006
Age.8:SexMale  0.304830 0.1347352 79  2.26243  0.0264
```

Systematic part of model:

$$\widehat{\text{Dist}}_{ij} = 21.2 + 0.5 \times (\text{Age}_{ij} - 8) + 1.4 \times \text{M}_i + 0.3 \times (\text{Age}_{ij} - 8) \times \text{M}_i$$

(identical estimates!)

Random Effects Covariance:

$$\hat{\boldsymbol{D}} = \begin{bmatrix} 1.80^2 & -0.09 \\ -0.09 & 0.18^2 \end{bmatrix}$$

$$\hat{\sigma}^2 = 1.31^2$$

Observations:

- $\hat{D}_{22} << \hat{D}_{11}$
- $\hat{D}_{12} = -0.09$: kids who start with larger distances may grow more slowly

What's Beneath the Hood?

Maximum Likelihood:

$$\boldsymbol{Y}_i|\boldsymbol{b}_i \sim MVN(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \epsilon_i)$$
$$\boldsymbol{b}_i \sim MVN(0, \boldsymbol{D})$$

So the likelihood function is a product of normal densities,

$$P(\boldsymbol{Y}_i, \boldsymbol{b}_i) = P(\boldsymbol{Y}_i|\boldsymbol{b}_i)P(\boldsymbol{b}_i),$$

which is easy to integrate and maximize (continuous outcomes only!).

Restricted Maximum Likelihood:

- ML estimation of Σ is biased (e.g., n vs. $n - p$ in denominator for σ^2)
- REML (default) provides less-biased estimation of Σ
  - Details are beyond the scope of this module

Likelihood-based inference for β

- Consider testing fixed effects in nested linear mixed-effects models, e.g.,

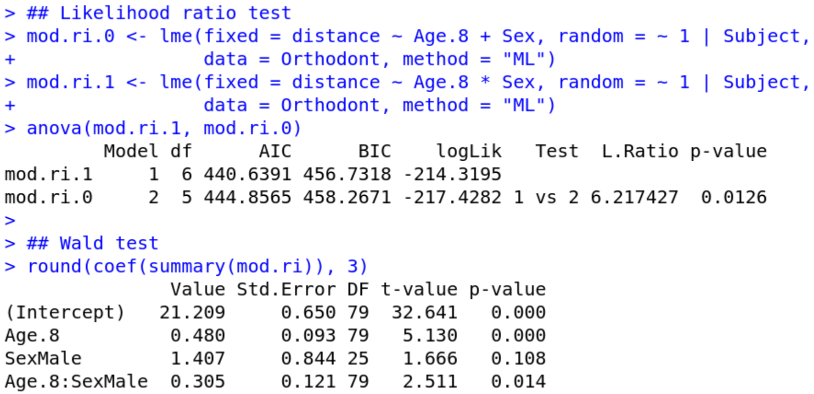
$$\text{Dist} = \beta_0 + \beta_1 \times (\text{Age}_{ij} - 8) + \beta_2 \times \text{M} \qquad \text{vs.}$$

$$\text{Dist} = \beta_0 + \beta_1 \times (\text{Age}_{ij} - 8) + \beta_2 \times \text{M} + \beta_3 \times (\text{Age}_{ij} - 8) \times \text{M}$$

  (equivalent to $H_0 : \beta_3 = 0$)

- Likelihood ratio test is valid with maximum likelihood estimation
  - Requires computation under the null and alternative hypotheses
- Likelihood ratio test may not be valid with other estimation methods (e.g., REML - R will warn you)
- Wald test (based on coefficient and standard error) is generally valid

Example: Dental Growth



```
> ## Likelihood ratio test
> mod.ri.0 <- lme(fixed = distance ~ Age.8 + Sex, random = ~ 1 | Subject,
+                 data = Orthodont, method = "ML")
> mod.ri.1 <- lme(fixed = distance ~ Age.8 * Sex, random = ~ 1 | Subject,
+                 data = Orthodont, method = "ML")
> anova(mod.ri.1, mod.ri.0)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.ri.1     1  6 440.6391 456.7318 -214.3195                        
mod.ri.0     2  5 444.8565 458.2671 -217.4282 1 vs 2 6.217427  0.0126
> 
> ## Wald test
> round(coef(summary(mod.ri)), 3)
              Value Std.Error DF t-value p-value
(Intercept)  21.209     0.650 79  32.641   0.000
Age.8         0.480     0.093 79   5.130   0.000
SexMale       1.407     0.844 25   1.666   0.108
Age.8:SexMale 0.305     0.121 79   2.511   0.014
```

Both tests agree: the association between age and distance differs by sex! (LR p=0.013, Wald p=0.014)

Choosing a Random Effects Structure

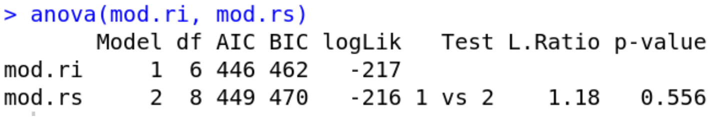
- Covariance model choice determines the standard error estimates for $\hat{\beta}$; correct model is required for correct standard error estimates
- Suppose we want to decide between two candidate random effect structures:

$$H_0 : \boldsymbol{D} = \left[\begin{array}{cc} D_{11} & 0 \\ 0 & 0 \end{array} \right] \quad \text{versus} \quad H_1 : \boldsymbol{D} = \left[\begin{array}{cc} D_{11} & D_{21} \\ D_{12} & D_{22} \end{array} \right],$$

- Could we formally test whether random intercepts are adequate, e.g., with LRT?

Likelihood-based inference for D

- Consider testing $H_0 : D_{12} = D_{22} = 0$
  - Fit both models with REML, compare with LRT
- This is possible in R



