Module 11
Mixed-effects Models for Longitudinal Data Analysis

Benjamin French, PhD
Department of Biostatistics, Vanderbilt University

SISCER 2021
July 20, 2021

Learning objectives

- This module will overview statistical methods for the analysis of longitudinal data, with a focus on mixed-effects models
- Focus will be on the practical application of appropriate analysis methods, using illustrative examples in R
- Some theoretical background and technical 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

Generalized linear mixed-effects models

Advanced topics

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

Summary

Overview

Introduction to longitudinal studies

Generalized linear mixed-effects models

Advanced topics

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

Summary

Longitudinal studies

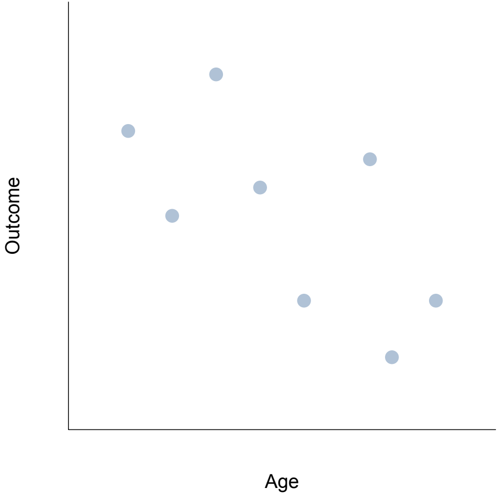
Repeatedly collect information on the same individuals over time

Benefits

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

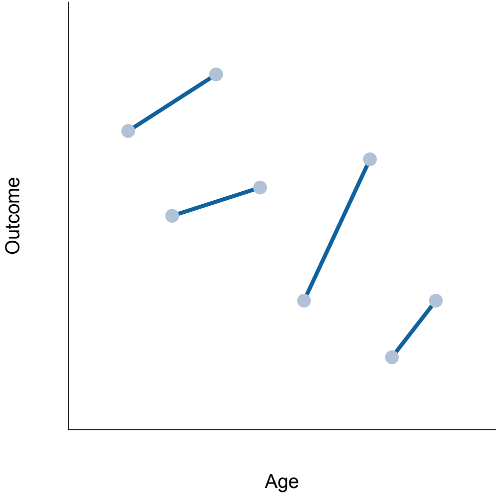
Longitudinal studies

Identify time effects: cohort, age



Longitudinal studies

Identify time effects: cohort, age



Longitudinal studies

Identify 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

Summarize 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

Help establish causal effect of exposure on 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

Longitudinal studies

Repeatedly collect information on the same individuals over time

Challenges

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

Longitudinal studies

Require specialized methods that account for longitudinal correlation

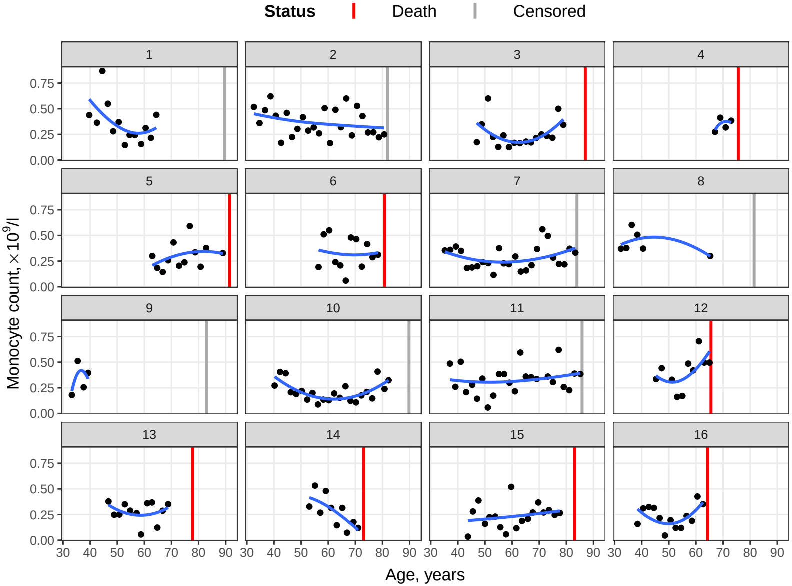
- Individuals are assumed to be independent
- Longitudinal dependence is a secondary feature
- Ignoring dependence may lead to incorrect inference
  - Longitudinal correlation usually positive
  - Estimated standard errors may be too small
  - Confidence intervals are too narrow; too often exclude true value

