Module 10

Generalized Estimating Equations for Longitudinal Data Analysis

Benjamin French, PhD
Department of Biostatistics, Vanderbilt University

SISCER 2021
July 19, 2021

Learning objectives

- This module will overview statistical methods for the analysis of longitudinal data, with a focus on estimating equations
- 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 estimating equations

Advanced topics

- Missing data
- Time-dependent exposures

Summary

Overview

Introduction to longitudinal studies

Generalized estimating equations

Advanced topics
- Missing data
- Time-dependent exposures

Summary

Longitudinal studies

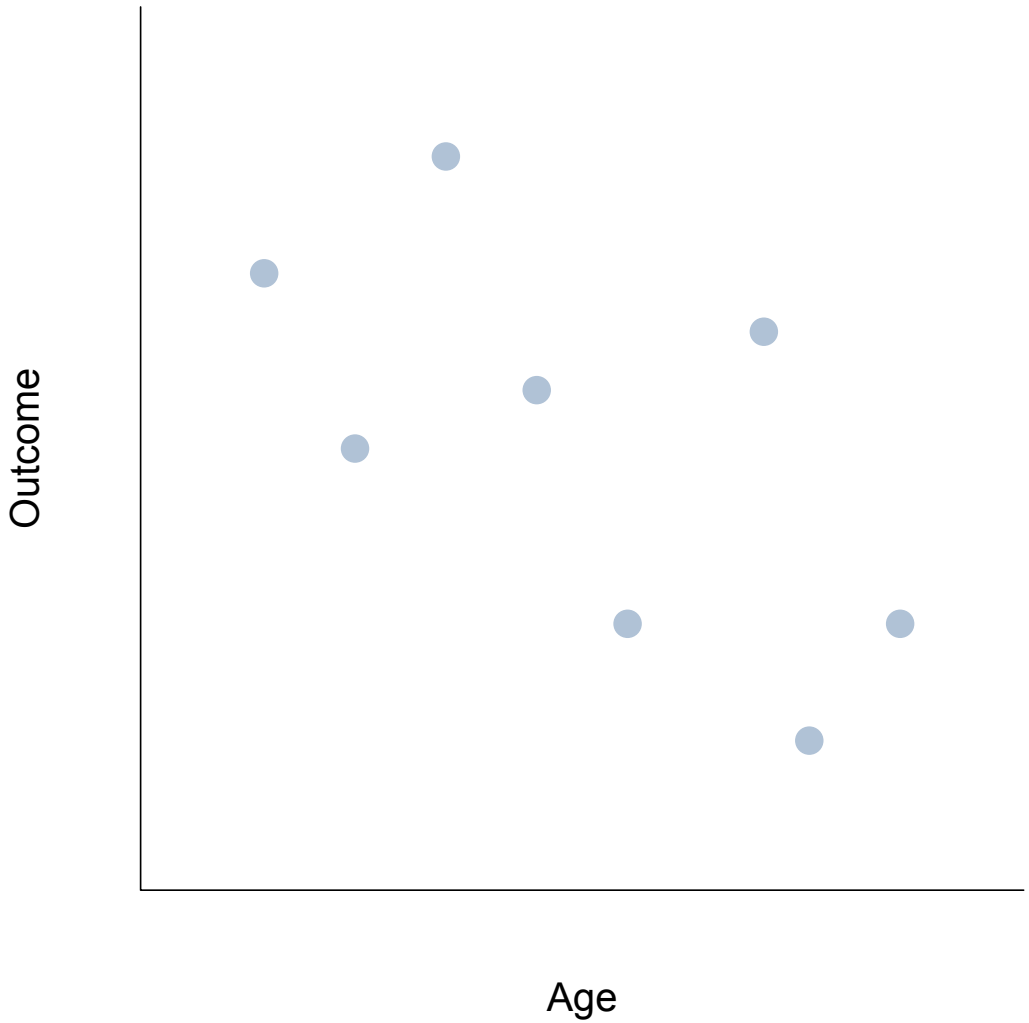
Repeatedly collect information on the same individuals over time