```
> anova(mod.ri, mod.rs)
       Model df AIC BIC logLik   Test L.Ratio p-value
mod.ri     1  6 446 462   -217
mod.rs     2  8 449 470   -216 1 vs 2    1.18   0.556
```

- Seems to suggest random intercept is sufficient (AIC is lower for RI than RS, p-value is big)
  - Consistent with our prior exploratory analysis

Choosing a Random Effects Structure

Problem: testing $D_{22} = 0$ is a nonstandard problem

- P-values tend to be conservative, could lead to over-simplifying correlation structure
- (Ad hoc fix proposed by Fitzmaurice, Laird, Ware: use $\alpha = 0.1$)

Alternatives:

- If only interested in inference on β:
  - Choose based on *a priori* scientific knowledge and exploratory analysis
  - Use robust standard errors (see module code; R package clubSandwich)
- If this is an exploratory analysis and you're interested in correlation structure, ok to test structures
  - Can cause type 1 error problems if your focus is inference on β

Dental Growth: Robust SE, Random Intercepts

Variable	$\hat{\beta}$	Estimation	P-value	CI
Intercept	21.2	Model-based	1.28×10^{-47}	(19.9, 22.5)
		Robust	6.35×10^{-12}	(19.9, 22.5)
(Age-8)	0.48	Model-based	2.02×10^{-6}	(0.29, 0.67)
		Robust	2.78×10^{-5}	(0.33, 0.63)
Male	1.41	Model-based	p=0.108	(-0.33, 3.14)
		Robust	p=0.095	(-0.27, 3.08)
(Age-8)*Male	0.30	Model-based	p=0.014	(0.06, 0.55)
		Robust	p=0.020	(0.05, 0.56)

Choosing a Random Effects Structure

A priori considerations:

- Random intercepts only:
  - Simpler (+/-) and easy to interpret (e.g., ICC)
  - Induces exchangeable correlation (+/-)
- Random intercepts and slopes:
  - More information about individual trajectories and covariance (+)
  - More complex (+/-)
    - Over-parameterization of covariance $\Longrightarrow$ inefficient estimation of fixed-effects parameters β

Dental Growth: Did We Answer Our Questions?

1. Estimate average growth curve for all children
    - Fixed effects coefficients
2. Estimate growth curves for individual children
    - Random intercept and slope model gives estimate for each child
3. Characterize heterogeneity in children's growth rates
    - Variability in random intercepts is quite high (distance at age 8)
    - Variability in random slopes is quite small (growth rates are fairly homogeneous)
4. Assess whether the growth rate differs between boys and girls
    - It does! Fixed effect interaction term is significantly nonzero
    - (and scientifically interesting - growth rate 0.8 mm/yr for boys, 0.5 mm/yr for girls)

Assumptions

Valid inference from a linear mixed-effects model relies on

- **Mean model**: As with any regression model for an average outcome, need to correctly specify the functional form of $\boldsymbol{X}_{ij}\boldsymbol{\beta}$ (here also $\boldsymbol{Z}_{ij}\boldsymbol{b}_i$)
  - Included important covariates in the model
  - Correctly specified any transformations or interactions
- **Covariance model**: Correct covariance model (random-effects specification) is required for correct standard error estimates for $\hat{\beta}$
  - Or robust SE
- **Normality**: Normality of ϵ_{ij} and $\boldsymbol{b}_i$ is required for normal likelihood function to be the correct likelihood function for Y_{ij}
  - Especially important for small samples & trusting individual trajectories
- N sufficiently large for **asymptotic inference** to be valid

Summary

- Mixed-effects models combine population-average (systematic) model components with subject-specific (random effects) components
- Estimation and inference for:
  - Average level or trajectory
  - Between-subject heterogeneity in level or trajectory
- Subject-specific random effects induce a correlation structure (conceptually nice, easy even for unbalanced data)
- Parametric approach; ML estimation is valid (but... assumptions)
- Could have multiple levels of random effects (e.g., clustering and longitudinal)

Issues

- Requires that any missing data are missing at random

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

- Recap
- Linear Mixed Model
- Wrap-Up: Methods for Continuous Outcomes
- Activity
- Review of Generalized Linear Models
- Generalized Estimating Equations

Day 3: Wednesday, July 12

Assumptions: GEE

Valid inference from a general linear model relies on

- **Mean model**: As with any regression model for an average outcome, need to correctly specify the functional form of $\mathbf{X}_{ij}\boldsymbol{\beta}$
    - Included important covariates in the model
    - Correctly specified any transformations or interactions
- **Covariance model**: Correct covariance model is required for correct standard error estimates for $\hat{\beta}$ if using **model-based variance estimate** otherwise we can use **robust/empirical variance estimate**
- N sufficiently large

Summary: GEE

- Primary focus is on the mean model
- Longitudinal correlation is secondary to mean model of interest and is treated as a nuisance
- Requires selection of a ‘working‘ correlation model
- Semi-parametric model: mean + correlation
- Working correlation model does not need to be correctly specified to obtain consistent estimator for $\boldsymbol{\beta}$ or valid standard errors for $\hat{\boldsymbol{\beta}}$ but efficiency gains possible if correlation model is correct
- Wald testing

Issues:

- Accommodates only one source of correlation: longitudinal or cluster
- Requires that any missing data are missing completely at random
- Issues arise with time-dependent exposures and covariance weighting

Assumptions: LMM

Valid inference from a linear mixed-effects model relies on

- **Mean model**: As with any regression model for an average outcome, need to correctly specify the functional form of $\boldsymbol{X}_{ij}\beta$ (here also $\boldsymbol{Z}_{ij}\boldsymbol{b}_i$)
  - Included important covariates in the model
  - Correctly specified any transformations or interactions
- **Covariance model**: Correct covariance model (random-effects specification) is required for correct standard error estimates for $\hat{\beta}$
  - Or robust SE
- **Normality**: Normality of ϵ_{ij} and $\boldsymbol{b}_i$ is required for normal likelihood function to be the correct likelihood function for Y_{ij}
  - Especially important for small samples & trusting individual trajectories
- N sufficiently large for **asymptotic inference** to be valid

Summary: LMM

- Mixed-effects models combine population-average (systematic) model components with subject-specific (random effects) components
- Estimation and inference for:
  - Average level or trajectory
  - Between-subject heterogeneity in level or trajectory
- Subject-specific random effects induce a correlation structure (conceptually nice, easy even for unbalanced data)
- Parametric approach; ML estimation is valid (but... assumptions)
- Could have multiple levels of random effects (e.g., clustering and longitudinal)

Issues

- Requires that any missing data are missing at random

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

- Recap
- Linear Mixed Model
- Wrap-Up: Methods for Continuous Outcomes
- Activity
- Review of Generalized Linear Models
- Generalized Estimating Equations

Day 3: Wednesday, July 12

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

- Recap
- Linear Mixed Model
- Wrap-Up: Methods for Continuous Outcomes
- Activity
- Review of Generalized Linear Models
- Generalized Estimating Equations

Day 3: Wednesday, July 12

Review of Generalized Linear Models

Generalized linear models (GLMs): class of models for regression analysis of observations that includes **linear regression models** for continuous responses, but also others:

- **Logistic regression** for binary response (e.g. yes/no or 0/1)
- **Log-linear** or **Poisson regression** for counts
- Others. For example ordinal models

Review of Generalized Linear Models

- Assume the outcomes are independent of each other
- Suppose we have N independent observations of a response variable Y
- Let Y_i denote the response variable for the ith subject
- $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ where X_{ik} denotes the kth covariate for the ith subject

Two part specification:

1. Random component (usually a distributional assumption)
2. Systematic component (how the mean relates to covariates)

Review of GLMs: Random Component

Assume we know the distribution of the outcome,

$$Y_i|\mu_i \sim \text{exponential family distribution}$$

- Linear regression: $Y_i \sim N(\mu_i, \sigma^2)$
- Logistic regression: $Y_i \sim \text{Bernoulli}(\mu_i)$
- Poisson regression: $Y_i \sim \text{Poisson}(\mu_i)$

Distribution	Mean	Variance Function
Normal	μ_i	σ^2
Bernoulli	μ_i	$\mu_i(1 - \mu_i)$
Poisson	μ_i	μ_i

Review of GLMs: Systematic Component

Assume the transformed mean of Y_i given $\mathbf{X}_i$ ($\mu_i = E[Y_i|\mathbf{X}_i]$) are related to the covariates in the following manner:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} = \mathbf{X}_i \boldsymbol{\beta}$$

The **link function** $g(\cdot)$ describes the relationship between the linear predictor ($\mathbf{X}_i \boldsymbol{\beta}$) and the expected value of Y_i (i.e., μ_i)

Distribution	Typical link function $g(\cdot)$
Normal	Identity: $g(\mu_i) = \mu_i$
Bernoulli	Logit: $g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
Poisson	Log: $g(\mu_i) = \log(\mu_i)$

Binary Outcome: Example

Clinical Trial of Contracepting Women:

- 1151 women randomized to either 100mg or 150mg of DMPA
- **Outcome of interest**: Amenorrhea during each 3-month interval after an injection
  - Since participants were measured multiple times we would want to use correlated data techniques to analyze all responses
  - We will focus on the response at the 4th (last scheduled) visit
- Is there a difference in the odds of amenorrhea after 1 year of injections by dose?

Data:

- 100mg: 50.1% (181 out of 361) experienced amenorrhea at 1 year
- 150mg: 53.5% (189 out of 353) experienced amenorrhea at 1 year

Binary Outcome: Example

We will use logistic regression:

- **Random component**: $Y_i \sim \text{Bernoulli}(\mu_i) \rightarrow \text{Var}(Y_i) = \mu_i(1-\mu_i)$
- **Systematic component**:

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 \text{Dose}_i$$

where $\mu_i = \Pr(Y_i = 1)$ where Y_i is an indicator for whether person i experienced amenorrhea at 1 year ($1 =$ yes; $0 =$ no)

In R:

```
glm(amenorrhea ~ dose, family = binomial(link = "logit"),
data = amenorrhea[amenorrhea$visit==4,])
```

Binary Outcome: Example

```
glm(formula = amenorrhea ~ dose, family = binomial(link = "logit"),
data = amenorrhea[amenorrhea$visit == 4, ])

Deviance Residuals:
Min       1Q  Median      3Q     Max
-1.24   -1.18    1.12    1.18    1.18

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.00554    0.10526    0.05      0.96
dose         0.13634    0.14990    0.91      0.36

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 988.87  on 713  degrees of freedom
Residual deviance: 988.04  on 712  degrees of freedom
(437 observations deleted due to missingness)
AIC: 992

