Module 4

Introduction to Longitudinal Data Analysis

Colleen Sitlani, PhD
University of Washington

Benjamin French, PhD
University of Pennsylvania

SISCR 2017
24 July 2017

Learning objectives

- This module will focus on the design of longitudinal studies, exploratory data analysis, and application of regression techniques based on estimating equations and mixed-effects models
- Focus will be on the practical application of appropriate analysis methods, using illustrative examples in R and Stata
- Some theoretical background and details will be provided; our goal is to translate statistical theory into practical application
- 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

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

Summary and resources

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

Summary and resources

Longitudinal studies

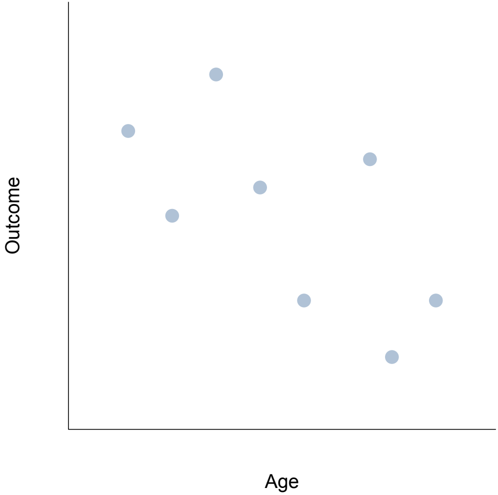
Repeatedly collect information on the same individuals over time

Benefits

- Record incident events
- Ascertain exposure prospectively
- Separate time effects: cohort, period, age
- Distinguish changes over time within individuals
- Offer attractive efficiency gains over cross-sectional studies
- Help establish causal effect of exposure on outcome

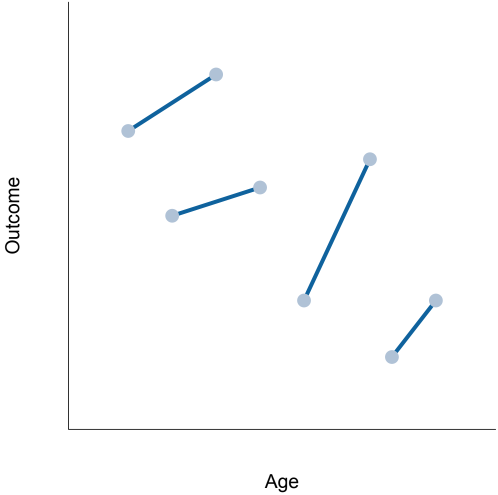
Longitudinal studies

Separate time effects: cohort, age



Longitudinal studies

Separate time effects: cohort, age



Longitudinal studies

Separate time effects: cohort, period, age

- Cohort effects
  - Differences between individuals at baseline
  - “Level”
  - **Example**: Younger individuals begin at a higher level
- Age effects
  - Differences within individuals over time
  - “Trend”
  - **Example**: Outcomes increase over time for everyone
- Period effects may also matter if measurement date varies

Longitudinal studies

Distinguish changes over time within individuals

- We can partition age into two components
  - Cross-sectional comparison

$$\mathrm{E}[Y_{i1}] = \beta_0 + \beta_C x_{i1}$$

  - Longitudinal comparison

$$\mathrm{E}[Y_{ij} - Y_{i1}] = \beta_L(x_{ij} - x_{i1})$$

  for observation $j = 1, \ldots, m_i$ on subject $i = 1, \ldots, n$

- Putting these two models together we obtain

$$\mathrm{E}[Y_{ij}] = \beta_0 + \beta_C x_{i1} + \beta_L(x_{ij} - x_{i1})$$

- β_L represents the expected change in the outcome per unit change in age for a given subject

Longitudinal studies

Offer attractive efficiency gains over cross-sectional studies

- Cross-sectional comparison of treatments A and B

$$\begin{aligned} \hat{\Delta} &= \bar{Y}_1^A - \bar{Y}_1^B \\ \mathrm{Var}[\hat{\Delta}] &= 2\sigma^2 \,/\, n \end{aligned}$$

- Longitudinal comparison of treatments A and B

$$\begin{aligned} \hat{\Delta}^\star &= (\bar{Y}_1^A - \bar{Y}_0^A) - (\bar{Y}_1^B - \bar{Y}_0^B) \\ \mathrm{Var}[\hat{\Delta}^\star] &= 2\sigma^2(2 - 2\rho) \,/\, n \end{aligned}$$

- Longitudinal estimate may be more precise
- May ameliorate bias because each subject "acts as their own control"

Longitudinal studies

Help establish causal effect of exposure on outcome

- Cross-sectional study

$$\begin{array}{rcl} \text{Egg} & \rightarrow & \text{Chicken} \\ \text{Chicken} & \rightarrow & \text{Egg} \end{array}$$

- Longitudinal study

$$\text{Bacterium} \rightarrow \text{Dinosaur} \rightarrow \text{Chicken}$$

⋆ 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

Challenges

- Determine causality when covariates vary over time
- Choose exposure lag when covariates vary over time
- Account for incomplete participant follow-up
- Require specialized methods that account for longitudinal correlation

Longitudinal studies

Require specialized methods that account for longitudinal correlation

- Individuals are assumed to be independent
- Longitudinal dependence may be a secondary feature
- 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

Motivating examples

Dental growth (Patthoff and Roy, 1964)

- Model growth among 11 females and 16 males, ages 8 to 14 years

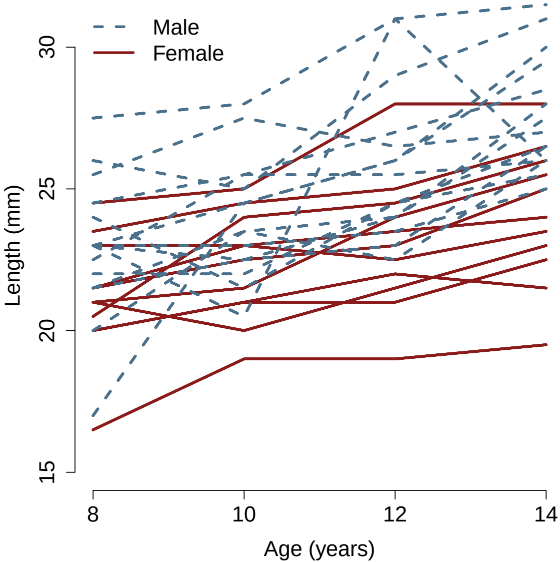
Treatment of lead-exposed children (TLC) (*Pediatric Research*, 2000)

- Assess treatment benefit via blood lead levels in $n = 100$ children

Dental growth

- Model growth among 11 females and 16 males, ages 8 to 14 years
- Distance between the pituitary gland and the pterygomaxillary fissure
- Characterize dental growth among children
    1. Estimate the average growth curve among all children
    2. Estimate the growth curve for individual children
    3. Characterize the degree of heterogeneity across children
    4. Identify factors that predict growth

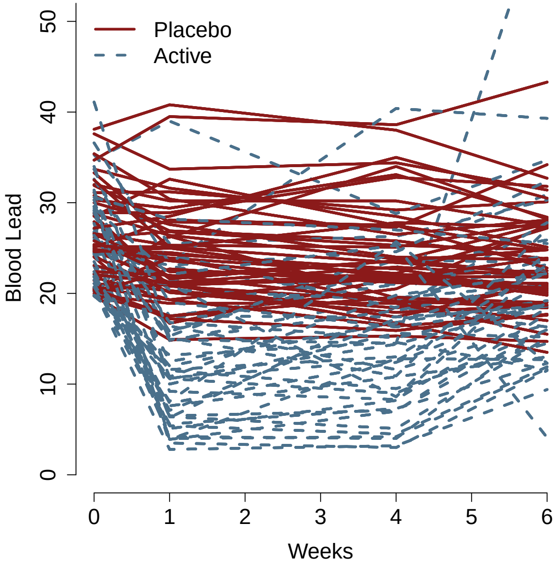
Dental growth: Data



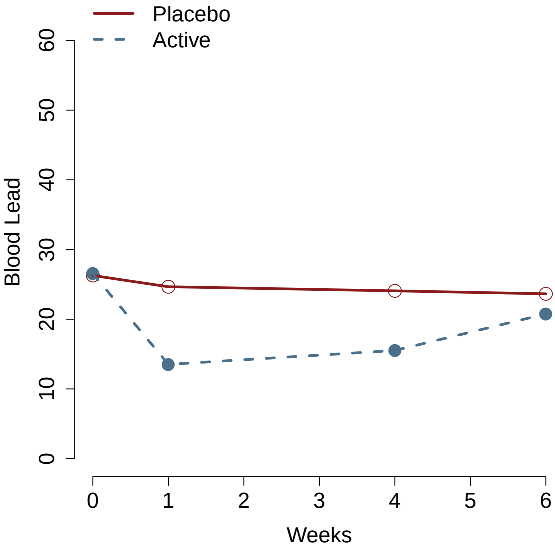
TLC trial

- Assess treatment benefit via blood lead levels in $n = 100$ children
- Placebo-controlled randomized trial of new chelating agent *succimer*
- 50 placebo and 50 active
- Balanced and complete data

TLC trial: Data



TLC trial: Means



Options for analysis of change

Does mean change differ across groups?

- Consider simple situation with
  - Baseline measurement ($t = 0$)
  - Single follow-up measurement ($t = 1$)
- Analysis options for simple pre-post design
  - Analysis of POST only
  - Analysis of CHANGE (post-pre)
  - Analysis of POST controlling for BASELINE
  - Analysis of CHANGE controlling for BASELINE

Randomized pre-post data: Table of means

Group	Baseline	Follow-up	Change
Control	μ_0	$\mu_0 + \Delta_T$	Δ_T
Treatment	μ_0	$\mu_0 + \Delta_T + \Delta_1$	$\Delta_T + \Delta_1$
Difference	0	Δ_1	Δ_1

Randomized pre-post data

- Randomization ensures same baseline mean
- Comparison of means at follow-up (POST) show impact of treatment

$$\begin{aligned}\bar{Y}_1(0) &= \text{sample mean of control at } t = 1\\ \bar{Y}_1(1) &= \text{sample mean of treated at } t = 1\\ \mathrm{E}[\bar{Y}_1(1) - \bar{Y}_1(0)] &= \Delta_1\end{aligned}$$

- Comparison of mean CHANGE shows same impact of treatment

$$\begin{aligned}\bar{C}_1(0) &= \mathrm{E}[Y_{i1}(0) - Y_{i0}(0)] = \bar{Y}_1(0) - \mu_0\\ \bar{C}_1(1) &= \mathrm{E}[Y_{i1}(1) - Y_{i0}(1)] = \bar{Y}_1(1) - \mu_0\\ \mathrm{E}[\bar{C}_1(1) - \bar{C}_1(0)] &= \Delta_1\end{aligned}$$

Randomized pre-post data

- With assumption of equal means at baseline, ANCOVA (POST controlling for BASELINE) also an option

$$E[Y_{i1} \mid X_i, Y_{i0}] = \beta_0 + \beta_1 \cdot X_i + \gamma \cdot Y_{i0}$$

- $\beta_1 = \Delta_1$ because, averaging over Y_{i0}

$$\begin{aligned} E[\bar{Y}_1(1) - \bar{Y}_1(0)] &= (\beta_0 + \beta_1 + \gamma \cdot E[Y_{i0} \mid X_{i0} = 1]) \\ &\qquad - (\beta_0 + \gamma \cdot E[Y_{i0} \mid X_{i0} = 0]) \\ &= \beta_1 \end{aligned}$$

- Equivalent to CHANGE controlling for BASELINE

$$\begin{aligned} E[Y_{i1} - Y_{i0} \mid X_i, Y_{i0}] &= E[Y_{i1} \mid X_i, Y_{i0}] - Y_{i0} \\ &= \beta_0 + \beta_1 \cdot X_i + (\gamma - 1) \cdot Y_{i0} \end{aligned}$$

Summary of options

Method	Expected value	Variance
POST	Δ_1	???
CHANGE	Δ_1	???
ANCOVA	Δ_1	???