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

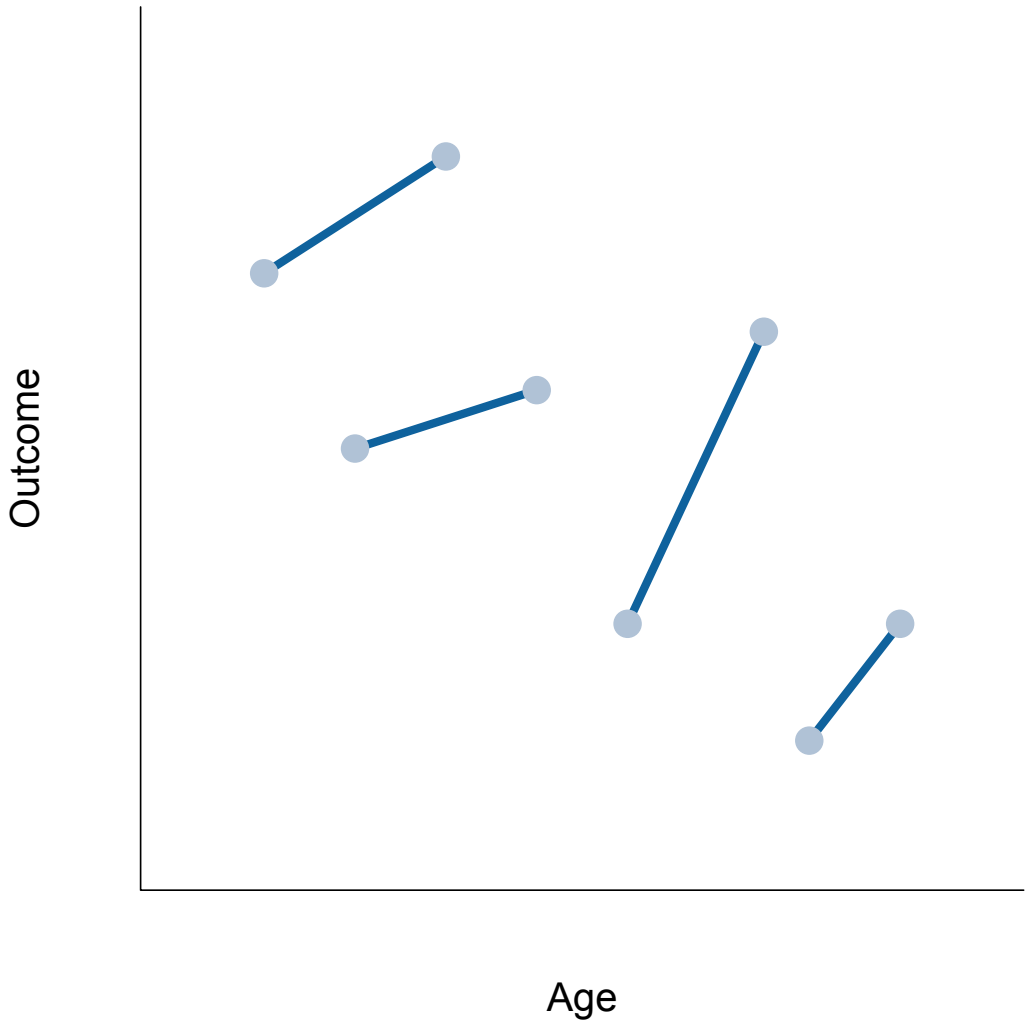
Longitudinal studies

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

$$\begin{aligned} \text{Egg} &\rightarrow \text{Chicken} \\ \text{Chicken} &\rightarrow \text{Egg} \end{aligned}$$