Number of Fisher Scoring iterations: 3
```

Binary Outcome: Example

- Estimate for `(Intercept)` = estimate of the log odds of amenorrhea for 100mg dose
- Estimate for `dose` = estimate of the log odds ratio of amenorrhea comparing 150mg dose to 100mg dose

Interpretation: the odds of amenorrhea is estimated to be 15% $(= \exp(0.136) - 1)$ higher for those on the 150mg dose as compared to those on the 100mg dose. We do not have evidence that amenorrhea after 1 year differs by dose $(p = 0.36)$.

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

- Recap
- Linear Mixed Model
- Wrap-Up: Methods for Continuous Outcomes
- Activity
- Review of Generalized Linear Models
- Generalized Estimating Equations

Day 3: Wednesday, July 12

GEE

⋆ Contrast average outcome values across populations of individuals defined by covariate values, while accounting for correlation

- Focus on a generalized linear model with regression parameters $\boldsymbol{\beta}$, which characterize the systemic variation in Y across covariates X

$$\begin{array}{rcll} \mathbf{Y}_i & = & \{Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\}^{\mathsf{T}} & \text{Outcomes} \\ \mathbf{X}_{ij} & = & \{1, X_{ij1}, X_{ij2}, \ldots, X_{ijp}\} & \text{Covariates} \\ \mathbf{X}_i & = & \{\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{in_i}\}^{\mathsf{T}} & \text{Design matrix} \\ \boldsymbol{\beta} & = & \{\beta_0, \beta_1, \beta_2, \ldots, \beta_p\}^{\mathsf{T}} & \text{Regression parameters} \end{array}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$

- Longitudinal correlation structure is a nuisance feature of the data

(Liang and Zeger, 1986)

Mean model

Assumptions

- Observations are independent across subjects
- Observations may be correlated within subjects

Mean model: Primary focus of the analysis

$$
\begin{aligned}
\mathrm{E}[Y_{ij} \mid \mathbf{X}_{ij}] &= \mu_{ij} \\
g(\mu_{ij}) &= \mathbf{X}_{ij}\boldsymbol{\beta}
\end{aligned}
$$

- May correspond to any generalized linear model with link $g(\cdot)$

Continuous outcome	Count outcome	Binary outcome
$\mathrm{E}[Y_{ij} \mid \mathbf{X}_{ij}] = \mu_{ij}$	$\mathrm{E}[Y_{ij} \mid \mathbf{X}_{ij}] = \mu_{ij}$	$\mathrm{P}[Y_{ij} = 1 \mid \mathbf{X}_{ij}] = \mu_{ij}$
$\mu_{ij} = \mathbf{X}_{ij}\beta$	$\log(\mu_{ij}) = \mathbf{X}_{ij}\beta$	$\mathrm{logit}(\mu_{ij}) = \mathbf{X}_{ij}\beta$

- Characterizes a marginal mean regression model
  - μ_{ij} does not condition on anything other than $\mathbf{X}_{ij}$

Covariance model

Longitudinal correlation is a nuisance; secondary to mean model of interest

1. Assume a form for variance that may depend on μ_{ij}

$$\begin{array}{lrcl} \text{Continuous outcome:} & \text{Var}[Y_{ij} \mid \mathbf{X}_{ij}] & = & \sigma^2 \\ \text{Count outcome:} & \text{Var}[Y_{ij} \mid \mathbf{X}_{ij}] & = & \mu_{ij} \\ \text{Binary outcome:} & \text{Var}[Y_{ij} \mid \mathbf{X}_{ij}] & = & \mu_{ij}(1-\mu_{ij}) \end{array}$$

   which may also include a scale or dispersion parameter $\phi > 0$

2. Select a model for longitudinal correlation with parameters α

$$\begin{array}{lrcl} \text{Independence:} & \text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] & = & 0 \\ \text{Exchangeable:} & \text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] & = & \alpha \\ \text{Auto-regressive:} & \text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] & = & \alpha^{|j-j'|} \\ \text{Unstructured:} & \text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] & = & \alpha_{jj'} \end{array}$$

Covariance model

Longitudinal correlation is a nuisance; secondary to mean model of interest

- Assume a form for variance that depends on μ
- Select a model for longitudinal correlation with parameters α

$$\mathrm{Var}[Y_{ij} \mid \mathbf{X}_{ij}] = V(\mu_{ij}) \qquad \rightarrow \qquad \mathbf{S}_i(\mu_i) = \mathrm{diag}\ V(\mu_{ij})$$

$$\mathrm{Corr}[Y_{ij},\, Y_{ij'} \mid \mathbf{X}_i] = \rho_{ijj'}(\boldsymbol{\alpha}) \qquad \rightarrow \qquad \mathbf{R}_i(\boldsymbol{\alpha}) = \text{matrix } \rho(\boldsymbol{\alpha})$$

$$\mathrm{Cov}[\mathbf{Y}_i \mid \mathbf{X}_i] = \mathbf{V}_i(\boldsymbol{\beta},\, \boldsymbol{\alpha}) = \mathbf{S}_i^{1/2}\mathbf{R}_i\mathbf{S}_i^{1/2}$$

Correlation models

Independence: $\text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] = 0$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Exchangeable: $\text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] = \alpha$

$$\begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix}$$

Correlation models

Auto-regressive: $\text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] = \alpha^{|j-j'|}$

$$\begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{m-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{m-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{m-1} & \alpha^{m-2} & \alpha^{m-3} & \cdots & 1 \end{bmatrix}$$

Unstructured: $\text{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i] = \alpha_{jj'}$

$$\begin{bmatrix} 1 & \alpha_{21} & \alpha_{31} & \cdots & \alpha_{m1} \\ \alpha_{12} & 1 & \alpha_{32} & \cdots & \alpha_{m2} \\ \alpha_{13} & \alpha_{23} & 1 & \cdots & \alpha_{m3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{1m} & \alpha_{2m} & \alpha_{3m} & \cdots & 1 \end{bmatrix}$$

Correlation models

Correlation between any two observations on the same subject. . .

- **Independence**: . . . is assumed to be zero
  - Appropriate with use of robust variance estimator (large N)
- **Exchangeable**: . . . is assumed to be constant
  - More appropriate for clustered data
- **Auto-regressive**: . . . is assumed to depend on time or distance
  - More appropriate for equally-spaced longitudinal data
- **Unstructured**: . . . is assumed to be distinct for each pair
  - Only appropriate for short series (small n) on many subjects (large N)

Semi-parametric

- Specification of a mean model and correlation model does not identify a complete probability model for the outcomes
- The [mean, correlation] model is semi-parametric because it only specifies the first two moments of the outcomes

Question: Without a likelihood function, how do we estimate $\boldsymbol{\beta}$ and generate valid statistical inference, while accounting for correlation?

Answer: Construct an unbiased estimating function

Estimating functions

The estimating function for estimation of β is given by

$$
\begin{aligned}
\mathcal{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \mathbf{D}_i^\top(\boldsymbol{\beta})\mathbf{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})[\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] \\
g(\boldsymbol{\mu}_i) &= \mathbf{X}_i\boldsymbol{\beta} \\
\mathbf{D}_i(\boldsymbol{\beta}) &= \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \\
\mathbf{D}_i(j, k) &= \frac{\partial \mu_{ij}}{\partial \beta_k}
\end{aligned}
$$

- $\mathbf{V}_i$ is the 'working' variance-covariance matrix: $\mathrm{Cov}[\mathbf{Y}_i \mid \mathbf{X}_i]$
  - Depends on the assumed form for the variance: $\mathrm{Var}[Y_{ij} \mid \mathbf{X}_{ij}]$
  - Depends on the specified correlation model: $\mathrm{Corr}[Y_{ij}, Y_{ij'} \mid \mathbf{X}_i]$
- $\mathcal{U}(\boldsymbol{\beta})$ depends on the model or value for $\boldsymbol{\alpha}$

Generalized estimating equations

Setting an **estimating function** equal to 0 defines an **estimating equation**

$$
\begin{aligned}
\mathbf{0} &= \mathcal{U}(\hat{\boldsymbol{\beta}}) \\
&= \sum_{i=1}^{N} \mathbf{D}_i^{\mathsf{T}}(\hat{\boldsymbol{\beta}})\mathbf{V}_i^{-1}(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha})[\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})]
\end{aligned}
$$

- ‘Generalized’ because it corresponds to a GLM with link function $g(\cdot)$
- Solution to the estimation equation defines an estimator $\hat{\boldsymbol{\beta}}$
- Note $\mathcal{U}(\hat{\boldsymbol{\beta}})$ depends on the model or value for $\boldsymbol{\alpha}$

Generalized estimating equations: Intuition

$$\mathbf{0} = \sum_{i=1}^{N} \underbrace{\mathbf{D}_i^\top(\hat{\boldsymbol{\beta}})}_{\boxed{3}} \underbrace{\mathbf{V}_i^{-1}(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha})}_{\boxed{2}} \underbrace{[\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})]}_{\boxed{1}}$$

$\boxed{1}$ The model for the mean $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, is compared to the observed data $\mathbf{Y}_i$. Setting the functions equal to $\mathbf{0}$ tries to minimize the difference between the **observed** and **expected**

$\boxed{2}$ Estimation uses the inverse of the variance (covariance) to weight the data from subject i; more weight is given to differences between observed and expected for those subjects who contribute more information

$\boxed{3}$ This is simply a 'change of scale' from the scale of the mean, $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, to the scale of the regression coefficients (covariates)

Because the GEE depends on both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, an iterative procedure is used

GEE: Properties of $\hat{\boldsymbol{\beta}}$

Question: What are the properties of $\hat{\boldsymbol{\beta}}$, the regression estimate?

Answer:

- The regression coefficient estimate will be correct (in large samples) even if you choose the wrong dependence model
- However, the variance of the regression estimate must capture the correlation in the data, either through choosing the correct covariance model, or using an **alternative variance estimate**
- Correctly specified (or close) covariance model will yield regression estimate that is most (or more) efficient (smaller variance)

GEE: Sandwich variance estimator

The empirical estimator is:

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \underbrace{\left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i^\top \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1}}_{\text{bread}} \underbrace{\left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i^\top \hat{\mathbf{V}}_i^{-1} \widetilde{\mathrm{Cov}}(\mathbf{Y}_i) \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)}_{\text{cheese}} \underbrace{\left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i^\top \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}\right)^{-1}}_{\text{bread}}$$

where we can use the residuals for $\widetilde{\mathrm{Cov}}(\mathbf{Y_i})$, i.e., has the entries $e_{ij}e_{ik}$ where $e_{ij} = Y_{ij} - \hat{\mu}_{ij}$.

Note that with linear regression (Gaussian family with identity link), $\mathbf{D}_i = \mathbf{X}_i$ and $\mathbf{V}_i = \phi\mathbf{R}_i$ where $\mathbf{R}_i$ is our working correlation matrix and $\phi = \sigma^2$

GEE: Sandwich variance estimator

$$\widehat{\text{Cov}}(\hat{\beta}) = \underbrace{\left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i^\top \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1}}_{\text{bread}} \underbrace{\left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i^\top \hat{\mathbf{V}}_i^{-1} \widetilde{\text{Cov}}(\mathbf{Y}_i) \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)}_{\text{cheese}} \underbrace{\left(\sum_{i=1}^{N} \hat{\mathbf{D}}_i^\top \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}\right)^{-1}}_{\text{bread}}$$

- Also known as sandwich, robust, or Huber-White variance estimator
- Requires large enough sample size ($N \geq 40$)
- Requires large enough sample size relative to cluster size ($N \gg n$)
- **Gives valid standard errors for the estimated regression coefficients even if the correlation model is wrong!**

GEE: Inference

Can perform Wald test and construct a Wald confidence interval:

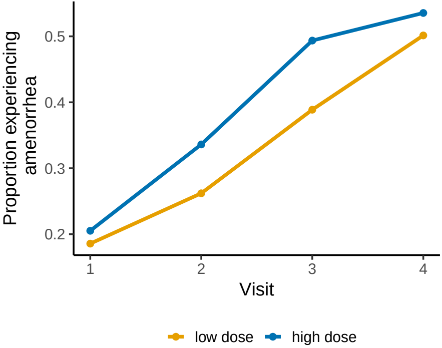
- $\hat{\beta}_k$/s.e. - valid test
- $\hat{\beta}_k \pm 1.96 \times$ s.e. - valid 95% confidence interval

Cannot perform a likelihood ratio test (no fully specified probability model) so no AIC or BIC either.

Case Study: Clinical Trial of Contracepting Women

Longitudinal clinical trial where people who menstruate received an injection of either 100 mg or 150 mg of DMPA at randomization and then every 90 days. Final follow-up visit after the 4th injection, 1 year after the randomization.

- $N = 1151$ people completed menstrual diaries
- Response: whether the person experienced amenorrhea in the previous 3 months
- Substantial dropout: more than 1/3 dropped out before completing the trial



Case Study: Clinical Trial of Contracepting Women

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 \mathrm{t}_{ij} + \beta_2 t_{ij}^2 + \beta_3 \mathrm{Dose}_i \times t_{ij} + \beta_4 \mathrm{Dose}_i \times t_{ij}^2$$

where $\mu_{ij} = \Pr(Y_{ij} = 1)$

- Y_{ij}: indicator for whether person i experienced amenorrhea in the jth injection interval (1 = yes; 0 = no)
- t_{ij}: measurement occasion (corresponds to the four consecutive 90-day injection intervals)
- Dose_i: indicator for whether the person i was randomized to 150 mg of DMPA or 100 mg (1 = 150 mg; 0 = 100 mg)

Case Study: Clinical Trial of Contracepting Women