Summary of options

- Assume n participants per group
- Assume same variance (σ^2) at $t = 0$ and $t = 1$
- Assume correlation between Y_{i0} and Y_{i1} is ρ

Method	Expected value	Variance
POST	Δ_1	$2 \cdot \sigma^2/n$
CHANGE	Δ_1	$2 \cdot \sigma^2(2 - 2\rho)/n$
ANCOVA	Δ_1	$2 \cdot \sigma^2(1 - \rho^2)/n$

Summary of options

- See Frison and Pocock (1992) for details regarding these results
- Implies we can order methods from worst to best w.r.t. precision
  - $\boxed{\rho > 1/2}$ POST $\prec$ CHANGE $\prec$ ANCOVA
  - $\boxed{\rho < 1/2}$ CHANGE $\prec$ POST $\prec$ ANCOVA

TLC trial: Randomized pre-post example

	week 1		week 4		week 6	
Method	est.	(s.e.)	est.	(s.e.)	est.	(s.e.)
POST	11.138	(1.332)	8.556	(1.377)	2.284	(1.532)
CHANGE	11.406	(1.120)	8.824	(1.152)	3.152	(1.257)
ANCOVA	11.341	(1.099)	8.765	(1.137)	3.120	(1.258)

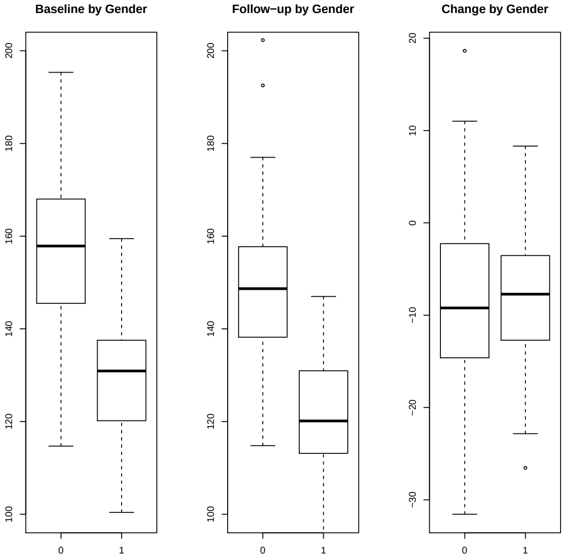
Summary: Change and randomized studies

- Key assumption: groups equivalent at baseline
- Methods that 'adjust' for baseline are generally preferable due to greater precision
  - CHANGE analysis adjusts for baseline by subtracting it from follow-up
  - ANCOVA analysis adjusts for baseline by controlling for it in a model
- Missing data will impact each approach

Non-randomized pre-post data: Example (created)

- Fitzmaurice (2001) *Nutrition* article discusses analysis of randomized and non-randomized studies of change
- Hypothetical study of weight loss pill call *Diagra* (by Fitzmaurice!)
- Non-randomized study
    - **MaWoD** = men and women on Diagra
    - Is weight change on Diagra the same for men and for women?
    - Groups **not equal** at baseline, in terms of outcome

MaWoD trial: Distributions

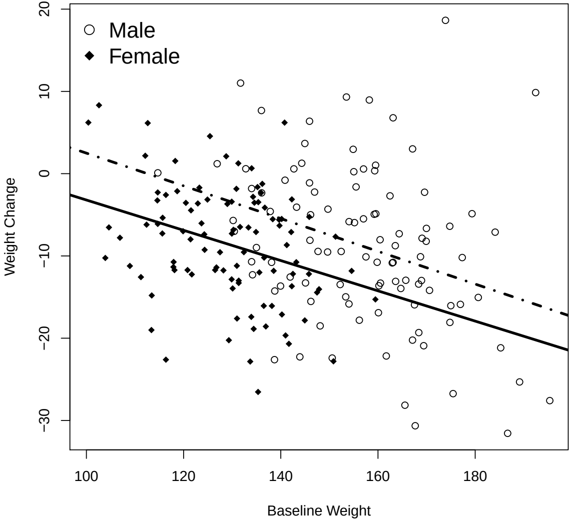


MaWoD trial

Method	est.	(s.e.)
POST	-29.99	(1.34)
CHANGE	-1.29	(0.89)
ANCOVA	-6.33	(1.25)

- POST not helpful, because groups different at baseline
- CHANGE useful to evaluate whether the data suggest a different reduction in males vs females
- ANCOVA compares men and women with the same weight at baseline; is this useful?

MaWoD trial: Change versus pre



Non-randomized pre-post data

- ANCOVA – by comparing the mean follow-up weight among men and women with equal weights at baseline, this is likely to be
  - A man who is *lighter* than average for men
  - A woman who is *heavier* than average for women
- Regression to the mean tells us that we should expect lighter men to get heavier and heavier women to get lighter
- Therefore, we expect the women to have a smaller mean weight at follow-up compared to the men

Summary: Non-randomized pre-post data

- Baseline equivalence no longer guaranteed
- Methods no longer answer same scientific question
  - POST: How different are groups at follow-up?
  - CHANGE: How different is the change in outcome for the two groups?
  - ANCOVA: What is the expected difference in the mean outcome at follow-up across the two groups, controlling for the baseline value of the outcome? [β_1 is a function of both Δ_1 and baseline difference]
- CHANGE typically most relevant; multivariable methods to come later characterize CHANGE across multiple timepoints

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

Summary and resources

More general longitudinal data

- Simple pre-post data can use analytic tools that don't incorporate correlation within individuals
- Material that follows leads toward GEE and mixed-effects models
  - Exploratory data analysis
  - Regression model specification
  - Parameter interpretation
  - Covariance and correlation

Notation

$\boxed{\text{Define}}$

$$
\begin{aligned}
m_i &= \text{number of observations for subject } i = 1, \ldots, n \\
Y_{ij} &= \text{outcome for subject } i \text{ at time } j = 1, \ldots, m_i \\
X_i &= (x_{i1}, x_{i2}, \ldots, x_{im_i}) \\
x_{ij} &= (x_{ij1}, x_{ij2}, \ldots, x_{ijp}) \\
&\quad \text{exposure, covariates}
\end{aligned}
$$

Stacks of data for each subject:

$$
Y_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im_i} \end{bmatrix}
\qquad
X_i = \begin{bmatrix} x_{i11} & x_{i12} & \cdots & x_{i1p} \\ x_{i21} & x_{i22} & \cdots & x_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{im_i1} & x_{im_i2} & \cdots & x_{im_ip} \end{bmatrix}
$$

Exploratory data analysis

Exploratory data analysis for longitudinal data

- Summary statistics over time (by groups)
- Individual plots of observed and fitted values
- Empirical covariance structure (variance and correlation)

Goal: Summarize mean and covariance structure

Exploratory data analysis: Guidelines

1. Show as much of the data as possible, rather than only summaries
2. Highlight aggregate patterns of potential scientific interest
3. Identify both cross-sectional and longitudinal patterns
4. Facilitate the identification of unusual individuals or observations

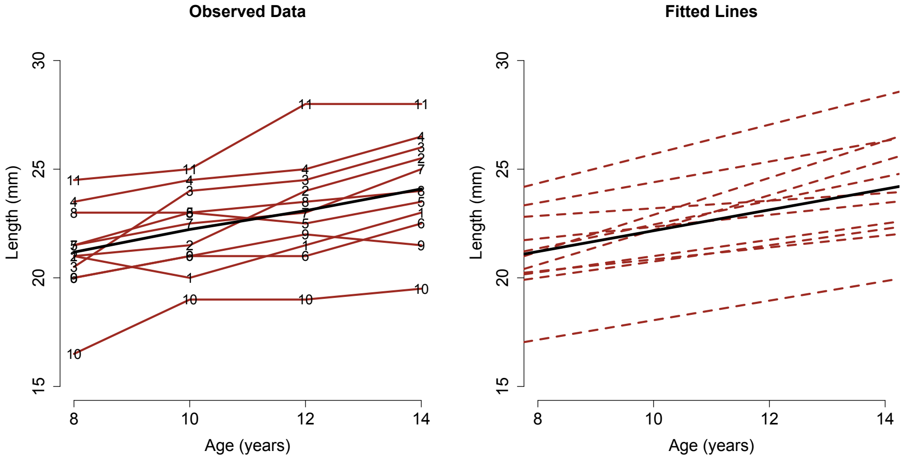
Dental growth: Summary statistics

	Mean Length (mm)			
	Age 8	Age 10	Age 12	Age 14
Males	22.9	24.0	25.9	27.6
Females	21.2	22.2	23.1	24.1
Difference	1.8	1.7	2.8	3.5

On average. . .

- **Trend**: Dental length increases over time for males and females
- **Cross-sectional**: Males have larger dental length at every age
- **Longitudinal**: Increase in average dental length is larger for males

Dental growth: Individual plots for females



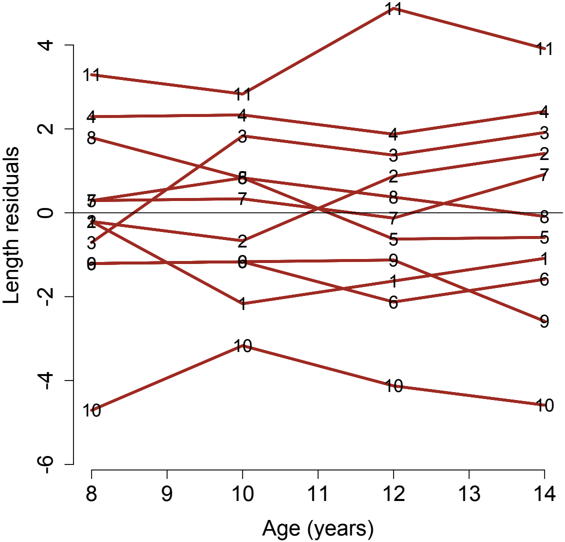
Dental growth: Individual plots for females

- **Trend**: Dental length in females increases over time
- **Tracking**: Females with large dental length at younger ages tend to have large dental length at older ages
- **Variability**: Dental length appears to be slightly more variable at older ages (verify using empirical estimates)
- **Outliers**
  - Subjects 1, 5, and 9 have a periodic decrease in dental length
  - Subject 10 appears to have small dental length, especially at age 8
  - Subject 11 appears to have large dental length, especially at age 12
  - **NB**: Outliers are hard to judge with only 11 subjects

Individual plots: Difficulties

- **Issue**: Individual plots may not be useful for large datasets
- **Issue**: Random selection of individual lines may be arbitrary
- **Solution**: Produce plots for well-defined groups
  - Example: Individual plots of dental growth for females
- **Issue**: Individual patterns may be difficult to detect in raw data
  - Example: Individual plots of dental growth for females
- **Solution**: Plot marginalized residuals versus time for individuals
  - Example: Individual plots of dental growth residuals for females

Dental growth: Individual plots of residuals



Dental growth: Individual plots of residuals

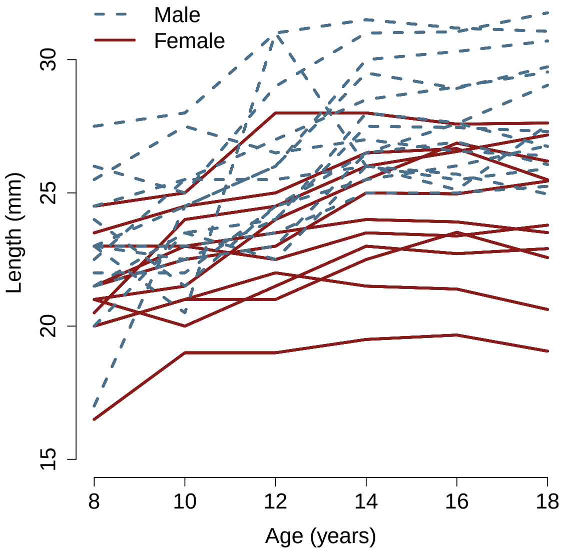
Question: What are the advantages in examining residuals?