Example 1

Longitudinal changes in peripheral monocytes (Yoshida et al., 2019)

- **Adult Health Study**
  - Subset of Life Span Study of atomic bomb survivors
  - Biennial clinic examinations since 1958
  - Detailed questionnaire and laboratory data
- DS02R1 radiation doses estimated from dosimetry system
- **Outcome of interest**
  - Monocyte count (longitudinal) as a measure of inflammation
- **Research questions**
  - What is the association between radiation and monocyte counts?
  - How does the association differ by sex and age?
  - Others?

AHS data



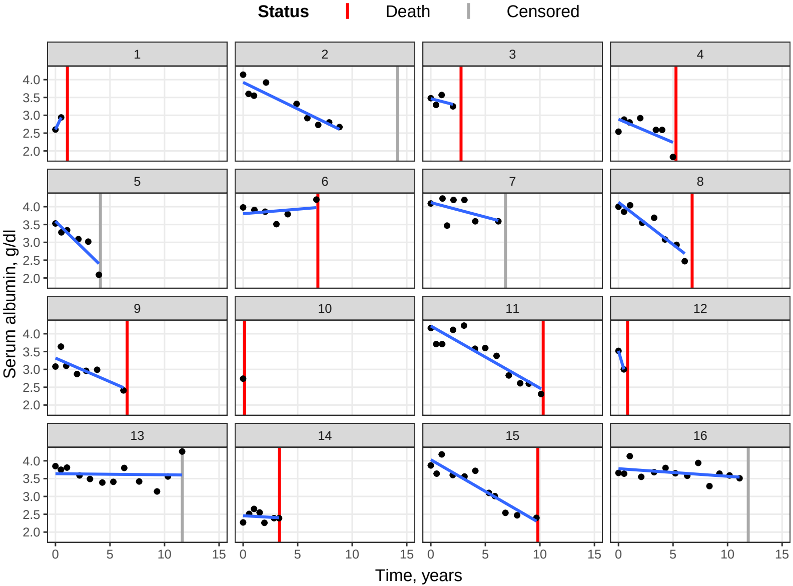
Status
Death
Censored
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
0.00
0.25
0.50
0.75
30
40
50
60
70
80
90
Monocyte count, $\times 10^9$/l
Age, years

Example 2

Mayo Clinic trial in primary biliary cirrhosis (Murtaugh et al., 1994)

- **Primary biliary cirrhosis**
  - Chronic and fatal but rare liver disease
  - Inflammatory destruction of small bile ducts within the liver
  - Patients referred to Mayo Clinic, 1974–1984
- 158 patients randomized to treatment with D-penicillamine; 154 randomized to placebo
- **Outcome of interest**
  - Serum albumin levels (longitudinal) as a measure of liver function
- **Research questions**
  - How do serum albumin levels change over time?
  - Does treatment improve serum albumin levels?
  - Others?

PBC data


Status
Death
Censored
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
4.0
3.5
3.0
2.5
2.0
Serum albumin, g/dl
0
5
10
15
Time, years

Analysis approaches

Must account for correlation due to repeated measurements over time

- Failure to account for correlation ⇒ incorrect standard estimates, resulting in incorrect confidence intervals and hypothesis tests

- **Approaches**: Include all observed data in a regression model for the mean response and account for longitudinal correlation
  - **Generalized estimating equations** (GEE): A marginal model for the mean response and a model for longitudinal correlation

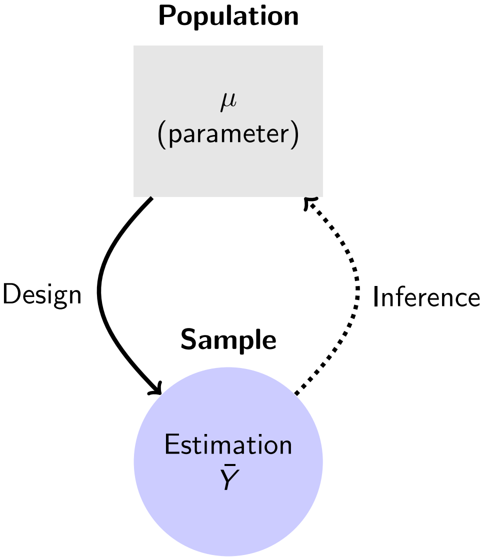
$$g(\mathrm{E}[Y_{ij} \mid x_{ij}]) = x_{ij}\beta \quad \text{and} \quad \mathrm{Corr}[Y_{ij}, Y_{ij'}] = \rho(\alpha),\, j \neq j'$$