```
m_ind <- geeglm(amenorrhea ~ visit + visit:dose + I(visit^2) + I(visit^2):dose,
id = id, data = amenorrhea, family = binomial, waves = visit,
corstr = "independence")
m_exc <- geeglm(amenorrhea ~ visit + visit:dose + I(visit^2) + I(visit^2):dose,
id = id, data = amenorrhea, family = binomial, waves = visit,
corstr = "exchangeable")
m_ar1 <- geeglm(amenorrhea ~ visit + visit:dose + visit2 + visit2:dose,
id = id, data = amenorrhea, family = binomial, waves = visit,
corstr = "ar1")
m_uns <- geeglm(amenorrhea ~ visit + visit:dose + I(visit^2) + I(visit^2):dose,
id = id, data = amenorrhea, family = binomial, waves = visit,
corstr = "unstructured")

m_glm <- glm(amenorrhea ~ visit + visit:dose + I(visit^2) + I(visit^2):dose,
data = amenorrhea, family = binomial)
```

Case Study: Clinical Trial of Contracepting Women

```
geeglm(formula = amenorrhea ~ visit + visit:dose + I(visit^2) +
I(visit^2):dose, family = binomial, data = amenorrhea, id = id,
waves = visit, corstr = "independence")

Coefficients:
Estimate Std.err   Wald Pr(>|W|)
(Intercept)      -2.1955  0.1784 151.37  < 2e-16 ***
visit             0.6698  0.1622  17.04  3.7e-05 ***
I(visit^2)       -0.0303  0.0328   0.86   0.3549
visit:dose        0.2973  0.1135   6.87   0.0088 **
dose:I(visit^2)  -0.0624  0.0296   4.44   0.0351 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

Estimate Std.err
(Intercept)        1  0.0289
Number of clusters:   1151  Maximum cluster size: 4
```

Case Study: Clinical Trial of Contracepting Women

```
geeglm(formula = amenorrhea ~ visit + visit:dose + I(visit^2) +
I(visit^2):dose, family = binomial, data = amenorrhea, id = id,
waves = visit, corstr = "exchangeable")

Coefficients:
Estimate Std.err   Wald Pr(>|W|)
(Intercept)      -2.2370  0.1765 160.64  < 2e-16 ***
visit             0.6967  0.1586  19.31  1.1e-05 ***
I(visit^2)       -0.0328  0.0320   1.05   0.3055
visit:dose        0.3284  0.1100   8.91   0.0028 **
dose:I(visit^2)  -0.0637  0.0286   4.97   0.0259 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