Answer

- Easier to identify individual patterns because it's generally easier to see variation across a flat line rather than a sloped line
- Facilitates the identification of unusual individuals or observations *given the average temporal trend*
  - Example: Dental length for subjects 8 and 10 increases over time, but their increase is smaller than the average increase

⋆ If we wish to study the random variation in the outcome over time, then we must remove the systemic variation due to temporal trends using residuals with a thorough and flexible adjustment for time

Dental data: Random extension



Categorical model for time

- $\boxed{\text{Mean Model}}$

$$E[Y_{ij} \mid \text{Age}_{ij}] = \beta_0 + \beta_1[\text{Age}_{ij} == 10] + \cdots + \beta_5[\text{Age}_{ij} == 18]$$

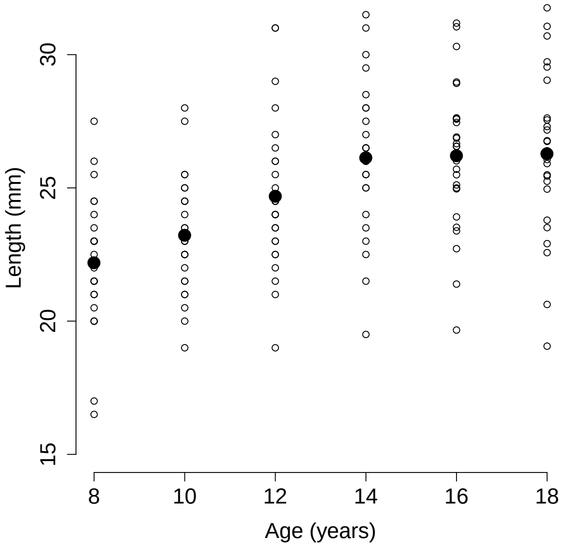
- $\boxed{\text{Rate of Change}}$
    - zero (flat) within each age, then jumps at new age

Fake dental data: Categorical model for time

```
lm(formula = length ~ as.factor(time), data = growthmore)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         22.192      0.532   41.73  < 2e-16 ***
as.factor(time)10    1.038      0.752    1.38   0.1694
as.factor(time)12    2.500      0.752    3.32   0.0011 **
as.factor(time)14    3.942      0.752    5.24  5.3e-07 ***
as.factor(time)16    4.005      0.752    5.32  3.6e-07 ***
as.factor(time)18    4.083      0.752    5.43  2.2e-07 ***
```

Fake dental data: Categorical model for time



Linear model for time

- $\boxed{\text{Mean Model}}$

$$\mathsf{E}[Y_{ij} \mid \mathsf{Age}_{ij}] = \beta_0 + \beta_1 \mathsf{Age}_{ij}$$

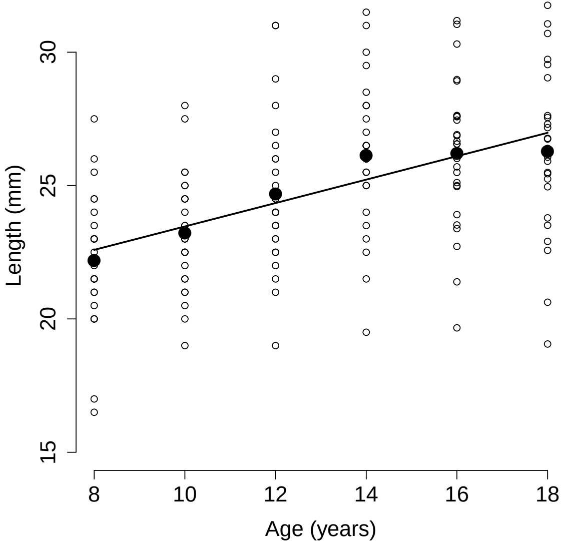
- $\boxed{\text{Rate of Change}}$
  - slope of the curve β_1
  - constant rate of change

Fake dental data: Linear model for time

```
lm(formula = length ~ time, data = growthmore)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.076      0.860   22.19  < 2e-16 ***
time           0.439      0.064    6.87  1.5e-10 ***
```

Fake dental data: Linear model for time



Quadratic model for time

- Mean Model

$$E[Y_{ij} \mid \text{Age}_{ij}] = \beta_0 + \beta_1 \text{Age}_{ij} + \beta_2 \text{Age}_{ij}^2$$

- Rate of Change
  - slope of the curve $\beta_1 + 2 \cdot \text{Age}_{ij} \cdot \beta_2$
  - non-constant rate of change

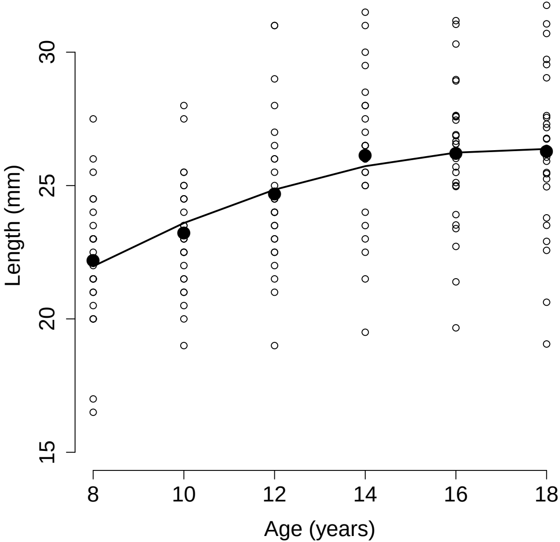
Fake dental data: Quadratic model for time

```
growthmore <- within(growthmore, {
  time2 <- timeˆ2})

mquad <- lm(length˜time+time2, data=growthmore)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7712     3.5096    3.35   0.0010 **
time          1.6464     0.5663    2.91   0.0042 **
time2        -0.0464     0.0216   -2.15   0.0335 *
```

Fake dental data: Quadratic model for time



Other models for time

- Linear spline
- Cubic spline
- Higher-order polynomials

- Useful for data that are not balanced
- Require careful handling when interactions with time are modeled

Choosing time scale(s)

- **Age**: use Age_{ij} as time variable
  - Assumes: growth from age 8 to age 10 experienced 1990–1992 is the same as that from age 8 to age 10 experienced 2000–2002
  - (e.g. no **period** effects)
- **Age-since-entry**: use $\text{Age}_{ij} - \text{Age}_{i1}$ as time variable
  - Assumes: growth experienced 1990–1992 is same for children who aged from 8 to 10 years old, and children who aged from 12 to 14 years old
  - (e.g. no **cohort** effects)
- **Age-at-entry**: use Age_{i1} as time variable
  - Assumes: children may be different at entry to study, but do not change further during follow-up
  - (e.g. no **aging** effects)

Dental growth: Scientific questions as regression

- Questions concerning the rate of growth refer to the time slope for dental length

$$\mathrm{E}[\mathrm{Length}_{ij} \mid x_{ij} = \{\mathrm{Age},\ \mathrm{Gender}\}] = \beta_0(x_{ij}) + \beta_1(x_{ij}) \cdot \mathrm{Time}_{ij}$$

- Does the rate of growth differ for males as compared to females?

$$\mathrm{E}[Y_{ij}] = \beta_0 + \beta_1(\mathrm{Age}_{ij} - 8) + \beta_2 \mathrm{Gender}_i + \beta_3(\mathrm{Age}_{ij} - 8) \cdot \mathrm{Gender}_i$$

Dental growth: Parameter interpretation

$$\mathrm{E}[Y_{ij}] = \beta_0 + \beta_1(\mathrm{Age}_{ij} - 8) + \beta_2\mathrm{Gender}_i + \beta_3(\mathrm{Age}_{ij} - 8) \cdot \mathrm{Gender}_i$$

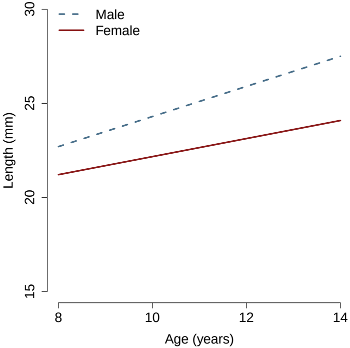
If Gender $= \{1 =$ male; $0 =$ female$\}$

- $\beta_1 =$ expected dental growth (per year) for females
- $\beta_2 =$ expected difference in dental length comparing 8-year-old males to 8-year-old females
- $\beta_3 =$ expected difference in dental growth (per year) between males and females

Dental growth: Regression model

```
model <- lm(length ~ I(age-8)*gender, data=growth)
```