  - **Generalized linear mixed-effects models** (GLMM): A conditional model for the mean response given subject-specific random effects, which induce a (possibly hierarchical) correlation structure

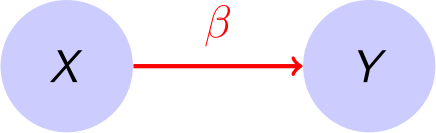
$$g(\mathrm{E}[Y_{ij} \mid x_{ij},\, b_i]) = x_{ij}\beta + z_{ij}b_i \quad \text{with} \quad b_i \sim N(0, D)$$

NB: Differences in interpretation of β between GEE and GLMM

Statistics


Population
μ
(parameter)
Design
Inference
Sample
Estimation
$\bar{Y}$

Regression



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

Estimation

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

Inference

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

Effect modification

- Association of interest varies across levels of another variable, or another variable modifies the association of the variable of interest
- Modeling of effect modification is achieved by interaction terms

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

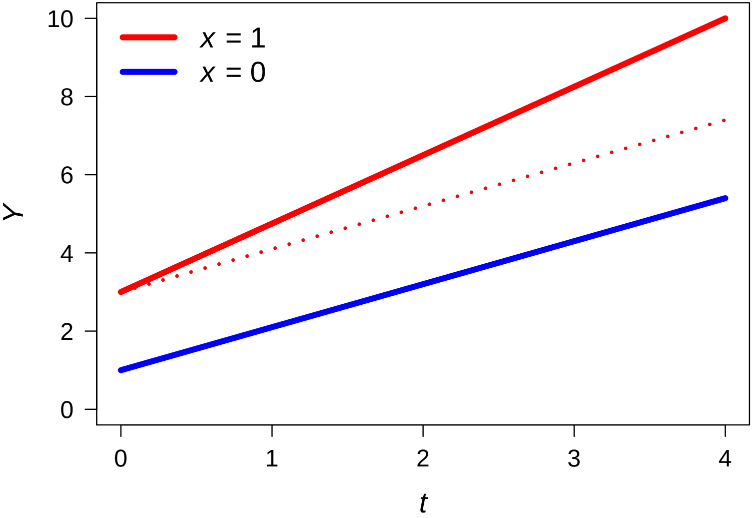
  with

  - A binary variable x for drug: 0 for placebo, 1 for treatment
  - A continuous variable t for time since randomization

- Wish to examine whether treatment modifies the association between time since randomization and serum albumin

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

Effect modification



Effect modification

- Contrasts for t (time) depend on the value for x (drug)

$$\begin{aligned} & \mathrm{E}[Y \mid x, t+1] - \mathrm{E}[Y \mid x, t] \\ & \quad = \{\beta_0 + \beta_1 \cdot x + \beta_2 \cdot (t+1) + \beta_3 \cdot x \cdot (t+1)\} \\ & \qquad - \{\beta_0 + \beta_1 \cdot x + \beta_2 \cdot t + \beta_3 \cdot x \cdot t\} \\ & \quad = \beta_2 + \beta_3 x \end{aligned}$$

- β_2 compares the mean albumin level between two placebo-treated populations whose time since randomization differs by 1 year ($x = 0$)
- $\beta_2 + \beta_3$ compares the mean albumin level between two drug-treated populations whose time since randomization differs by 1 year ($x = 1$)
- Hence β_3 represents a difference evaluating whether the association between time and serum albumin differs between treatment groups
- A hypothesis test of $\beta_3 = 0$ can be used to evaluate the difference

Overview

Introduction to longitudinal studies

Generalized linear mixed-effects models

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

Summary

Mixed-effects models

⋆ 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
  - 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

(Laird and Ware, 1982)