Estimate Std.err
(Intercept)        1  0.0287
Link = identity

Estimated Correlation Parameters:
Estimate Std.err
alpha    0.363  0.0243
Number of clusters:   1151  Maximum cluster size: 4
```

Case Study: Clinical Trial of Contracepting Women

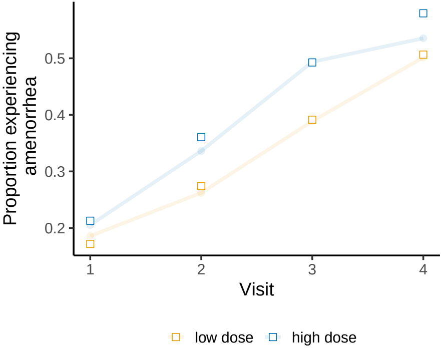
	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)	$\hat{\beta}_4$ (SE)
Ind	-2.2 (0.18)	0.67 (0.16)	-0.03 (0.03)	0.30 (0.11)	-0.06 (0.03)
Exch	-2.2 (0.18)	0.70 (0.16)	-0.03 (0.03)	0.33 (0.11)	-0.06 (0.03)
AR1	-2.2 (0.18)	0.71 (0.16)	-0.03 (0.03)	0.36 (0.11)	-0.08 (0.03)
Unst	-2.2 (0.18)	0.70 (0.16)	-0.03 (0.03)	0.34 (0.11)	-0.07 (0.03)
GLM	-2.2 (0.21)	0.67 (0.19)	-0.03 (0.04)	0.30 (0.11)	-0.06 (0.03)

- Working independence and using `glm()` give exactly the same point estimates
- `glm()` standard errors are not correct (here, too large)
- Working independence, exchangeable, AR1, and unstructured provide similar results

Case Study: Clinical Trial of Contracepting Women

Using working exchangeable:

- The ratio of the population odds of amenorrhea at 12 months comparing high dose to low dose is estimated to be 1.34 (95% CI: 1.01 - 1.78)
- Conducting a multivariate Wald test ($H_0 : \beta_3 = \beta_4 = 0$) we find a statistically significant effect of dose, p-value $= 0.002$



GEE: Which Correlation Model to Choose?

Question: Which correlation model should I choose?

Answer: Ideally, to preserve CI statements, Type I error, the choice of working correlation should be based on external information or substantive grounds rather than exploratory analysis

Question: If the correlation model does not need to be correctly specified to obtain a consistent estimator for β or valid standard errors for $\hat{\beta}$, why not always use an independence working correlation model?

Answer: Selecting a non-independence or weighted correlation model

- Permits use of the model-based variance estimator
- May provide improved efficiency for $\hat{\beta}$

GEE: Summary

- Primary focus of the analysis is a marginal mean regression model that corresponds to any GLM
- Longitudinal correlation is secondary to the mean model of interest and is treated as a nuisance feature of the data
- Requires selection of a 'working' correlation model
- Lack of a likelihood function implies that likelihood ratio test statistics are unavailable; hypothesis testing with GEE uses Wald statistics
- Working correlation model does not need to be correctly specified to obtain a consistent estimator for β or valid standard errors for $\hat{\beta}$, but efficiency gains are possible if the correlation model is correct

Issues

- Accommodates only one source of correlation: Longitudinal **or** cluster
- GEE requires that any missing data are missing completely at random
- Issues arise with time-dependent exposures and covariance weighting

End of Day 2

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

- Recap
- Generalized Linear Mixed Models
- Wrap-Up: Methods for Binary and Count Outcomes
- Activity
- Advanced topics
- Summary

Recap: GLMs for Independent Responses

- Suppose we have N independent observations of a response variable Y
- Let Y_i denote the response variable for the ith subject
- $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ where X_{ik} denotes the kth covariate for the ith subject

Assume the transformed mean of Y_i given $\mathbf{X}_i$ ($\mu_i = E[Y_i|\mathbf{X}_i]$) are related to the covariates in the following manner:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} = \mathbf{X}_i \boldsymbol{\beta}$$

Distribution	Mean	Variance	Typical link function $g(\cdot)$
Normal	μ_i	σ^2	Identity: $g(\mu_i) = \mu_i$
Bernoulli	μ_i	$\mu_i(1-\mu_i)$	Logit: $g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
Poisson	μ_i	μ_i	Log: $g(\mu_i) = \log(\mu_i)$

Recap: GLMs Interpretations

- Linear regression: β_k: difference in means between groups differing in X_k by 1 unit (holding everything else constant)

$$\mu_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_p X_p$$

- Logistic regression: e^{β_k}: odds ratio between groups differing in X_k by 1 unit (holding everything else constant)

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \cdots + \beta_p X_p$$

Recap: GEE Approach

⋆ Contrast average outcome values across populations of individuals defined by covariate values, while accounting for correlation

- Specify the **marginal** mean model, the primary focus of the analysis

$$\begin{aligned} \mathrm{E}[Y_{ij} \mid \mathbf{X}_{ij}] &= \mu_{ij} \\ g(\mu_{ij}) &= \mathbf{X}_{ij}\boldsymbol{\beta} \end{aligned}$$

- Choose a working covariance model (form for variance could depend on mean)
  - Longitudinal correlation is a nuisance feature
  - Provides a way to weight the data
- Estimation proceeds via estimating equations (no likelihood)
- Even if wrong covariance model assumed, regression coefficient estimate valid, valid SEs can be obtained via the sandwich (large samples)

Recap: Linear Mixed Model (LMM)

- In contrast to GEE, the source of correlation is of interest in LMMs
  - Assume it is due to between-subject heterogeneity

- Model specification:
  1. Model for response given random effects:

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_i + \epsilon_{ij}$$

  2. Model for random effects:

$$\begin{aligned} \boldsymbol{b}_i &\sim N(0, \boldsymbol{D}) \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

  with $\boldsymbol{b}_i$ and ϵ_{ij} assumed to be independent

Common LMM Specifications

Common linear mixed effects models include:

- **Random intercepts**

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{i0} + \epsilon_{ij} \\ &= (\beta_0 + b_{i0}) + \beta_1 t_{ij} + \epsilon_{ij} \end{aligned}$$

- **Random intercepts and slopes**

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij} \\ &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{ij} + \epsilon_{ij} \end{aligned}$$

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

- Recap
- Generalized Linear Mixed Models
- Wrap-Up: Methods for Binary and Count Outcomes
- Activity
- Advanced topics
- Summary

Non-Continuous Outcomes: What Changes?

- We've seen specification of GLMs:

$$\begin{aligned}&\boldsymbol{Y}_i|\mu_i \sim \text{exponential family distribution}\\ &g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}\end{aligned}$$

- We'll focus on binary outcomes, but distribution and link function may be swapped out for other outcome types

$$\boldsymbol{Y}_i|\mu_i \sim \text{Bernoulli}(\mu_i)$$
$$\text{logit}(\mu_i) = \boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$$

- Can't we just add $\boldsymbol{Z}_i\boldsymbol{b}_i$ to $\boldsymbol{\eta}_i$, and update our likelihood function?
    - ... yes and no.

Generalized Linear Mixed Effects Models

LMM:

$$\begin{aligned} Y_i|\mu_i, \boldsymbol{b}_i &\sim \text{Normal}(\mu_i, \sigma^2 \boldsymbol{I}) \\ \mu_i|\boldsymbol{b}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim \text{Normal}(0, \boldsymbol{D}) \end{aligned}$$

GLMM:

$$\begin{aligned} Y_i|\mu_i, \boldsymbol{b}_i &\sim \text{Bernoulli}(\mu_i) \\ \text{logit}(\mu_i|\boldsymbol{b}_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim \text{Normal}(0, \boldsymbol{D}) \end{aligned}$$

Generalized Linear Mixed Effects Models

- Conceptually very similar!

- Computationally, GLMM is much more complicated
  - Likelihood: product of non-Normal exponential density (for $\boldsymbol{Y}_i|\boldsymbol{b}_i$) and Normal density (for $\boldsymbol{b}_i$)
  - Typically fit using approximation or numerical techniques
  - (NB: centering and scaling predictors can help with convergence issues)

- Interpretation is (importantly) different

Conditional and Marginal Effects

- Parameter estimates obtained from a marginal model (as obtained via GEE) estimate population-averaged contrasts

- Parameter estimates obtained from a conditional model (as obtained via GLMM) estimate subject-specific contrasts

- In a linear model for a Gaussian outcome with an identity link, these are equivalent; not the case with non-linear models
    - Depends on the outcome distribution
    - Depends on the specified random effects

Conditional = Marginal for LMM, Not GLMM. Why?

For an LMM, we use the identity link function. Then:

- Taking the expected value over $\boldsymbol{b}_i$:

$$\mathrm{E}_b\left\{\mathrm{E}(Y_{ij}|X_{ij},\boldsymbol{b}_i)\right\} = \mathrm{E}_b\left\{X_{ij}^\top\boldsymbol{\beta} + Z_{ij}^\top\boldsymbol{b}_i\right\} = Z_{ij}^\top\boldsymbol{\beta} = \mathrm{E}(Y_{ij}|X_{ij})$$

- So the **average conditional mean** $\mathrm{E}_b\left\{\mathrm{E}(Y_{ij}|X_{ij},\boldsymbol{b}_i)\right\}$ is the same as the **marginal mean** $\mathrm{E}(Y_{ij}|X_{ij})$

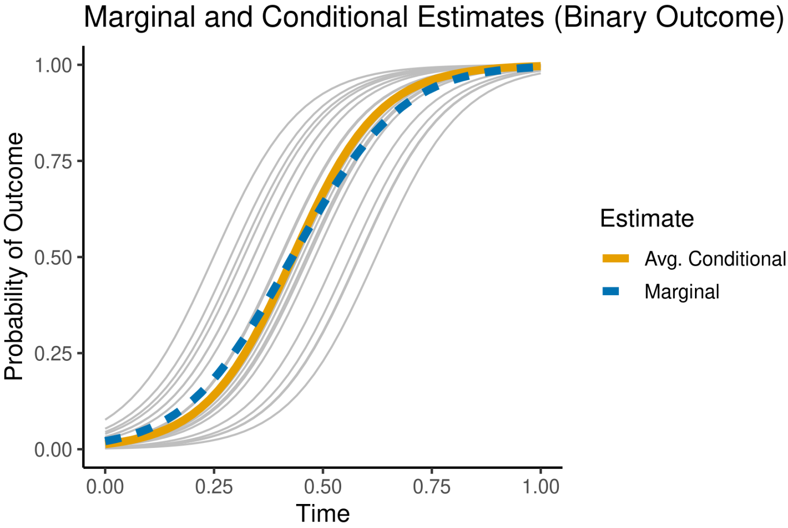
For a GLMM, we use a non-identity link function. Then:

- Taking the expected value over $\boldsymbol{b}_i$:

$$\mathrm{E}_b\left\{g(\mathrm{E}(Y_{ij}|X_{ij},\boldsymbol{b}_i))\right\} = \mathrm{E}_b\left\{g(X_{ij}^\top\boldsymbol{\beta} + Z_{ij}^\top\boldsymbol{b}_i)\right\} \neq g(\mathrm{E}(Y_{ij}|X_{ij}))$$

- So average conditional mean is **not** the same as marginal mean, in general

Conditional $\neq$ Marginal for GLMM



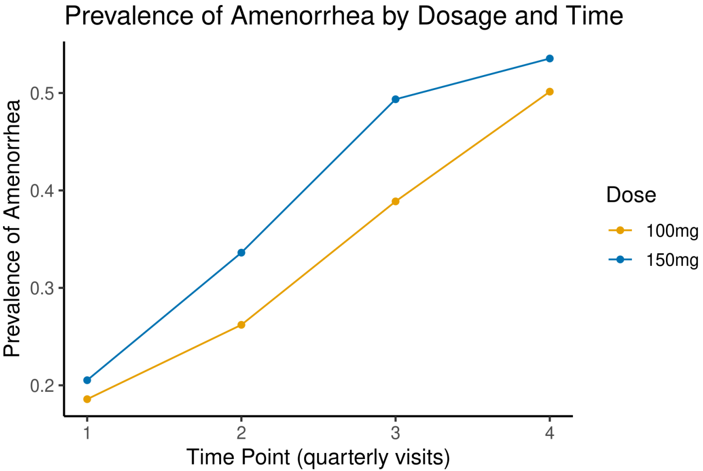
Note: marginal is "attenuated" (smaller estimate) compared to conditional

Interpretation of GLMM

		Fitted model	
Outcome	Coefficient	Random intercept	Random intercept/slope
Continuous	Intercept	Marginal	Marginal
	Slope	Marginal	Marginal
Binary	Intercept	Conditional	Conditional
	Slope	Conditional	Conditional
Count	Intercept	Conditional	Conditional
	Slope	Marginal	Conditional

⋆ Marginal = population-averaged; conditional = subject-specific

Example: Amenorrhea in Contracepting Women



Scientific Questions

- How do subject-specific risks of amenorrhea change over the course of the study?
- What is the influence of dosage on amenorrhea risk?

GLMM for Amenorrhea (random intercepts)

Same fixed effects component of model as GEE, plus random intercepts:

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 \mathrm{t}_{ij} + \beta_2 t_{ij}^2 + \beta_3 \mathrm{Dose}_i \times t_{ij} + \beta_4 \mathrm{Dose}_i \times t_{ij}^2 + b_i$$

where $\mu_{ij} = \Pr(Y_{ij} = 1)$

In R:

```
library(lme4)
glmm.ri <- glmer(amenorrhea ~ visit + visit:dose +
I(visit^2) + I(visit^2):dose + (1 | id),
data = ctcw, nAGQ = 10,
family = "binomial")
```

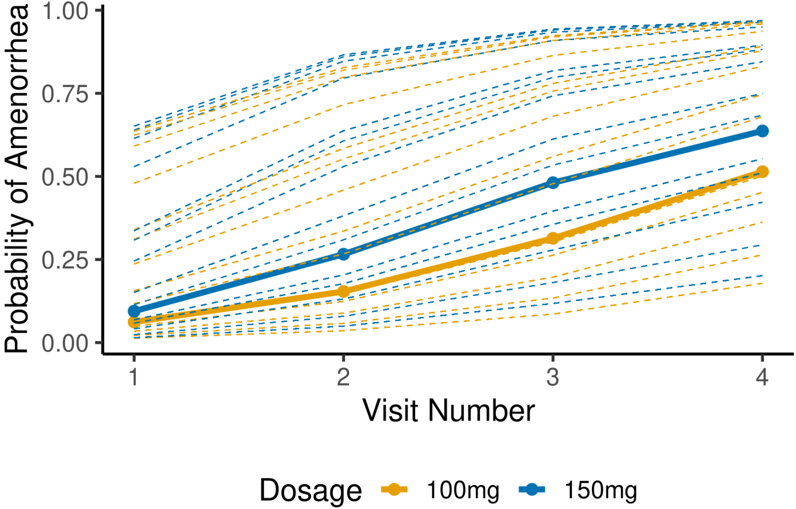
Case Study: Clinical Trial of Contracepting Women

```
Generalized linear mixed model fit by maximum likelihood
(Adaptive Gauss-Hermite Quadrature, nAGQ = 10) [’glmerMod’]
Family: binomial  ( logit )
Formula: amenorrhea ~ visit + visit:dose + I(visit^2) + I(visit^2):dose + (1 | id)
Data: ctcw

Random effects:
Groups Name        Variance Std.Dev.
id     (Intercept) 5.054    2.248
Number of obs: 3616, groups:  id, 1151

Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.80348    0.30470 -12.483  < 2e-16 ***
visit            1.13259    0.26811   4.224  2.4e-05 ***
I(visit^2)      -0.04187    0.05479  -0.764  0.44475
visit:dose       0.56417    0.19214   2.936  0.00332 **
dose:I(visit^2) -0.10952    0.04959  -2.209  0.02721 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Case Study: Clinical Trial of Contracepting Women



GLMM for Amenorrhea (random intercepts)

$$\text{logit}(P(\text{Amenorrhea}_i|\text{Dose}_i, \text{Visit}_{ij}, \boldsymbol{b}_i)) = -3.80 + 1.13 \times \text{Visit}_{ij}$$
$$-0.04 \times \text{Visit}_{ij}^2 + 0.56 \times \text{Dose}_i \times \text{Visit}_{ij} - 0.11 \times \text{Dose}_i \times \text{Visit}_{ij}^2 + b_i$$

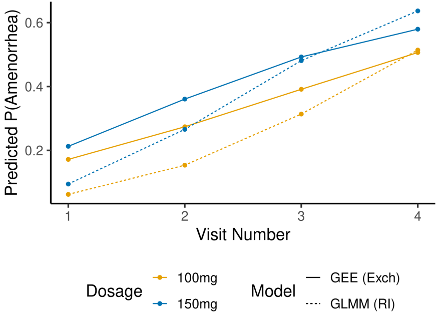
- On average, the odds of amenorrhea at 12 months (Visit 4) for a woman who took the high dose is 1.66 times higher than a woman with the same baseline risk who took the low dose (95% CI: 1.03-2.65).
- Conducting a likelihood ratio test ($H_0 : \beta_3 = \beta_4 = 0$), we find a statistically significant effect of dose (p=0.002)

(NB: conditional interpretations for all parameters in logistic GLMM)

GLMM vs. GEE for Amenorrhea

- GEE: marginal estimates (population-averaged odds ratios)
- GLMM: conditional estimates (within-subject odds ratios)
- Marginal is attenuated relative to conditional
- Note: unlike LMM, random intercept in GLMM does not induce exchangeable correlation
    - So these models do not assume the same correlation structure