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             21.209      0.570   37.21   <2e-16 ***
I(age - 8)               0.480      0.152    3.15   0.0022 **
gendermale               1.491      0.750    1.99   0.0497 *
I(age - 8):gendermale    0.320      0.201    1.60   0.1133
```



Dependence and correlation

$\boxed{\text{Issue}}$ Response variables measured on the same subject are correlated

- Observations are **independent** when deviation in one variable does not predict deviation in the other variable
  - Given two sujects with the same age and gender, then the dental length for patient ID=14 is not predictive of the dental length for patient ID=9
- Observations are **dependent** or **correlated** when one variable does predict the value of another variable
  - The dental length for patient ID=14 at age 10 is predictive of the dental length for patient ID=14 at age 12

Dependence and correlation: Variance review

- Recall: The variance of a variable Y_{ij} (fix time j) is defined as:

$$\begin{aligned}\sigma_j^2 &= \mathrm{E}[(Y_{ij} - \mu_j)^2] \\ &= \mathrm{E}[(Y_{ij} - \mu_j)(Y_{ij} - \mu_j)]\end{aligned}$$

- The variance measures the average distance that an observation falls away from the mean

Dependence and correlation: Covariance

- $\boxed{\text{Define:}}$ The **covariance** of two variables Y_{ij} and Y_{ik} is

$$\sigma_{jk} = \mathrm{E}[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)]$$

- The covariance measures whether, on average, departures in one variable $Y_{ij} - \mu_j$ ‘go together with’ departures in a second variable $Y_{ik} - \mu_k$
- In simple linear regression of Y_{ij} on Y_{ik} the regression coefficient β_1 in $\mathrm{E}[Y_{ij} \mid Y_{ik}] = \beta_0 + \beta_1 \cdot Y_{ik}$ is the covariance divided by the variance of Y_{ik}

$$\beta_1 = \frac{\sigma_{jk}}{\sigma_k^2}$$

Dependence and correlation: Correlation

- $\boxed{\text{Define:}}$ The **correlation** of two variables Y_{ij} and Y_{ik} is

$$\rho_{jk} = \frac{\mathrm{E}[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)]}{\sigma_j \sigma_k}$$

- The correlation is a measure of dependence that takes values between -1 and $+1$
- Recall that a correlation of 0 implies that two measures are unrelated (linearly)
- Recall that a correlation of 1 implies that the two measures fall perfectly on a line – one exactly predicts the other!

Covariance: Something new to model

$$
\begin{aligned}
\mathrm{Cov}[Y_i] &= \begin{bmatrix} \mathrm{Var}[Y_{i1}] & \mathrm{Cov}[Y_{i1}, Y_{i2}] & \cdots & \mathrm{Cov}[Y_{i1}, Y_{im_i}] \\ \mathrm{Cov}[Y_{i2}, Y_{i1}] & \mathrm{Var}[Y_{i2}] & \cdots & \mathrm{Cov}[Y_{i2}, Y_{im_i}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[Y_{im_i}, Y_{i1}] & \mathrm{Cov}[Y_{im_i}, Y_{i2}] & \cdots & \mathrm{Var}[Y_{im_i}] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_{m_i}\rho_{1m_i} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \cdots & \sigma_2\sigma_{m_i}\rho_{2m_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m_i}\sigma_1\rho_{m_i1} & \sigma_{m_i}\sigma_2\rho_{m_i2} & \cdots & \sigma_{m_i}^2 \end{bmatrix}
\end{aligned}
$$

Dental growth: Covariances

	Females				Males			
	Age 8	Age 10	Age 12	Age 14	Age 8	Age 10	Age 12	Age 14
Age 8	4.51	3.35	4.33	4.36	6.39	2.30	3.74	1.56
Age 10	3.35	3.62	4.03	4.08	2.30	4.48	1.96	2.58
Age 12	4.33	4.03	5.59	5.47	3.74	1.96	7.16	3.05
Age 14	4.36	4.08	5.47	5.94	1.56	2.58	3.05	4.20

Dental growth: Correlations

	Females				Males			
	Age 8	Age 10	Age 12	Age 14	Age 8	Age 10	Age 12	Age 14
Age 8	1.0	0.83	0.86	0.84	1.0	0.43	0.55	0.30
Age 10	0.83	1.0	0.90	0.88	0.43	1.0	0.35	0.59
Age 12	0.86	0.90	1.0	0.95	0.55	0.35	1.0	0.56
Age 14	0.84	0.88	0.95	1.0	0.30	0.59	0.56	1.0

Dental growth: Comments on covariance structure

- Covariance of raw outcomes same as covariance of residuals due to lack of covariates
- In females, some indication that the variance increases with the mean
- Similar magnitude of variance in males vs females
- Clear correlation among observations on the same individual, though correlation in males lower than that in females
- **NB**
  - Must also examine sample size in each cell to assess relative confidence in each estimate (here we have balanced and complete data)
  - Producing covariance and correlation matrices requires categorizing continuous time into a reasonable number of categories

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

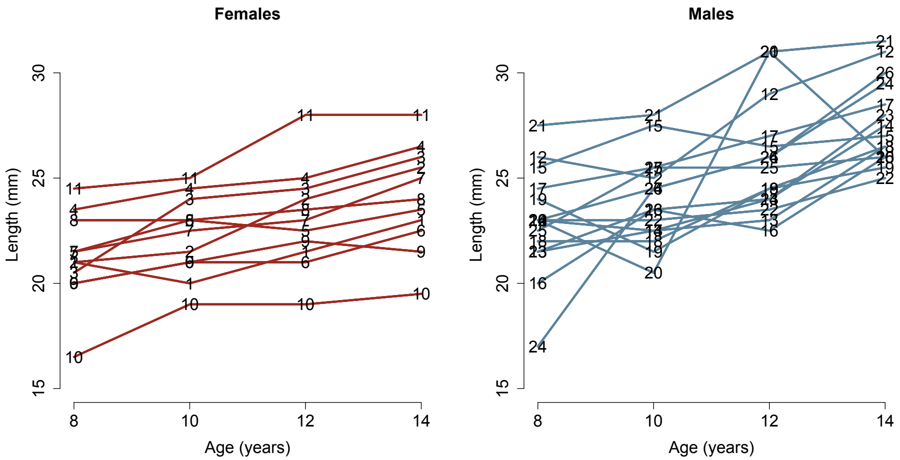
Summary and resources

Dental growth

Goal: Characterize dental growth among children, ages 8 to 14 years

1. Estimate the average growth curve among all children
2. Estimate the growth curve for individual children
3. Characterize the degree of heterogeneity across children
4. Identify factors that predict growth

Dental growth



GEE (Liang and Zeger, 1986)

9705 citations as of June 2017

⋆ 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 β, which characterize the systemic variation in Y across covariates X

$$\begin{aligned} Y_i &= (Y_{i1}, Y_{i2}, \ldots, Y_{im_i})^{\mathsf{T}} \\ X_i &= (x_{i1}, x_{i2}, \ldots, x_{im_i})^{\mathsf{T}} \\ x_{ij} &= (x_{ij1}, x_{ij2}, \ldots, x_{ijp}) \\ \beta &= (\beta_1, \beta_2, \ldots, \beta_p)^{\mathsf{T}} \end{aligned}$$

for $i = 1, \ldots, n$; $j = 1, \ldots, m_i$; and $k = 1, \ldots, p$

- Longitudinal correlation structure is a nuisance feature of the data

Mean model

Assumptions

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

Mean model: Primary focus of the analysis

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

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

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

- Characterizes a **marginal** mean regression model

Marginal mean

Definition: μ_{ij} does not condition on anything other than x_{ij}

- **Mixed-effects model**: Use subject-specific random effects γ_i to induce a correlation structure

$$g(\mathrm{E}[Y_{ij} \mid x_{ij}, \gamma_i]) = x_{ij}(\beta^\star + \gamma_i)$$

- **Transition model**: Model the conditional expectation as a function of covariates and previous outcomes $\mathcal{Y}_{ij}$

$$g(\mathrm{E}[Y_{ij} \mid x_{ij}, \mathcal{Y}_{ij}]) = x_{ij}\beta^{\star\star} + \mathcal{Y}_{ij}\alpha$$

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}

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

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

2. Select a model for longitudinal **correlation** with parameters α

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

Covariance model: General notation

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

$$\begin{aligned}
\mathrm{Var}[Y_{ij} \mid X_i] &= V(\mu_{ij}) \\
S_i(\mu_i) &= \text{diag } V(\mu_{ij})
\end{aligned}$$

$$\begin{aligned}
\mathrm{Corr}[Y_{ij},\, Y_{ij'} \mid X_i] &= \rho(\alpha) \\
R_i(\alpha) &= \text{matrix } \rho(\alpha)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}[Y_i \mid X_i] &= V_i(\beta, \alpha) \\
&= S_i^{1/2} R_i S_i^{1/2}
\end{aligned}$$

Correlation models

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

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

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

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

Correlation models

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

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

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

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

Correlation models

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

- **Independence**: ...is assumed to be zero
  - Always 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 m) 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
- Additional assumptions are required to identify a complete probability model and a corresponding parametric likelihood function (GLMM)

Question: Without a likelihood function, how do we estimate β 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

$$\begin{aligned} \mathcal{U}_\beta(\beta, \alpha) &= \sum_{i=1}^{n} D_i^\mathsf{T} V_i^{-1}(Y_i - \mu_i) \\ \mu_i &= g^{-1}(X_i\beta) \\ D_i &= \frac{\partial \mu_i}{\partial \beta} \end{aligned}$$

- V_i is the 'working' variance-covariance matrix: $\text{Cov}[Y_i \mid X_i]$
  - Depends on the assumed form for the variance: $\text{Var}[Y_{ij} \mid x_{ij}]$
  - Depends on the specified correlation model: $\text{Corr}[Y_{ij}, Y_{ij'} \mid X_i]$
- V_i may also be written as a covariance weight matrix: $W_i = V_i^{-1}$
- $\mathcal{U}_\beta(\beta, \alpha)$ depends on the model or value for α

Generalized estimating equations

Setting an estimation function equal to 0 defines an estimating equation

$$
\begin{aligned}
0 &= \mathcal{U}_\beta(\hat{\beta}, \alpha) \\
&= \sum_{i=1}^{n} D_i^{\mathsf{T}} V_i^{-1} (Y_i - \hat{\mu}_i)
\end{aligned}
$$

with $\hat{\mu}_i = g^{-1}(X_i \hat{\beta})$

- 'Generalized' because it corresponds to a GLM with link function $g(\cdot)$
- Solution to the estimation equation defines an estimator $\hat{\beta}$
- $\mathcal{U}_\beta(\hat{\beta}, \alpha)$ depends on the model or value for α
  - Moment-based estimation of α based on residuals
  - A second set of estimating equations for α

Generalized estimating equations: Intuition

$$0 \;=\; \sum_{i=1}^{n} \underbrace{D_i^{\mathsf{T}}}_{\boxed{3}} \underbrace{V_i^{-1}}_{\boxed{2}} (\underbrace{Y_i - \hat{\mu}_i}_{\boxed{1}})$$

$\boxed{1}$ The model for the mean, $\mu_i(\beta)$, is compared to the observed data, Y_i; setting the equations to equal 0 tries to minimize the difference between **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, $\mu_i(\beta)$, to the scale of the regression coefficients (covariates)

Properties of $\hat{\beta}$

Suppose Y_i is continuous so that $\mathrm{E}[Y_i \mid X_i] = X_i\beta$ and $\mathrm{Cov}[Y_i \mid X_i] = V_i$

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} Y_i$$

- $\hat{\beta}$ is **unbiased** assuming $\mathrm{E}[Y_i \mid X_i] = X_i\beta$ is correct

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} \mathrm{E}[Y_i] \\ &= \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i \beta \\ &= \beta \end{aligned}$$

Properties of $\hat{\beta}$

- $\hat{\beta}$ is **efficient** assuming $\text{Cov}[Y_i \mid X_i] = V_i$ is correct

$$\begin{aligned}\text{Cov}[\hat{\beta}] &= \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i\right)^{-1} \\ &\quad \times \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} \text{Cov}[Y_i] V_i^{-1} X_i\right) \\ &\quad \times \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i\right)^{-1} \\ &= \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i\right)^{-1}\end{aligned}$$

which is known as the model-based variance estimator

Properties of $\hat{\beta}$

If $\text{Cov}[Y_i \mid X_i] \neq V_i$, then use an empirical estimator

$$
\begin{aligned}
\text{Cov}[\hat{\beta}] \quad = \quad & \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i \right)^{-1} \\
& \times \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^{\mathsf{T}} V_i^{-1} X_i \right) \\
& \times \left(\sum_{i=1}^{n} X_i^{\mathsf{T}} V_i^{-1} X_i \right)^{-1}
\end{aligned}
$$

- Also known as sandwich, robust, or Huber-White variance estimator
- Requires sufficiently large sample size ($n \geq 40$)
- Requires sufficiently large sample size relative to cluster size ($n \gg m$)

$\text{Cov}[\hat{\beta}]$

$(Y_i - \mu_i)(Y_i - \mu_i)^\mathsf{T}$ is a poor estimate of $\text{Cov}[Y_i]$ for each i

- However, a good estimate for each i is not required
- Rather, need a good estimate of the average (total) covariance

$$\begin{aligned} B_n &= \frac{1}{n}\sum_{i=1}^{n} D_i^\mathsf{T} V_i^{-1} \text{Cov}[Y_i] V_i^{-1} D_i \\ \hat{B}_n &= \frac{1}{n}\sum_{i=1}^{n} D_i^\mathsf{T} V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^\mathsf{T} V_i^{-1} D_i \end{aligned}$$

- $\hat{B}_n$ can be well estimated with sufficient independent replication, i.e. sufficiently large sample size relative to cluster size

Properties of $\hat{\beta}$

- $\hat{\beta}$ is a consistent estimator for β even if the model for longitudinal correlation is incorrectly specified, i.e. $\hat{\beta}$ is 'robust' to correlation model mis-specification
- However, the variance of $\hat{\beta}$ must capture the correlation in the data, either by choosing the correct correlation model, or via an alternative variance estimator
- Selecting an approximately correct correlation model will yield a more efficient estimator for β, i.e. $\hat{\beta}$ has the smallest variance (standard error) if the correlation model is correctly specified

Comments

- GEE is specified by a mean model and a correlation model
    1. A regression model for the average outcome, e.g. linear, logistic
    2. A model for longitudinal correlation, e.g. independence, exchangeable
- GEE also computes an empirical variance estimator (aka sandwich, robust, or Huber-White variance estimator)
- Empirical variance estimator provides valid standard errors for $\hat{\beta}$ even if the correlation model is incorrect, but requires $n \geq 40$ and $n \gg m$

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 structure?

Answer: Selecting a non-independence or weighted correlation structure

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

Variance estimators

- **Independence estimating equation**: An estimation equation with a working independence correlation structure
  - Model-based standard errors are generally not valid
  - Empirical standard errors are valid given large n and $n \gg m$
- **Weighted estimation equation**: An estimation equation with a non-independence working correlation structure
  - Model-based standard errors are valid if correlation model is correct
  - Empirical standard errors are valid given large n and $n \gg m$

	Variance estimator	
Estimating equation	Model-based	Empirical
Independence	$-$	$+/-$
Weighted	$-/+$	$+$

Inference for β: Wald test

Consider testing linear hypotheses of the form

$$H\colon Q\beta = 0$$

where Q a matrix of full rank with $\dim(Q) = r \times p$ and $r < p$

- Obtain $\hat{\beta}$ and $\text{Cov}[\hat{\beta}]$; under the null hypothesis

$$\sqrt{n}\, Q\hat{\beta} \sim N_r(0,\ Q\text{Cov}[\hat{\beta}]Q^{\mathsf{T}})$$

- Testing may proceed using a multivariable Wald statistic

$$n\,(Q\hat{\beta})^{\mathsf{T}}(Q\text{Cov}[\hat{\beta}]Q^{\mathsf{T}})^{-1}Q\hat{\beta} \sim \chi^2_r$$

- Requires computation under the alternative hypothesis

NB: Likelihood ratio test not available; not relied on a likelihood function

Dental growth

Characterize dental growth among males and females, ages 8 to 14 years

$$\mathrm{E}[Y_{ij}] = \beta_0 + \beta_1(\mathrm{Age}_{ij} - 8) + \beta_2\mathrm{Gender}_i + \beta_3(\mathrm{Age}_{ij} - 8) \cdot \mathrm{Gender}_i$$

- Consider various specifications for the 'working' correlation structure
  - Independence
  - Exchangeable
  - Auto-regressive
  - Unstructured

NB: In practice, selection of a working correlation structure should be guided by a priori knowledge and/or exploratory analysis

Dental growth: R

- Use the `geeglm` command in the `geepack` library

```
library(geepack)
?geeglm

m_ind <- geeglm(length ~ I(age-8)*gender, id=id,
                corstr="independence", data=growth)
m_exc <- geeglm(length ~ I(age-8)*gender, id=id,
                corstr="exchangeable", data=growth)
m_ar1 <- geeglm(length ~ I(age-8)*gender, id=id,
                corstr="ar1", data=growth)
m_uns <- geeglm(length ~ I(age-8)*gender, id=id,
                corstr="unstructured", data=growth)

m_ols <- lm(length ~ I(age-8)*gender, data=growth)
```

Dental growth: R

```
geeglm(formula = length ~ I(age - 8) * gender, data = growth,
    id = id, corstr = "independence")

 Coefficients:
                      Estimate Std.err    Wald Pr(>|W|)
(Intercept)            21.2091  0.5604 1432.19  < 2e-16 ***
I(age - 8)              0.4795  0.0631   57.70  3.1e-14 ***
gendermale              1.4909  0.7940    3.53   0.0604 .
I(age - 8):gendermale   0.3205  0.1214    6.97   0.0083 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     4.91    1.02

Correlation: Structure = independence
Number of clusters:   26   Maximum cluster size: 4
```

Dental growth: R

```
geeglm(formula = length ~ I(age - 8) * gender, data = growth,
    id = id, corstr = "exchangeable")