Set-up

For subject i the mixed-effects model is characterized by

$$y_i = \{y_{i1}, y_{i2}, \ldots, y_{im_i}\}^{\mathsf{T}}$$

$$\begin{aligned} \beta &= \{\beta_0, \beta_1, \beta_2, \ldots, \beta_p\}^{\mathsf{T}} && \text{Fixed effects} \\ x_{ij} &= \{1, x_{ij1}, x_{ij2}, \ldots, x_{ijp}\} \\ X_i &= \{x_{i1}, x_{i2}, \ldots, x_{im_i}\}^{\mathsf{T}} && \text{Design matrix for fixed effects} \end{aligned}$$

$$\begin{aligned} b_i &= \{b_{i0}, b_{i1}, b_{i2}, \ldots, b_{iq}\}^{\mathsf{T}} && \text{Random effects} \\ z_{ij} &= \{1, z_{ij1}, z_{ij2}, \ldots, z_{ijq}\} \\ Z_i &= \{z_{i1}, z_{i2}, \ldots, z_{im_i}\}^{\mathsf{T}} && \text{Design matrix for random effects} \end{aligned}$$

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

Linear mixed-effects model

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

1. Model for response given random effects

$$y_{ij} = x_{ij}\beta + z_{ij}b_i + \epsilon_{ij}$$

   with

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

2. Model for random effects

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

   with b_i and ϵ_{ij} assumed to be independent

Choices for random effects

Consider the linear mixed-effects models that include

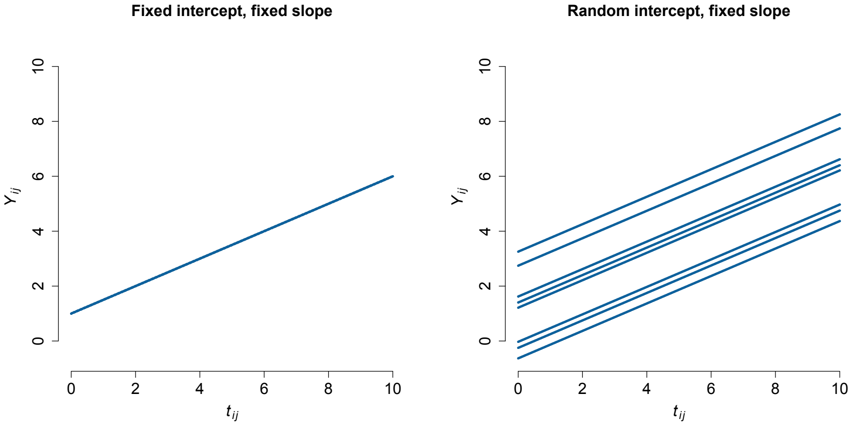
- **Random intercepts**

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

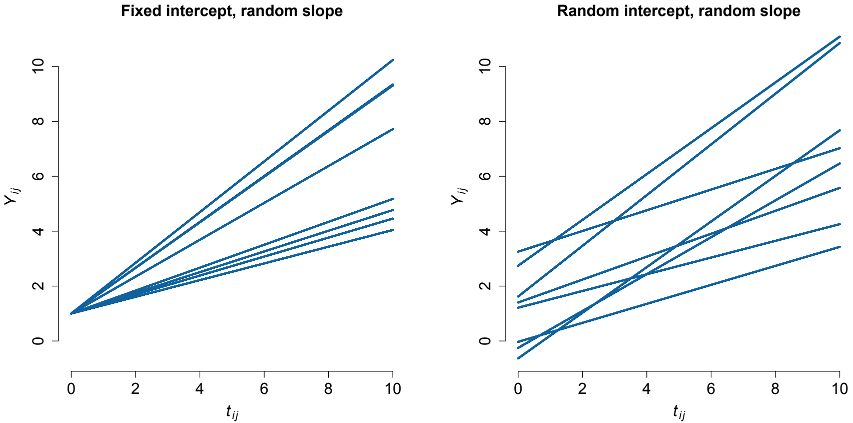
- **Random intercepts and slopes**

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

Choices for random effects



Choices for random effects



Choices for random effects: D

D quantifies random variation in trajectories across subjects

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

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

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

Induced correlation structure

What is the correlation between measurements on the same subject?