Model	GEE (Exch)	GLMM (RI)
$\hat{\beta}_0$ (SE)	-2.2 (0.18)	-3.8 (0.30)
$\hat{\beta}_1$ (SE)	0.70 (0.16)	1.13 (0.27)
$\hat{\beta}_2$ (SE)	-0.03 (0.03)	-0.04 (0.05)
$\hat{\beta}_3$ (SE)	0.33 (0.11)	0.56 (0.19)
$\hat{\beta}_4$ (SE)	-0.06 (0.03)	-0.11 (0.05)



Assumptions and Notes

- Same set of assumptions as LMM:
  - Correct mean model (fixed effects + random effects)
  - Correct covariance specification (random effects)
  - Correct distributional assumptions for $\boldsymbol{Y}_i|\boldsymbol{b}_i$ and $\boldsymbol{b}_i$
- Interpretations:
  - Conditional = marginal for LMM, so can interpret either way
  - Conditional $\neq$ marginal for GLMM; interpret appropriately (usually conditional/subject-specific interpretation)

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

- Recap
- Generalized Linear Mixed Models
- Wrap-Up: Methods for Binary and Count Outcomes
- Activity
- Advanced topics
- Summary

Summary and Comparison of Methods I

	GEE	**GLMM**
Mean model	Marginal	Conditional
Correlation	Model for correlation	Model for population heterogeneity; correlation induced by random effects
Corr. sources	One source (+ or -)	Multiple sources (+ only)
Model type	Semi-parametric (mean & corr.)	Parametric (exponential family)
Target quantities	Mean model	Mean model and within vs. between-subject heterogeneity

Summary and Comparison of Methods II

	GEE	**GLMM**
Estimation	Unbiased estimating equations	Maximum likelihood
Testing	Wald tests	Likelihood ratio or Wald tests
Inference	Marginal (population-averaged)	Conditional (subject-specific)
Missing data assumptions	Missing completely at random (MCAR)	Missing at random (MAR)
Robustness	Robust to correlation model mis-specification	Requires correct parametric model specification
Other notes	Large sample ($N \geq 40$)	Induced marginal mean structure and 'attenuation'
	Model-based or sandwich variance estimator	Typically model-based variance, but can use sandwich (robust) estimator
	Efficiency of non-independence correlation models	

Overview

Day 1: Monday, July 10

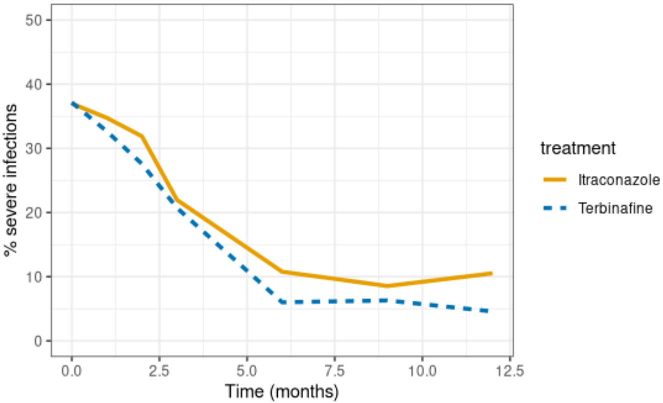
Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

- Recap
- Generalized Linear Mixed Models
- Wrap-Up: Methods for Binary and Count Outcomes
- Activity
- Advanced topics
- Summary

Toenail Infection Analysis

- RCT comparing two oral treatments for a toenail infection
- n=294, most observed at 7 visits (0, 4, 8, 12, 24, 36, 48 weeks)
- Goal: Compare percentage of severe infections over time and between treatment groups



Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

Recap

Generalized Linear Mixed Models

Wrap-Up: Methods for Binary and Count Outcomes

Activity

Advanced topics

- Missing data
- Time-dependent exposures

Summary

Missing data

- Missing values arise in longitudinal studies whenever the intended measurements are not obtained
  - Collect fewer data than planned ⇒ decreased efficiency (power)
  - Missingness can depend on outcome values ⇒ potential bias
- Important to distinguish between missing data and unbalanced data, although missing data necessarily result in unbalanced data
- Missing data require consideration of the factors that influence the missingness of intended observations
- Also important to distinguish between intermittent missing values (non-monotone) and dropouts in which all observations are missing after subjects are lost to follow-up (monotone)

Pattern	t_1	t_2	t_3	t_4	t_5
Monotone	3.8	3.1	2.0	□	□
Non-monotone	4.1	□	3.8	□	□

Mechanisms

In order to obtain valid inference from incomplete data, the mechanism producing the missingness must be considered

- **Missing completely at random** (MCAR)
  Missingness does not depend on **either** the observed or missing responses
- **Missing at random** (MAR)
  Missingness depends **only** on the observed responses
- **Missing not at random** (MNAR)
  Missingness depends on the missing responses

MNAR also referred to as informative or non-ignorable missingness; thus MAR and MCAR as non-informative or ignorable missingness (Rubin, 1976)

Examples and implications

- **MCAR**: Administrative censoring at a fixed calendar time
  - Generalized estimating equations are valid
  - Mixed-effects models are valid
  - Note: If missingness depends on covariates, the covariates need to be incorporated into the analysis (missingness would be a problem if you do not condition on them)
- **MAR**: Study protocol that a subject be removed once the value of an outcome variable is below a certain threshold
  - Generalized estimating equations are not valid
  - Mixed-effects models are valid (as long as model is correctly specified)
- **MNAR**: Outcome is a measure of 'quality-of-life' and subjects fail to complete the questionnaire when their quality-of-life is compromises
  - Generalized estimating equations are not valid
  - Mixed-effects models are not valid

⋆ Without knowing the reasons for missingness or having the missing data cannot know which mechanism for sure.

Missing Data: Analytic Approaches

1. **Complete-case** analysis: use observations from only those that have complete data
   - Valid if data are MCAR
2. **Available data** analysis: use all available observations
   - Valid if data are MCAR
   - Tends to be more efficient than complete case
   - Likelihood-based methods valid if data are MAR and correctly specified
3. **Imputation**: fill in missing values
   - Multiple imputation helps to account for uncertainty
4. **Weighting** methods: accounts for under-representation of certain responses in the observed data by weighting the observed data according to probability of remaining in the study
5. Others...

WGEE

- Extend marginal GEE approach to longitudinal studies with MAR dropout
- Define $R_{ij} = 1$ if jth observation from subject i is observed.
- Observations in the estimating equation are weighted inversely proportional to the probability of being observed

$$\sum_{i=1}^{N} \mathbf{D}_i(\hat{\beta})^\top \mathbf{V}_i^{-1}(\hat{\beta}, \boldsymbol{\alpha}) \boldsymbol{\Delta}_i(\boldsymbol{\theta})[\mathbf{Y}_i - \boldsymbol{\mu}_i(\hat{\beta})] = \mathbf{0}$$

where

$$\boldsymbol{\Delta}_i(\boldsymbol{\theta}) = \begin{pmatrix} R_{i1}w_{i1} & 0 & \cdots & 0 \\ 0 & R_{i2}w_{i2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{in}w_{in} \end{pmatrix}$$

(Robins et al., 1995)

WGEE

- Valid under MAR provided the model for probability of dropout is correct
  - Requires either *a priori* knowledge of the weights w_{ij} or
  - Correctly specified dropout model (the probability of remaining in the study at the current time point, given dropout not occurring at previous time points):

$$\pi_{ij} = \Pr(R_{ij} = 1|, R_{i,j-1} = 1, \mathbf{X}_i, Y_{i1}, \ldots, Y_{i,j-1}) \quad \rightarrow \quad w_{ij} = \left[\prod_{k=1}^{j} \pi_{ik}\right]^{-1}$$

- Usual comments regarding GEE apply:
  - Correct specification for the mean μ and sufficiently large N
  - Use of robust variance estimator provides robustness to misspecification of the correlation structure
  - Choice of working correlation matrix affects efficiency

Example

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 \mathrm{t}_{ij} + \beta_2 t_{ij}^2 + \beta_3 \mathrm{Dose}_i \times t_{ij} + \beta_4 \mathrm{Dose}_i \times t_{ij}^2$$

	$\hat\beta_0$ (SE)	$\hat\beta_1$ (SE)	$\hat\beta_2$ (SE)	$\hat\beta_3$ (SE)	$\hat\beta_4$ (SE)
CC	-2.4 (0.22)	0.77 (0.19)	-0.04 (0.04)	0.25 (0.13)	-0.04 (0.03)
AD	-2.2 (0.18)	0.70 (0.16)	-0.03 (0.03)	0.33 (0.11)	-0.06 (0.03)
WGEE	-2.3 (0.18)	0.71 (0.16)	-0.03 (0.03)	0.34 (0.11)	-0.07 (0.03)