 Coefficients:
                      Estimate Std.err    Wald Pr(>|W|)
(Intercept)            21.2091  0.5604 1432.19  < 2e-16 ***
I(age - 8)              0.4795  0.0631   57.70  3.1e-14 ***
gendermale              1.4909  0.7940    3.53   0.0604 .
I(age - 8):gendermale   0.3205  0.1214    6.97   0.0083 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     4.91    1.02

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha     0.61   0.134
Number of clusters:   26   Maximum cluster size: 4
```

Dental growth

	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)
Independence	21.2 (0.56)	0.48 (0.06)	1.49 (0.79)	0.32 (0.12)
Exchangeable	21.2 (0.56)	0.48 (0.06)	1.49 (0.79)	0.32 (0.12)
Auto-regressive	21.2 (0.59)	0.48 (0.06)	1.67 (0.85)	0.30 (0.13)
Unstructured	21.2 (0.56)	0.48 (0.06)	1.50 (0.78)	0.32 (0.12)
OLS	21.2 (0.57)	0.48 (0.15)	1.49 (0.75)	0.32 (0.20)

- Independence and OLS point estimates are identical
  - Independence estimating equation is identical to the score equation
- OLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_3$ are too big
  - Age is within-subject or time-dependent
- Independence and exchangeable provide identical results
  - Data are balanced and complete
- Unstructured provides similar results
- Auto-regressive provides different results

Dental growth

$$\text{Exchangeable} \quad : \quad \begin{bmatrix} 1 & & & \\ 0.61 & 1 & & \\ 0.61 & 0.61 & 1 & \\ 0.61 & 0.61 & 0.61 & 1 \end{bmatrix}$$

$$\text{Auto-regressive} \quad : \quad \begin{bmatrix} 1 & & & \\ 0.75 & 1 & & \\ 0.56 & 0.75 & 1 & \\ 0.42 & 0.56 & 0.75 & 1 \end{bmatrix}$$

$$\text{Unstructured} \quad : \quad \begin{bmatrix} 1 & & & \\ 0.51 & 1 & & \\ 0.75 & 0.53 & 1 & \\ 0.52 & 0.60 & 0.76 & 1 \end{bmatrix}$$

Dental growth: Stata

```
* Declare the dataset to be "panel" data, grouped by id
* with time variable age
xtset id age

* Generate a new variable for centered age
gen cage = age-8

* Fit models with an exchangeable correlation structure
help xtgee
xi: xtgee length i.gender*cage, corr(exch) robust
lincom cage + _IgenXcage_2

* Examine working correlation structure
estat wcorr
```

Dental growth: Stata

```
GEE population-averaged model                   Number of obs      =       104
Group variable:                         id      Number of groups   =        26
Link:                             identity      Obs per group: min =         4
Family:                           Gaussian                     avg =       4.0
Correlation:                   independent                     max =         4
                                                Wald chi2(3)       =    148.85
Scale parameter:                  4.909594      Prob > chi2        =    0.0000

Pearson chi2(104):                  510.60      Deviance           =    510.60
Dispersion (Pearson):             4.909594      Dispersion         =  4.909594

                                     (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
      length |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Igender_2 |   1.490909   .8096977     1.84   0.066    -.0960691    3.077887
        cage |   .4795455   .0643829     7.45   0.000     .3533573    .6057336
_IgenXcage_2 |   .3204545   .1237715     2.59   0.010     .0778669    .5630422
       _cons |   21.20909   .5715302    37.11   0.000     20.08891    22.32927
------------------------------------------------------------------------------
```

Dental growth: Stata

```
. lincom cage + _IgenXcage_2

 ( 1)  cage + _IgenXcage_2 = 0

------------------------------------------------------------------------------
      length |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |         .8   .1057082     7.57   0.000     .5928157    1.007184
------------------------------------------------------------------------------


. estat wcorr

Estimated within-id correlation matrix R:

      |        c1         c2         c3         c4
------+--------------------------------------------
   r1 |         1
   r2 |  .6103379          1
   r3 |  .6103379   .6103379          1
   r4 |  .6103379   .6103379   .6103379          1
```

Summary

- In the GEE approach the 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
- Semi-parametric: Only the mean and correlation models are specified
- 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

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

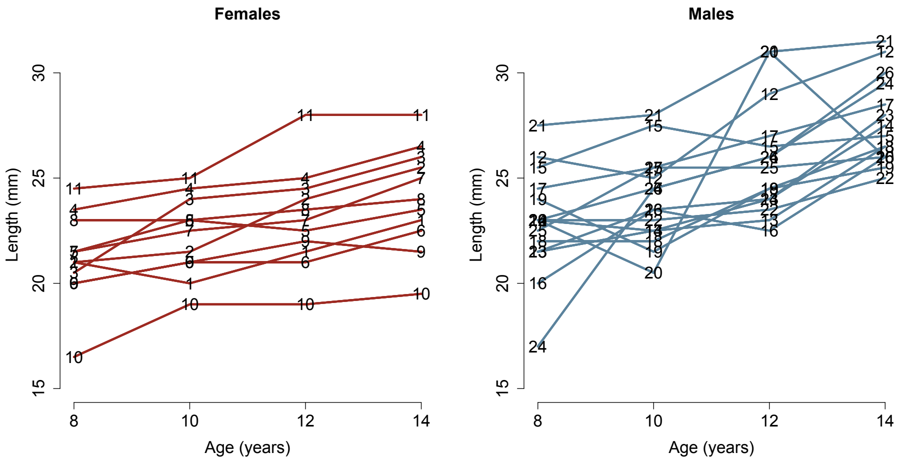
Summary and resources

Dental growth

Goal: Characterize dental growth among children, ages 8 to 14 years

1. Estimate the average growth curve among all children
2. Estimate the growth curve for individual children
3. Characterize the degree of heterogeneity across children
4. Identify factors that predict growth

Dental growth



Mixed-effects models (Laird and Ware, 1982)

4780 citations as of June 2017

- Contrast outcomes both within and between **individuals**
  - Assume that each subject has a regression model characterized by subject-specific parameters: a combination of **fixed-effects** parameters common to all individuals in the population and **random-effects** parameters unique to each individual subject
  - Although covariates allow for differences across subjects, typically cannot measure all factors that give rise to subject-specific variation
  - Subject-specific random effects induce a correlation structure

Set-up

For subject i the mixed-effects model is characterized by

$$\begin{aligned}
Y_i &= (Y_{i1}, Y_{i2}, \ldots, Y_{im_i})^\mathsf{T} \\
\\
\beta^\star &= (\beta_1^\star, \beta_2^\star, \ldots, \beta_p^\star)^\mathsf{T} && \text{Fixed effects} \\
x_{ij} &= (x_{ij1}, x_{ij2}, \ldots, x_{ijp}) \\
X_i &= (x_{i1}, x_{i2}, \ldots, x_{im_i})^\mathsf{T} && \text{Design matrix for fixed effects} \\
\\
\gamma_i &= (\gamma_{1i}, \gamma_{2i}, \ldots, \gamma_{qi})^\mathsf{T} && \text{Random effects} \\
z_{ij} &= (z_{ij1}, z_{ij2}, \ldots, z_{ijq}) \\
Z_i &= (z_{i1}, z_{i2}, \ldots, z_{im_i})^\mathsf{T} && \text{Design matrix for random effects}
\end{aligned}$$

for $i = 1, \ldots, n$; $j = 1, \ldots, m_i$; and $k = 1, \ldots, p$ with $q \leq p$

Linear mixed-effects model

Consider a linear mixed-effects model for a continuous outcome Y_{ij}

- **Stage 1**: Model for response given random effects

$$Y_{ij} = x_{ij}\beta + z_{ij}\gamma_i + \epsilon_{ij}$$

  where

  - x_{ij} is a vector a covariates
  - z_{ij} is a subset of x_{ij}
  - β is a vector of fixed-effects parameters
  - γ_i is a vector of random-effects parameters
  - ϵ_{ij} is observation-specific measurement error

- **Stage 2**: Model for random effects

$$\begin{aligned} \gamma_i &\sim N(0, G) \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

  where γ_i and ϵ_{ij} are assumed to be independent

Choices for random effects

Consider the linear mixed-effects models that include

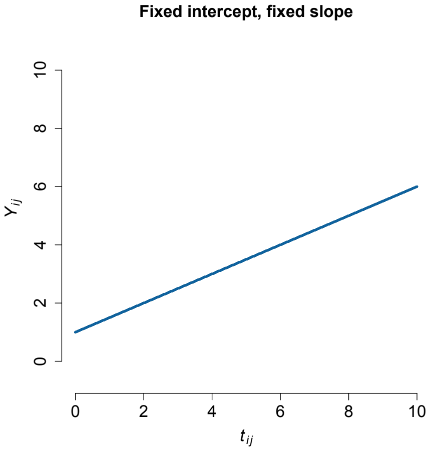
- **Random intercepts**

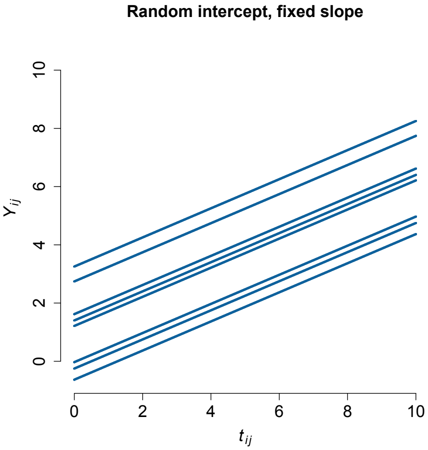
$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \gamma_{0i} + \epsilon_{ij} \\ &= (\beta_0 + \gamma_{0i}) + \beta_1 t_{ij} + \epsilon_{ij} \end{aligned}$$

- **Random intercepts and slopes**

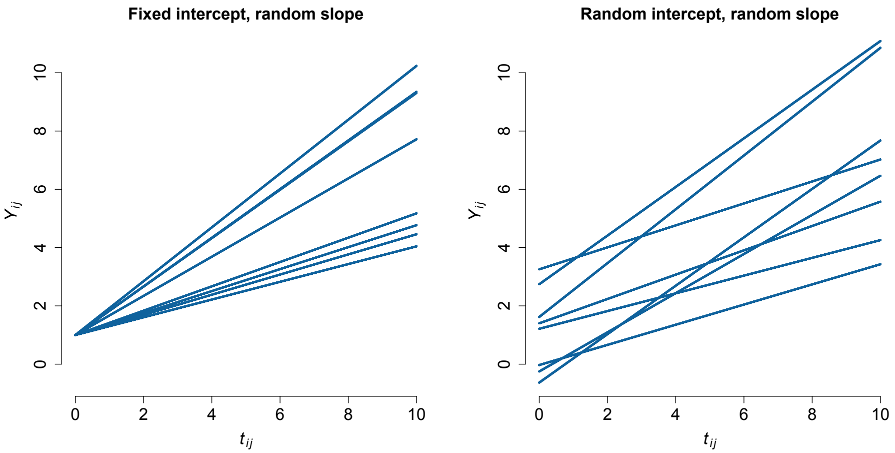
$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \gamma_{0i} + \gamma_{1i} t_{ij} + \epsilon_{ij} \\ &= (\beta_0 + \gamma_{0i}) + (\beta_1 + \gamma_{1i}) t_{ij} + \epsilon_{ij} \end{aligned}$$

Choices for random effects


Fixed intercept, fixed slope


Random intercept, fixed slope

Choices for random effects



Fixed intercept, random slope

Random intercept, random slope

Choices for random effects: G

G quantifies random variation in trajectories across subjects

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$$

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

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

Basic models: Correlation

What is the correlation between measurements on the same subject?

- Random intercepts model
  - Assuming $\text{Var}[\epsilon_{ij}] = \sigma^2$ and $\text{Cov}[\epsilon_{ij}, \epsilon_{ij'}] = 0$

$$
\begin{aligned}
Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \gamma_{0i} + \epsilon_{ij} \\
Y_{ij'} &= \beta_0 + \beta_1 t_{ij'} + \gamma_{0i} + \epsilon_{ij'}
\end{aligned}
$$

$$
\begin{aligned}
\text{Var}[Y_{ij}] &= \text{Var}_\gamma[\text{E}_Y(Y_{ij} \mid \gamma_{0i})] + \text{E}_\gamma[\text{Var}_Y(Y_{ij} \mid \gamma_{0i})] \\
&= G_{11} + \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\text{Cov}[Y_{ij},\ Y_{ij'}] &= \text{Cov}_\gamma[\text{E}_Y(Y_{ij} \mid \gamma_{0i}),\ \text{E}_Y(Y_{ij'} \mid \gamma_{0i})] \\
&\quad + \text{E}_\gamma[\text{Cov}_Y(Y_{ij},\ Y_{ij'} \mid \gamma_{0i})] \\
&= G_{11}
\end{aligned}
$$

Basic models: Correlation

- Random intercepts model (continued)

$$\begin{aligned}\text{Corr}[Y_{ij}, Y_{ij'}] &= \frac{G_{11}}{\sqrt{G_{11}+\sigma^2}\sqrt{G_{11}+\sigma^2}} \\ &= \frac{G_{11}}{G_{11}+\sigma^2} \\ &= \frac{\text{`Between'}}{\text{`Between'} + \text{`Within'}} \\ &\geq 0 \text{ (and } \leq 1\text{)}\end{aligned}$$

  - Any two measurements on the same subject have the same correlation; does not depend on time nor the distance between measurements
  - Equivalent to an exchangeable correlation structure
  - Longitudinal correlation is constrained to be positive ($G_{11} \geq 0$, $\sigma^2 \geq 0$)

Basic models: Correlation

- Random intercepts and slopes model
  - Assuming $\text{Var}[\epsilon_{ij}] = \sigma^2$ and $\text{Cov}[\epsilon_{ij}, \epsilon_{ij'}] = 0$

$$\begin{aligned} Y_{ij} &= (\beta_0 + \beta_1 t_{ij}) + (\gamma_{0i} + \gamma_{1i} t_{ij}) + \epsilon_{ij} \\ Y_{ij'} &= (\beta_0 + \beta_1 t_{ij'}) + (\gamma_{0i} + \gamma_{1i} t_{ij'}) + \epsilon_{ij'} \end{aligned}$$

$$\begin{aligned} \text{Var}[Y_{ij}] &= \text{Var}_\gamma[\text{E}_Y(Y_{ij} \mid \gamma_i)] + \text{E}_\gamma[\text{Var}_Y(Y_{ij} \mid \gamma_i)] \\ &= G_{11} + 2G_{12} t_{ij} + G_{22} t_{ij}^2 + \sigma^2 \end{aligned}$$

$$\begin{aligned} \text{Cov}[Y_{ij}, Y_{ij'}] &= \text{Cov}_\gamma[\text{E}_Y(Y_{ij} \mid \gamma_i), \text{E}_Y(Y_{ij'} \mid \gamma_i)] \\ &\quad + \text{E}_\gamma[\text{Cov}_Y(Y_{ij},\ Y_{ij'} \mid \gamma_i)] \\ &= G_{11} + G_{12}(t_{ij} + t_{ij'}) + G_{22} t_{ij} t_{ij'} \end{aligned}$$

Basic models: Correlation

- Random intercepts and slopes model (continued)

$$
\begin{aligned}
&\mathrm{Corr}[Y_{ij}, Y_{ij'}] \\
&= \frac{G_{11} + G_{12}(t_{ij} + t_{ij'}) + G_{22} t_{ij} t_{ij'}}{\sqrt{G_{11} + 2G_{12} t_{ij} + G_{22} t_{ij}^2 + \sigma^2}\sqrt{G_{11} + 2G_{12} t_{ij'} + G_{22} t_{ij'}^2 + \sigma^2}} \\
&\equiv \rho_{ijj'}
\end{aligned}
$$

  - Any two measurements on the same subject may not have the same correlation; depends on the specific observation times

Generalized linear mixed-effects models

A GLMM is defined by **random** and **systematic** components

- **Random**: Conditional on γ_i the outcomes $Y_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathsf{T}}$ are mutually independent and have an exponential family density

  $$f(Y_{ij} \mid \beta^{\star}, \gamma_i, \phi) = \exp\{[Y_{ij}\theta_{ij} - \psi(\theta_{ij})]/\phi + c(Y_{ij}, \phi)\}$$

  for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$ with a scale parameter $\phi > 0$ and $\theta_{ij} \equiv \theta_{ij}(\beta^{\star}, \gamma_i)$

Generalized linear mixed-effects models

A GLMM is defined by **random** and **systematic** components

- **Systematic**: $\mu^\star_{ij}$ is modeled via a linear predictor containing fixed regression parameters $\beta^\star$ common to all individuals in the population and subject-specific random effects γ_i with a known link function $g(\cdot)$

  $$g(\mu^\star_{ij}) = x_{ij}\beta^\star + z_{ij}\gamma_i \;\Leftrightarrow\; \mu^\star_{ij} = g^{-1}(x_{ij}\beta^\star + z_{ij}\gamma_i)$$

  where the random effects γ_i are mutually independent with a common underlying multivariate distribution, typically assumed to be

  $$\gamma_i \sim N_q(0,\ G)$$

  so that G quantifies random variation across subjects

Likelihood-based estimation of β

Requires specification of a complete probability distribution for the data

- Likelihood-based methods are designed for fixed effects, so integrate over the assumed distribution for the random effects

$$\mathcal{L}_Y(\beta, \sigma, G) = \prod_{i=1}^{n} \int f_{Y|\gamma}(Y_i \mid \gamma_i, \beta, \sigma) \times f_\gamma(\gamma_i \mid G) d\gamma_i$$

  where f_γ is typically the density function of a Normal random variable
- For linear models the required integration is straightforward because Y_i and γ_i are both normally distributed (easy to program)
- For non-linear models the integration is difficult and requires either approximation or numerical techniques (hard to program)

Likelihood-based estimation of β

Two likelihood-based approaches to estimation using a GLMM

1. **Conditional likelihood**: Treat the random effects as if they were fixed parameters and **eliminate** them by conditioning on their sufficient statistics; does not require a specified distribution for γ_i
   - `xtreg` and `xtlogit` with `fe` option in Stata
2. **Maximum likelihood**: Treat the random effects as unobserved nuisance variables and **integrate** over their assumed distribution to obtain the marginal likelihood for β; typically assume $\gamma_i \sim N(0, G)$
   - `xtreg` and `xtlogit` with `re` option in Stata
   - `mixed` and `melogit` in Stata
   - `lmer` and `glmer` in R package `lme4`

‘Fixed effects’ versus ‘random effects’

‘Fixed-effects’ approach provided by conditional likelihood estimation

- Comparisons are made within individuals who act as their own control and differences are averaged across all individuals in the sample
- May eliminate potentially large sources of bias by controlling for all stable characteristics of the individuals under study $(+)$
- Variation across subjects is ignored, which may provide standard error estimates that are too big; conservative inference $(-)$
- Although controlled for by conditioning, cannot estimate coefficients for covariates that have no within-subject variation $(-/+)$

‘Fixed effects’ versus ‘random effects’

‘Random-effects’ approach provided by maximum likelihood estimation

- Comparisons are based on within- and between-subject contrasts
- Requires a specified distribution for subject-specific effects; correct specification is required for valid likelihood-based inference $(-/+)$
- Do not control for unmeasured characteristics because random effects are almost always assumed to be uncorrelated with covariates $(-)$
- Can estimate effects of within- and between-subject covariates $(+)$

Inference for β

Consider testing fixed effects in nested linear mixed-effects models

$$H\colon\ \beta = \left[\begin{array}{c} \beta_1 \\ 0 \end{array}\right] \quad \text{versus} \quad K\colon\ \beta = \left[\begin{array}{c} \beta_1 \\ \beta_2 \end{array}\right],$$

i.e., $H\colon\ \beta_2 = 0$

- Likelihood ratio test is valid if ML estimation is used
- Likelihood ratio test may not be valid with other estimation methods
- Wald test is generally valid

Inference for G

Consider testing whether a random intercept model is adequate

$$H\colon\ G = \begin{bmatrix} G_{11} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{versus} \quad K\colon\ G = \begin{bmatrix} G_{11} & \\ G_{12} & G_{22} \end{bmatrix},$$

i.e., $H\colon\ G_{12} = G_{22} = 0$

- Adequate covariance modeling is useful for the interpretation of the random variation in the data
- Over-parameterization of the covariance structure leads to inefficient estimation of fixed effects parameters β
- Covariance model choice determines the standard error estimates for $\hat{\beta}$; correct model is required for correct standard error estimates

Inference for G

- $G_{22} = 0$ is on the boundary of the parameter space
  - Violates the standard assumption used to establish the typical χ^2 distribution of the likelihood ratio test statistic
  - Null hypothesis is accepted too often, leading to an incorrect simplification of the covariance structure of the data

  (see Stata output for dental growth example)

- Correct distribution of test statistic is a mixture of χ^2 distributions
  - Example: Consider testing H: $G_{11} = 0$
  - Correct distribution is a mixture of χ^2_1 and χ^2_0, each with weight 0.5
  - χ^2_0 gives probability mass 1 to the value 0

- Generally recommend against this inferential procedure
  - Specification for the covariance structure should be guided by *a priori* scientific knowledge and exploratory data analysis

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 $x_{ij}\beta$ (here also $z_{ij}\gamma_i$)
    - 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 $\hat{\beta}$
- **Normality**: Normality of ϵ_{ij} and γ_i is required for normal likelihood function to be the correct likelihood function for Y_{ij}
- n sufficiently large for **asymptotic inference** to be valid

⋆ These assumptions must be verified to evaluate any fitted model

Dental growth

Characterize dental growth among males and females, ages 8 to 14 years

$$\mathrm{E}[Y_{ij}] = \beta_0 + \beta_1(\mathrm{Age}_{ij} - 8) + \beta_2 \mathrm{Gender}_i + \beta_3(\mathrm{Age}_{ij} - 8) \cdot \mathrm{Gender}_i$$

- Consider various specifications for the random effects structure
  - Random intercepts
  - Random intercepts and slopes (for age)

  NB: In practice, selection of a random effects structure
  should be guided by a priori knowledge and/or exploratory analysis,
  or specified as relevant to the scientific question of interest

Dental growth: R

- Use the lmer command in the lme4 library

```
library(lme4)
?lmer

m_ri <- lmer(length ~ (1 | id) + I(age-8)*gender, data=growth)

m_rs <- lmer(length ~ (I(age-8) | id) + I(age-8)*gender, data=growth)
```

Dental growth: R

```
> summary(m_ri)

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 3.27     1.81
 Residual             1.96     1.40
Number of obs: 104, groups: id, 26

Fixed effects:
                      Estimate Std. Error t value
(Intercept)            21.2091     0.6500    32.6
I(age - 8)              0.4795     0.0945     5.1
gendermale              1.4909     0.8558     1.7
I(age - 8):gendermale   0.3205     0.1244     2.6
```

Dental growth: R

```
> summary(m_rs)

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 3.3209   1.822
          I(age - 8)  0.0331   0.182    -0.15
 Residual             1.7543   1.325
Number of obs: 104, groups: id, 26

Fixed effects:
                      Estimate Std. Error t value
(Intercept)             21.209      0.643    33.0
I(age - 8)               0.480      0.105     4.6
gendermale               1.491      0.847     1.8
I(age - 8):gendermale    0.320      0.138     2.3
```

Dental growth: R

```
> anova(m_ri, m_rs)
refitting model(s) with ML (instead of REML)
Data: growth
Models:
m_ri: length ~ (1 | id) + I(age - 8) * gender
m_rs: length ~ (I(age - 8) | id) + I(age - 8) * gender
     Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
m_ri  6 426 442   -207      414
m_rs  8 430 451   -207      414  0.66      2       0.72
```

Dental growth

- $\hat{G}_{12} < 0$ indicates subjects with high length have low rate of growth
- $\hat{G}_{11}$ indicates mild variability in level of dental length
- $\hat{G}_{22}$ indicates mild variability in change in length over time
- AIC and LR indicate model 1 is a reasonable fit to the data

$$\mathrm{Corr}[Y_{ij}, Y_{ij'}] = \frac{1.73^2}{1.73^2 + 1.38^2} = 0.61$$

  - Consistent with exploratory and GEE analyses that indicated exchangeable correlation structure is adequate

- $\hat{\beta}_3$ indicates increase in average dental length is larger for males
- Reject the null hypothesis that $\beta_3 = 0$ with $p = 0.0125$

Dental growth: Stata

```
* Declare the dataset to be "panel" data, grouped by id
* with time variable age
xtset id age

* Fit models with random intercepts and slopes
help mixed
gen cage = age-8
mixed length i.gender##c.cage || id:, stddeviations
est store ri
estat ic

mixed length i.gender##c.cage || id: cage, ///
cov(unstructured) stddeviations
est store rs
estat ic

* Use likelihood ratio test and AIC to compare models
lrtest ri rs
```

Dental growth: Stata

```
Mixed-effects ML regression                     Number of obs      =       104
Group variable: id                              Number of groups   =        26