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

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{i0} + \epsilon_{ij} \\ y_{ij'} &= \beta_0 + \beta_1 t_{ij'} + b_{i0} + \epsilon_{ij'} \end{aligned}$$

$$\begin{aligned} \text{Var}[Y_{ij}] &= \text{Var}_b[\text{E}_Y(Y_{ij} \mid b_{i0})] + \text{E}_b[\text{Var}_Y(Y_{ij} \mid b_{i0})] \\ &= D_{11} + \sigma^2 \end{aligned}$$

$$\begin{aligned} \text{Cov}[Y_{ij},\ Y_{ij'}] &= \text{Cov}_b[\text{E}_Y(Y_{ij} \mid b_{i0}),\, \text{E}_Y(Y_{ij'} \mid b_{i0})] \\ &\quad + \text{E}_b[\text{Cov}_Y(Y_{ij},\ Y_{ij'} \mid b_{i0})] \\ &= D_{11} \end{aligned}$$

Induced correlation structure

- **Random intercepts model** (continued)

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

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

Induced correlation structure

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

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

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

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

Induced correlation structure

- **Random intercepts and slopes model** (continued)

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

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

Likelihood-based estimation of β

Requires specification of a complete probability distribution for the data

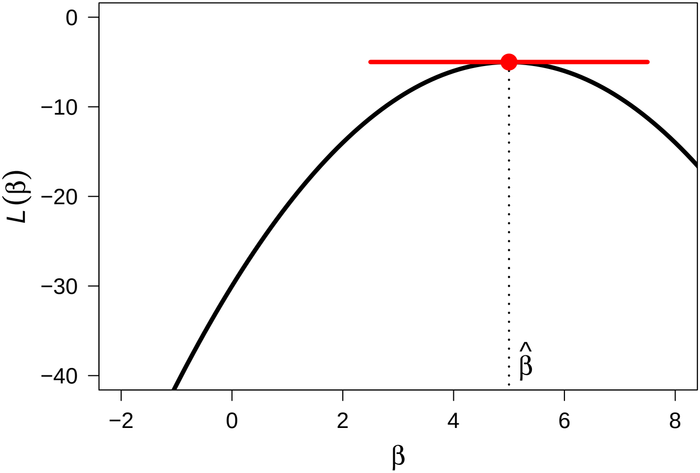
- Likelihood-based methods are designed for fixed effects, so integrate over the assumed distribution for the random effects

$$\mathcal{L}(\beta, \sigma, D) = \prod_{i=1}^{n} \int f_Y(y_i \mid b_i, \beta, \sigma) \times f_b(b_i \mid D) db_i$$

  where f_b is typically the density function of a Normal random variable

  - For linear models the required integration is straightforward because y_i and b_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)
  - Conditional likelihood methods treat the random effects as fixed and condition on statistics for them

Likelihood-based estimation of β



Likelihood-based inference for β

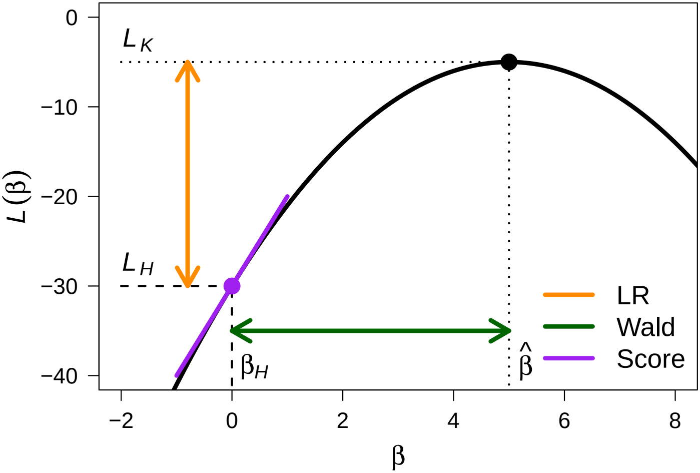
Consider testing fixed effects in nested linear mixed-effects models

$$H\colon \beta = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix} \quad \text{versus} \quad K\colon \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

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

- Likelihood ratio test is valid with maximum likelihood estimation
  - Requires computation under the null and alternative hypotheses
- Likelihood ratio test may not be valid with other estimation methods
- Wald test (based on coefficient and standard error) is generally valid

Likelihood-based inference for β



Likelihood-based inference for D

Consider testing whether a random intercept model is adequate

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

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

- 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
- 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 $z_{ij}b_i$)
  - Included important covariates in the model
  - Correctly specified any transformations or interactions
- **Covariance model**: Correct covariance model (random-effects specification) is required for correct standard error estimates for $\hat{\beta}$
- **Normality**: Normality of ϵ_{ij} and b_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