- Complete case (CC) and available data (AD) analyses valid if data are MCAR
- For WGEE, assumed the probability of remaining in the study depends on measurement occasion, dose group, and previous response

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \theta_0 + \theta_1 \mathrm{Dose}_i + \theta_2 t_{ij} + \theta_3 Y_{i,j-1} + \theta_4 \mathrm{Dose}_i \times Y_{i,j-1}$$

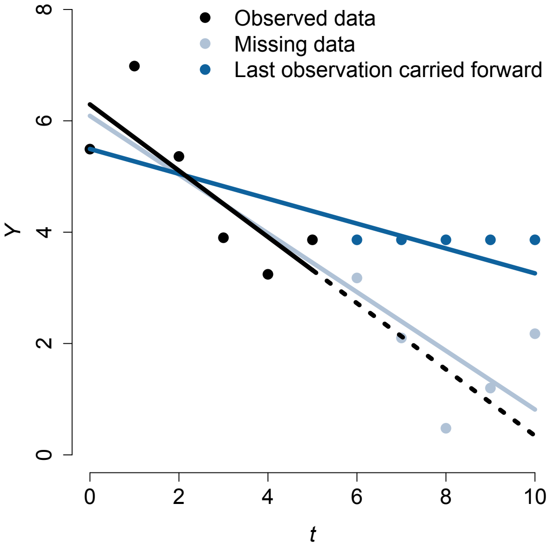
Last observation carried forward

- Extrapolate the last observed measurement to the remainder of the intended serial observations for subjects with any missing data

ID	t_1	t_2	t_3	t_4	t_5
1	3.8	3.1	2.0	$\boxed{2.0}$	$\boxed{2.0}$
2	4.1	3.5	3.8	2.4	2.8
3	2.7	2.4	2.9	3.5	$\boxed{3.5}$

- May result in serious bias in either direction (even when missingness is MCAR)
- May result in anti-conservative *p*-values; variance is understated
- Has been thoroughly repudiated, but still a standard method used by the pharmaceutical industry and appears in published articles
- A refinement would extrapolate based on a regression model for the average trend, which may reduce bias, but still understates variance

Last observation carried forward



Time-dependent exposures

Important analytical issues arise with time-dependent exposures

1. May be necessary to correctly specify the lag relationship over time between outcome $Y_i(t)$ and exposure $X_i(t)$, $X_i(t-1)$, $X_i(t-2), \ldots$ to characterize the underlying biological latency in the relationship
   - **Example**: Air pollution studies may examine the association between mortality on day t and pollutant levels on days t, $t-1$, $t-2, \ldots$
2. May exist exposure endogeneity in which the outcome at time t predicts the exposure at times $t' > t$; motivates consideration of alternative targets of inference and corresponding estimation methods
   - **Example**: If $Y_i(t)$ is a symptom measure and $X_i(t)$ is an indicator of drug treatment, then past symptoms may influence current treatment

Definitions

Factors that influence $X_i(t)$ require consideration when selecting analysis methods to relate a time-dependent exposure to longitudinal outcomes

- **Exogenous**: An exposure is exogenous w.r.t. the outcome process if the exposure at time t is conditionally independent of the history of the outcome process $\mathcal{Y}_i(t) = \{Y_i(s) \mid s \leq t\}$ given the history of the exposure process $\mathcal{X}_i(t) = \{X_i(s) \mid s \leq t\}$

$$[X_i(t) \mid \mathcal{Y}_i(t), \mathcal{X}_i(t)] = [X_i(t) \mid \mathcal{X}_i(t)]$$

- **Endogenous**: Not exogenous

$$[X_i(t) \mid \mathcal{Y}_i(t), \mathcal{X}_i(t)] \neq [X_i(t) \mid \mathcal{X}_i(t)]$$

Examples

Exogeneity may be assumed based on the design or evaluated empirically

- **Observation time**: Any analysis that uses scheduled observation time as a time-dependent exposure can safely assume exogeneity because time is 'external' to the system under study and thus not stochastic
- **Cross-over trials**: Although treatment assignment over time is random, in a randomized study treatment assignment and treatment order are independent of outcomes by design and therefore exogenous
- **Empirical evaluation**: Endogeneity may be empirically evaluated using the observed data by regressing current exposure $X_i(t)$ on previous outcomes $Y_i(t-1)$, adjusting for previous exposure $X_i(t-1)$

$$g(\mathrm{E}[X_i(t)]) = \theta_0 + \theta_1 Y_i(t-1) + \theta_2 X_i(t-1)$$

and using a model-based test to evaluate the null hypothesis: $\theta_1 = 0$

Implications

The presence of endogeneity determines specific analysis strategies

- If exposure is exogenous, then the analysis can focus on specifying the lag dependence of $Y_i(t)$ on $X_i(t)$, $X_i(t-1)$, $X_i(t-2), \dots$
- If exposure is endogenous, then analysts must focus on selecting a meaningful target of inference and valid estimation methods

Targets of inference

With longitudinal outcomes and a time-dependent exposure there are several possible conditional expectations that may be of scientific interest

- **Fully conditional model**: Include the entire exposure process

$$\mathrm{E}[Y_i(t) \mid X_i(1), X_i(2), \ldots, X_i(T_i)]$$

- **Partly conditional models**: Include a subset of exposure process

$$\begin{aligned}&\mathrm{E}[Y_i(t) \mid X_i(t)]\\&\mathrm{E}[Y_i(t) \mid X_i(t-k)] \text{ for } k \leq t\\&\mathrm{E}[Y_i(t) \mid \mathcal{X}_i(t) = \{X_i(1), X_i(2), \ldots, X_i(t)\}]\end{aligned}$$

⋆ An appropriate target of inference that reflects the scientific question of interest must be identified prior to selection of an estimation method

Key assumption

Suppose that primary scientific interest lies in a cross-sectional mean model

$$\mathrm{E}[Y_i(t) \mid X_i(t)] = \beta_0 + \beta_1 X_i(t)$$

To ensure consistency of a generalized estimating equation or likelihood-based mixed-model estimator for β, it is sufficient to assume that

$$\mathrm{E}[Y_i(t) \mid X_i(t)] = \mathrm{E}[Y_i(t) \mid X_i(1), X_i(2), \ldots, X_i(T_i)]$$

Otherwise an **independence** estimating equation should be used

- Known as the full covariate conditional mean assumption
- Implies that with time-dependent exposures must assume exogeneity when using a covariance-weighting estimation method
- The full covariate conditional mean assumption is often overlooked and should be verified as a crucial element of model verification

(Pepe and Anderson, 1994)

Overview

Day 1: Monday, July 10

Day 2: Tuesday, July 11

Day 3: Wednesday, July 12

- Recap
- Generalized Linear Mixed Models
- Wrap-Up: Methods for Binary and Count Outcomes
- Activity
- Advanced topics
- Summary

Summary and Comparison of Methods: Big Picture

	GEE	**GLMM**
Robustness	Provide valid estimates and standard errors for regression parameters of interest even if the correlation model is incorrectly specified (+) Empirical variance estimator requires large sample size (−)	Provide valid estimates and standard errors for regression parameters only under stringent model assumptions (−)
Inference	Always provide population-averaged inference regardless of the outcome distribution; ignores subject-level heterogeneity (+/−)	Provide population-averaged or subject-specific inference depending on the outcome distribution and specified random effects (+/−)
Correlation	Accommodate only one source of correlation (−/+)	Accommodate multiple sources of correlation (+/−)
Missing Data	Require that any missing data are missing completely at random (−)	Require that any missing data are missing at random (−/+)

Advice

- Analysis of longitudinal data is often complex and difficult
- You now have versatile methods of analysis at your disposal
- Each of the methods you have learned has strengths and weaknesses
- Do not be afraid to apply different methods as appropriate
- Always be mindful of the scientific question(s) of interest
- Sensitivity analyses are helpful

Resources

Introductory

- Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Wiley, 2011.
- Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/ Hierarchical Models*. Cambridge University Press, 2007.
- Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. Wiley, 2006.

Advanced

- Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*, 2nd Edition. Oxford University Press, 2002.
- Molenbergs G, Verbeke G. *Models for Discrete Longitudinal Data*. Springer Series in Statistics, 2006.
- Verbeke G, Molenbergs G. *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, 2000.