                                                Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4


                                                Wald chi2(3)       =    137.79
Log likelihood = -207.08327                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      length |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
       male  |   1.490909   .8265567     1.80   0.071    -.1291124    3.110931
        cage |   .4795455   .0932514     5.14   0.000      .296776    .6623149
             |
gender#c.cage |
       male  |   .3204545   .1227712     2.61   0.009     .0798274    .5610817
             |
       _cons |   21.20909   .6278149    33.78   0.000      19.9786    22.43959
------------------------------------------------------------------------------
```

Dental growth: Stata

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Identity                 |
                   sd(_cons) |   1.731043   .2792446      1.261815    2.374762
-----------------------------+------------------------------------------------
                sd(Residual) |   1.383142     .11074      1.182269    1.618146
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    46.46 Prob >= chibar2 = 0.0000


Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
          ri |    104           .   -207.0833      6     426.1665    442.0329
-----------------------------------------------------------------------------
```

Dental growth: Stata

```
Mixed-effects ML regression                     Number of obs      =       104
Group variable: id                              Number of groups   =        26

                                                Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4


                                                Wald chi2(3)       =    118.63
Log likelihood = -206.75403                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      length |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
       male  |   1.490909   .8134256     1.83   0.067    -.1033757    3.085194
        cage |   .4795455   .1006929     4.76   0.000      .282191    .6768999
             |
gender#c.cage |
       male  |   .3204545   .1325684     2.42   0.016     .0606253    .5802838
             |
       _cons |   21.20909   .6178411    34.33   0.000     19.99814    22.42004
------------------------------------------------------------------------------
```

Dental growth: Stata

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                    sd(cage) |   .1543156   .1146815      .0359608    .6622021
                   sd(_cons) |   1.723651   .3449757      1.164362     2.55159
            corr(cage,_cons) |  -.0934221   .5302289     -.8151116    .7418963
-----------------------------+------------------------------------------------
                sd(Residual) |    1.32451   .1298788       1.09292    1.605175
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =    47.12   Prob > chi2 = 0.0000


Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)      df          AIC         BIC
-------------+---------------------------------------------------------------
          rs |    104           .    -206.754       8     429.5081    450.6632
-----------------------------------------------------------------------------
```

Dental growth: Stata

```
. lrtest ri rs

Likelihood-ratio test                                 LR chi2(2)  =      0.66
(Assumption: ri nested in rs)                         Prob > chi2 =    0.7195