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
  - Random-effects parameters unique to each individual subject
- Estimation and inference can focus both on average outcome levels and trends, and on heterogeneity across subjects in levels and trends
- Subject-specific random effects induce a correlation structure
- Parametric likelihood approach permits use of likelihood ratio test, but requires several assumptions that must be verified in practice

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

Generalized linear mixed-effects models

Advanced topics

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

Summary

Conditional and marginal effects

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

Interpretation of GLMM

		Fitted 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

Example

Consider a logistic regression model with subject-specific intercepts

$$\text{logit}(\text{P}[Y_{ij} = 1 \mid b_{i0}]) = \beta_0^\star + \beta_1^\star x_{ij} + b_{i0}$$

where each subject has their own baseline risk of the disease ($x_{ij} = 0$)

$$\frac{\exp(\beta_0^\star + b_{i0})}{1 + \exp(\beta_0^\star + b_{i0})}$$

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

Example

The population rate of infection is the average risk across individuals

$$\begin{aligned}\mathrm{P}[Y_{ij}=1] &= \int \mathrm{P}[Y_{ij}=1 \mid b_{i0}]\, dF(b_{i0}) \\ &= \int \frac{\exp(\beta_0^\star + \beta_1^\star x_{ij} + b_{i0})}{1+\exp(\beta_0^\star + \beta_1^\star x_{ij} + b_{i0})} f(b_{i0} \mid \tau)\, db_{i0}\end{aligned}$$

where typically $b_{i0} \sim N(0, D_{11})$

- Assuming $\{\beta_0^\star, \beta_1^\star\} = \{-2, 0.4\}$ so that the exposure odds ratio is

  $$\exp(0.4) = 1.5$$

  and $D_{11} = 2$ the population rates (via integration) are

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

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

Missing data

- Missing values arise in longitudinal studies whenever the intended serial observations collected on a subject over time are incomplete
  - Collect fewer data than planned ⇒ decreased efficiency (power)
  - Missingness can depend on outcome values ⇒ potential bias
- Important to distinguish between missing data and unbalanced data, although missing data necessarily result in unbalanced data
- Missing data require consideration of the factors that influence the missingness of intended observations
- Also important to distinguish between intermittent missing values (non-monotone) and dropouts in which all observations are missing after subjects are lost to follow-up (monotone)

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

Mechanisms

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 (Rubin, 1976)

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
  - 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 can be evaluated using the observed data

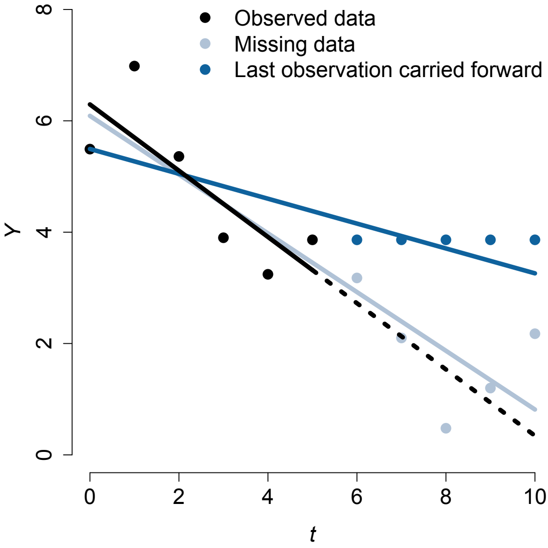
Last observation carried forward

- 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



Time-dependent exposures

Important analytical issues arise with time-dependent exposures

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

Definitions

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

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

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

- **Endogenous**: Not exogenous

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

Examples

Exogeneity may be assumed based on the design or evaluated empirically

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

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

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

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

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

Key assumption

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

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

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

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

Otherwise an independence estimating equation should be used

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

Overview

Introduction to longitudinal studies

Generalized linear mixed-effects models

Advanced topics

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

Summary

Key points

- Conditional mean regression model
- Model for population heterogeneity
- Subject-specific random effects induce a correlation structure
- Multiple sources of positive correlation
- 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 b_i
- 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
- R package `lme4`; Stata commands `mixed`, `melogit`

Big picture

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

Advanced

- Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*, 2$^{\text{nd}}$ 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.