Note: The reported degrees of freedom assumes the null hypothesis is not on the
boundary of the parameter space.  If this is not true, then the reported test is
conservative.
```

Summary

- Mixed-effects models assume that each subject has a regression model characterized by subject-specific parameters; a combination of fixed effects parameters common to all individuals in the population and random subject-specific perturbations
- Likelihood-based estimation and inference requires a complete parametric probability distribution for subject-specific random effects and error terms that must be verified for valid inference
- Estimates for the random effects are available (a.k.a. prediction), e.g., provider profiling
- See help files for specification of hierarchical random effects

Issues

- Interpretation depends on outcomes and random-effects specification
- GLMM requires that any missing data are missing at random
- Issues arise with time-dependent exposures and covariance weighting

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

Summary and resources

Conditional and marginal effects

- Parameter estimates obtained from a **marginal** model (as obtained via a GEE) estimate **population-averaged** contrasts
- Parameter estimates obtained from a **conditional** model (as obtained via a GLMM) estimate **subject-specific** contrasts
- In a linear model for a Gaussian outcome with an identity link these contrasts are equivalent; not the case with non-linear models
  - Depends on the outcome distribution
  - Depends on the specified random effects

Conditional and marginal effects

Parameters in the LMM may be interpreted as population-level contrasts

- **Random intercepts**

$$
\begin{aligned}
&\mathrm{E}[Y_{ij} \mid t_{ij} = t+1] - \mathrm{E}[Y_{ij} \mid t_{ij} = t] \\
&\quad = \mathrm{E}_\gamma[\mathrm{E}_Y(Y_{ij} \mid t_{ij} = t+1,\, \gamma_{0i})] - \mathrm{E}_\gamma[\mathrm{E}_Y(Y_{ij} \mid t_{ij} = t,\, \gamma_{0i})] \\
&\quad = \mathrm{E}_\gamma[\beta_0 + \beta_1(t+1) + \gamma_{0i}] - \mathrm{E}_\gamma[\beta_0 + \beta_1 t + \gamma_{0i}] \\
&\quad = \beta_1
\end{aligned}
$$

- **Random intercepts and slopes**

$$
\begin{aligned}
&\mathrm{E}[Y_{ij} \mid t_{ij} = t+1] - \mathrm{E}[Y_{ij} \mid t_{ij} = t] \\
&\quad = \mathrm{E}_\gamma[\mathrm{E}_Y(Y_{ij} \mid t_{ij} = t+1,\, \gamma_{0i},\, \gamma_{1i})] - \mathrm{E}_\gamma[\mathrm{E}_Y(Y_{ij} \mid t_{ij} = t,\, \gamma_{0i},\, \gamma_{1i})] \\
&\quad = \mathrm{E}_\gamma[\beta_0 + \beta_1(t+1) + \gamma_{0i} + \gamma_{1i}(t+1)] - \mathrm{E}_\gamma[\beta_0 + \beta_1 t + \gamma_{0i} + \gamma_{1i} t] \\
&\quad = \beta_1
\end{aligned}
$$

Conditional and marginal effects

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

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

Conditional and marginal effects: Example

Consider a logistic regression model with **subject-specific** intercepts

$$\text{logit}(\text{P}[Y_{ij} = 1 \mid \gamma_{0i}]) = \beta_0^\star + \beta_1^\star x_{ij} + \gamma_{0i}$$

where each subject has their own baseline risk of the disease ($Y_{ij} = 1$)

$$\frac{\exp(\beta_0^\star + \gamma_{0i})}{1 + \exp(\beta_0^\star + \gamma_{0i})}$$

which is multiplied by $\exp(\beta_1^\star)$ if the subject becomes exposed ($x_{ij} = 1$)

Conditional and marginal effects: Example

The **population** rate of infection is the average risk across individuals

$$\begin{aligned} \mathrm{P}[Y_{ij} = 1] &= \int \mathrm{P}[Y_{ij} = 1 \mid \gamma_{0i}]\, dF(\gamma_{0i}) \\ &= \int \frac{\exp(\beta_0^\star + \beta_1^\star x_{ij} + \gamma_{0i})}{1 + \exp(\beta_0^\star + \beta_1^\star x_{ij} + \gamma_{0i})} f(\gamma_{0i} \mid \tau)\, d\gamma_{0i} \end{aligned}$$

where typically $\gamma_{0i} \sim N(0, \tau^2)$

- Assuming $[\beta_0^\star, \beta_1^\star] = [-2, 0.4]$ and $\tau^2 = 2$ the **population** rates are

$$\begin{aligned} \mathrm{P}[Y_{ij} = 1 \mid x_{ij} = 0] &= 0.18 \\ \mathrm{P}[Y_{ij} = 1 \mid x_{ij} = 1] &= 0.23 \end{aligned}$$

where the odds ratio associated with exposure is $\exp(0.4) = 1.5$

Conditional and marginal effects: Example

A **marginal** model ignores heterogeneity among individuals and considers the **population-averaged** rate rather than the **conditional** rate

$$\text{logit}(\text{P}[Y_{ij} = 1]) = \beta_0 + \beta_1 x_{ij}$$

where the infection rate among a **population** of unexposed individuals is

$$\text{P}[Y_{ij} = 1 \mid x_{ij} = 0] = 0.18$$

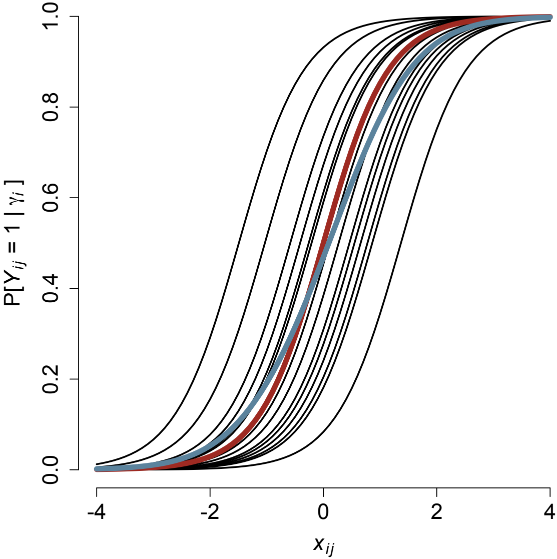
and the **population-averaged** odds ratio associated with exposure is

$$\frac{\text{P}[Y_{ij} = 1 \mid x_{ij} = 1]/(1 - \text{P}[Y_{ij} = 1 \mid x_{ij} = 1])}{\text{P}[Y_{ij} = 1 \mid x_{ij} = 0]/(1 - \text{P}[Y_{ij} = 1 \mid x_{ij} = 0])} = 1.36$$

so that $[\beta_0, \beta_1] = [\text{logit}(0.18), \log(1.36)] = [-1.23, 0.31]$

- $\star$ **Marginal** parameters are “attenuated” w.r.t. **conditional** parameters

Conditional and marginal effects



Conditional and marginal effects

After "Will the real subject-specific odds ratio please stand up?" by Thomas Lumley

Suppose we are evaluating an anti-smoking intervention and observe

$$\begin{aligned} Y_i &= \text{Indicator whether subject } i \text{ smoked during the past week} \\ x_i &= \text{Indicator whether subject } i \text{ received the intervention} \end{aligned}$$

for $i = 1, \ldots, n$

- Logistic regression model is given by

$$\text{logit}(\text{E}[Y_i]) = \beta_0 + \beta_1 x_i$$

- Effect of the intervention is measured by the odds ratio $\exp(\beta_1)$

Conditional and marginal effects

After "Will the real subject-specific odds ratio please stand up?" by Thomas Lumley

I forgot to tell you that each person is evaluated three times so that

$$\begin{aligned} \text{logit}(\text{E}[Y_{ij}]) &= \beta_0 + \beta_1 x_{ij} \\ \text{logit}(\text{E}[Y_{ij} \mid \gamma_i]) &= \beta_0^\star + \beta_1^\star x_{ij} + \gamma_i \end{aligned}$$

where γ_i quantifies variation across subjects

- First is a marginal model; second is a conditional model
- $\exp(\beta_1^\star)$ is the subject-specific odds ratio measuring intervention effect
- $\beta_1^\star$ measures actual intervention effect and β_1 has been attenuated

Conditional and marginal effects

After “Will the real subject-specific odds ratio please stand up?” by Thomas Lumley

I also forgot to tell you that this is group-discussion intervention so that

$$\begin{aligned}\text{logit}(\mathrm{E}[Y_{gij}]) &= \beta_0 + \beta_1 x_{gij} \\ \text{logit}(\mathrm{E}[Y_{gij} \mid \gamma_i, \gamma_g]) &= \beta_0^{\star\star} + \beta_1^{\star\star} x_{gij} + \gamma_i + \gamma_g\end{aligned}$$

where γ_g quantifies variation across groups

- $\exp(\beta_1^{\star\star})$ is the real subject-specific odds ratio
- $\exp(\beta_1^{\star})$ is an attenuated version; it is the group-specific odds ratio

Conditional and marginal effects

After "Will the real subject-specific odds ratio please stand up?" by Thomas Lumley

I also forgot to tell you that the discussion was facilitated by a physician, where the study was actually randomized by medical practice, so that

$$\begin{aligned} \text{logit}(\mathrm{E}[Y_{pgij}]) &= \beta_0 + \beta_1 x_{pgij} \\ \text{logit}(\mathrm{E}[Y_{pgij} \mid \gamma_i, \gamma_g, \gamma_p]) &= \beta_0^{\star\star\star} + \beta_1^{\star\star\star} x_{pgij} + \gamma_i + \gamma_g + \gamma_p \end{aligned}$$

where γ_p quantifies variation across physicians

- Now the subject-specific odds ratio is really $\exp(\beta_1^{\star\star\star})$
- Marginal odds ratio is still boringly stuck at $\exp(\beta_1)$

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

Summary and resources

Missing data

- Missing values arise in longitudinal studies whenever the intended serial observations collected on a subject over time are incomplete
- 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	□	□

Strategies

1. **Complete-case** analyses based only on complete measurement series
   - Easy to implement; may be valid with small amount of missing data
   - Otherwise may lead to serious bias and loss of efficiency
2. **Imputation**-based procedures to fill-in any missing data
   - Examples: Hot deck, mean, regression, and multiple imputation
   - Allows use of standard estimation methods on resulting complete data
3. **Weighted** procedures to adjust for non-response as if part of design
   - Developed from sample-survey techniques for non-response weighting
   - Example: Weighted generalized estimating equations (WGEE)
4. **Model**-based procedures based on a model for the observed data
   - Examples: Selection, pattern mixture, and random effects models
   - Facilitate evaluation of assumptions underlying the fitted models
5. **Others** that should rarely, if ever, be used
   - Example: Last observation carried forward

Taxonomy (Little and Rubin, 2002)

Partition the complete set of intended observations into the observed and missing data; what factors influence missingness of intended observations?

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

MNAR also referred to as informative or non-ignorable missingness;
thus MAR and MCAR as non-informative or ignorable missingness

Examples and implications

- **MCAR**: Administrative censoring at a fixed calendar time
  - Generalized estimating equations are valid
  - Mixed-effects models are valid
- **MAR**: Individuals with no current weight loss in a weight-loss study
  - Generalized estimating equations are not valid
  - Weighted estimating equations are valid
  - Mixed-effects models are valid
- **MNAR**: Subjects in a prospective study based on disease prognosis
  - Generalized estimating equations are not valid
  - Mixed-effects models are not valid

⋆ MAR and MCAR may be evaluated using the observed data

Implication of MCAR and MAR

Likelihood-based inference based on the observed data is valid

$$\begin{aligned} f(Y^o, M) &= \int f(Y^c, M)\, dY^m \\ &= \int f(Y^c)\, f(M \mid Y^c)\, dY^m \\ &= f(M) \int f(Y^c)\, dY^m \quad \text{or} \quad f(M \mid Y^o) \int f(Y^c)\, dY^m \\ &= f(M) f(Y^o) \quad \text{or} \quad f(M \mid Y^o)\, f(Y^o) \\ &\propto f(Y^o) \end{aligned}$$

although this result relies on assumptions that the

- Likelihood for the observed data is correctly specified (as always)
- Distributions are separately parameterized; otherwise efficiency losses
- Unconditional distribution $f(Y^o)$ represents the target of inference

GEE

Estimating equations based on the observed data are valid under MCAR

$$\mathcal{U}_\beta(\beta,\alpha;Y_i^o,X_i)=\sum_{i=1}^{n}(1-M_i)\mathcal{U}_\beta(\beta,\alpha;Y_i^c,X_i)$$

so that for $\mathrm{E}[\mathcal{U}_\beta(\beta,\alpha;Y_i^o,X_i)]=0$ and hence consistency of $\hat{\beta}$ we obtain

$$\begin{aligned}
&\mathrm{E}_{Y^c,X,M}[(1-M_i)\mathcal{U}_\beta(\beta,\alpha;Y_i^c,X_i)]\\
&\quad= \mathrm{E}_{Y^c,X}\{\mathrm{E}_{M|Y^c,X}[(1-M_i)\mathcal{U}_\beta(\beta,\alpha;Y_i^c,X_i)]\}\\
&\quad= \mathrm{E}_{Y^c,X}\{\mathcal{U}_\beta(\beta,\alpha;Y_i^c,X_i)\mathrm{E}_{M|Y^c,X}[(1-M_i)]\}\\
&\quad= \mathrm{E}_{Y^c,X}\{\mathcal{U}_\beta(\beta,\alpha;Y_i^c,X_i)\mathrm{P}[M_i=0\mid Y_i^c,X_i]\}\\
&\quad= \mathrm{E}_{Y^c,X}\{\mathcal{U}_\beta(\beta,\alpha;Y_i^c,X_i)\mathrm{P}[M_i=0\mid X_i]\}\\
&\quad= \mathrm{E}_{X}\{\mathrm{P}[M_i=0\mid X_i]\mathrm{E}_{Y^c|X}[\mathcal{U}_\beta(\beta,\alpha;Y_i^c,X_i)]\}\\
&\quad= 0
\end{aligned}$$

GEE: Comments

- Under MCAR point estimators and robust standard error estimators are consistent even if the correlation structure is incorrectly specified
- Under MAR point estimators are consistent only if the correlation structure is correctly specified, although the robust standard error estimators may be inconsistent (Kenward and Molenberghs, 1998)
- Requires correct specification for μ and sufficiently large n (as always)
- Weighted estimating equations (WGEE) are valid under MAR

WGEE (Robins et al., 1995)

Extend marginal GEE approach to situations with MAR missing data

- Also known as the inverse probability of censoring weighted GEE
- Provides unbiased inference in longitudinal studies with drop-outs
- Observations (or person-visits) in the estimating function are assigned a weight inversely proportional to their probability of being observed

$$\mathcal{U}_\beta(\beta, \alpha, \theta) = \sum_{i=1}^{n} D_i(\beta)^\mathsf{T} V_i(\beta, \alpha)^{-1} W_i(\theta) [Y_i^c - \mu_i(\beta)]$$

so that the drop-out process is taken into account by specification of an $(m \times m)$ diagonal matrix of visit-specific weights

$$W_i(\theta) = \mathrm{diag}[(1 - M_{i1}) w_{i1}, \ldots, (1 - M_{im}) w_{im}]$$

where $M_{ij} = 0$ if the i^{th} individual's outcome is observed at visit j; hence the weight is w_{ij} for observed visits and 0 for unobserved visits

WGEE: Comments

- Accommodates drop-outs but not intermittent missing data patterns

$$
\begin{aligned}
Y_i^c &= \{Y_i^o,\, Y_i^m\} \\
Y_i^o &= \{Y_{i1}, \ldots, Y_{ik-1}\} \\
Y_i^m &= \{Y_{ik}, \ldots, Y_{im}\}
\end{aligned}
$$

- Valid under MAR even if the correlation model is incorrectly specified, provided the model for the probability of missing outcome is correct
  - As with GEE use of the robust variance estimator in WGEE provides robustness to misspecification of the correlation structure
  - With consistent estimation of weights provided by a correctly specified drop-out model, WGEE does not require a correct specification for the correlation structure to estimate consistently β and its covariance
- As with GEE choice of the working correlation matrix affects efficiency
- Requires correct specification for μ and sufficiently large n (as always)
- Estimation of (β, α) requires either *a priori* knowledge of the weights or estimation of w_{ij} using a correctly specified drop-out model

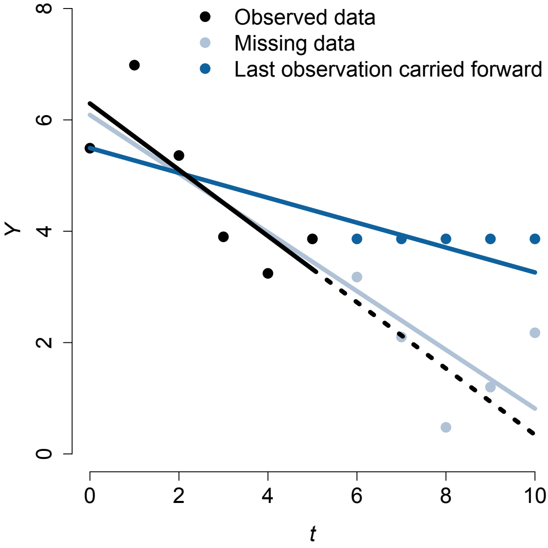
Last observation carried forward

- 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
- 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



Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

Summary and resources

Longitudinal studies

Help establish the causal effect of exposure on outcome by determining the temporal order of exposure and outcome (exposure precedes outcome)

- Cross-sectional study

Egg → Chicken

Chicken → Egg

- Longitudinal study

Bacterium → Dinosaur → Chicken

⋆ There are several other challenges to generating causal inference from longitudinal data, particularly observational longitudinal data

Issues

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 $X_i(t)$ is exogenous with respect to 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 $X_i(t)$, $X_i(t-1)$, $X_i(t-2), \ldots$
- 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

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

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

Pepe and Anderson (1994)

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

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

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

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

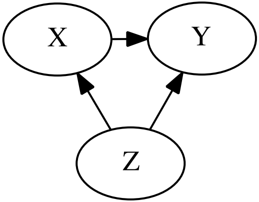
Otherwise an independence estimating equation should be used

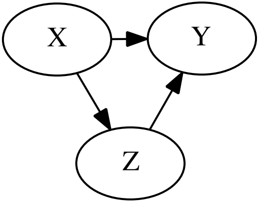
- Known as the **full covariate conditional mean** assumption
- Implies that with time-dependent exposures must assume exogeneity when using a covariance-weighting estimation method
- The **full covariate conditional mean** assumption is often overlooked and should be verified as a crucial element of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

Time-dependent confounders

Traditional epidemiology classifies a variable that is related to both exposure and outcome as either a confounder or intermediary variable

- **Confounder**: A variable Z that is associated with exposure X and outcome Y; if ignored will lead to biased exposure effect estimates
- **Intermediary**: A variable Z that is in the causal pathway between exposure X and outcome Y; should not be controlled for in analysis





⋆ A longitudinal outcome can be both a confounder and an intermediary

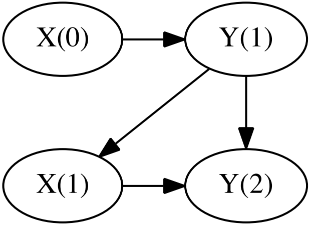
Time-dependent confounders: Example

Consider an observational study of HIV-infected patients in which interest lies in the benefit on CD4+ cell count attributable to AZT treatment

- Current CD4+ count is likely to predict future CD4+ count
- Current CD4+ count may also predict future treatment choices
- Current CD4+ count is the outcome associated with prior treatment, but is also a predictor of and thus a confounder for future treatment
- A regression of current CD4+ count on prior treatment may reveal a lower mean CD4+ count among treated subjects, reflecting the fact that patients who are more sick are more likely to receive treatment

Time-dependent confounders: Example

Feedback: Outcome is a both a confounder and an intermediary



- $Y(1)$ is a confounder for $X(1) \rightarrow Y(2)$
- $Y(1)$ is an intermediary for $X(0) \rightarrow Y(2)$

$\star$ No standard regression methods can be used to generate causal inference

Summary

- Parameter estimates obtained from a marginal model (GEE) estimate population-averaged contrasts; parameter estimates obtained from a conditional model (GLMM) estimate subject-specific contrasts; in some situations these contrasts are equivalent
- Any time-dependent exposures motivate consideration of alternative targets of inference and specific assumptions that must be verified for certain estimation methods to be appropriate
- The presence of missing data determines situations in which certain estimation methods are valid (GEE for MCAR; GLMM for MAR)
- Never use last observation carried forward, even if the FDA says so

Overview

Introduction to longitudinal studies

Longitudinal regression models

Generalized estimating equations

Generalized linear mixed-effects models

Advanced topics

- Conditional and marginal effects
- Missing data
- Time-dependent exposures

Summary and resources

Big picture: GEE

- Marginal mean regression model
- Model for longitudinal correlation
- Semi-parametric model: mean + correlation
- Form an unbiased estimating function
- Estimates obtained as solution to estimating equation
- Model-based or empirical variance estimator
- Robust to correlation model mis-specification
- Large sample: $n \geq 40$
- Testing with Wald tests
- Marginal or population-averaged inference
- Efficiency of non-independence correlation structures
- Missing completely at random (MCAR)
- Time-dependent covariates and endogeneity
- Only one source of positive or negative correlation
- R package `geepack`; Stata command `xtgee`

Big picture: GLMM

- Conditional mean regression model
- Model for population heterogeneity
- Subject-specific random effects induce a correlation structure
- Fully parametric model based on exponential family density
- Estimates obtained from likelihood function
- Conditional (fixed effects) and maximum (random effects) likelihood
- Approximation or numerical integration to integrate out γ
- Requires correct parametric model specification
- Testing with likelihood ratio and Wald tests
- Conditional or subject-specific inference
- Induced marginal mean structure and 'attenuation'
- Missing at random (MAR)
- Time-dependent covariates and endogeneity
- Multiple sources of positive correlation
- R package `lme4`; Stata commands `mixed`, `melogit`

Final summary

Generalized estimating equations

- Provide valid estimates and standard errors for regression parameters of interest even if the correlation model is incorrectly specified $(+)$
- Empirical variance estimator requires sufficiently large sample size $(-)$
- Always provide population-averaged inference regardless of the outcome distribution; ignores subject-level heterogeneity $(+/-)$
- Accommodate only one source of correlation $(-/+)$
- Require that any missing data are missing completely at random $(-)$

Final summary

Generalized linear mixed-effects models

- Provide valid estimates and standard errors for regression parameters only under stringent model assumptions that must be verified $(-)$
- Provide population-averaged or subject-specific inference depending on the outcome distribution and specified random effects $(+/-)$
- Accommodate multiple sources of correlation $(+/-)$
- 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
- Statistical modeling should be informed by exploratory analyses
- Always be mindful of the scientific question(s) of interest

Resources

Introductory

- Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Wiley, 2004.
- Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/ Heirarchical 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.

Thank you!