- 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

- 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

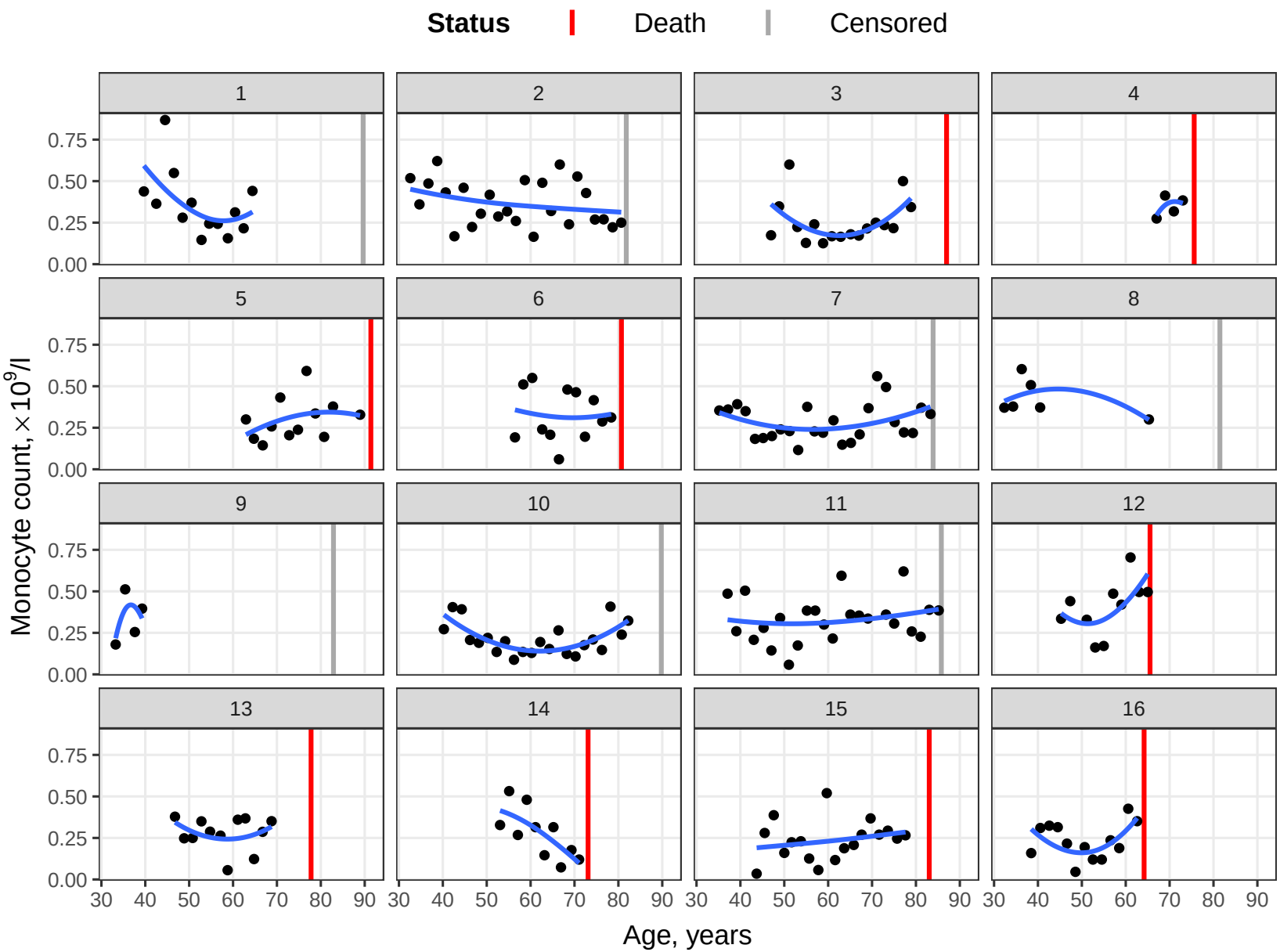
- Individuals are assumed to be independent
- Longitudinal dependence is a secondary feature
- Ignoring dependence may lead to incorrect inference
  - Longitudinal correlation usually positive
  - Estimated standard errors may be too small
  - Confidence intervals are too narrow; too often exclude true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

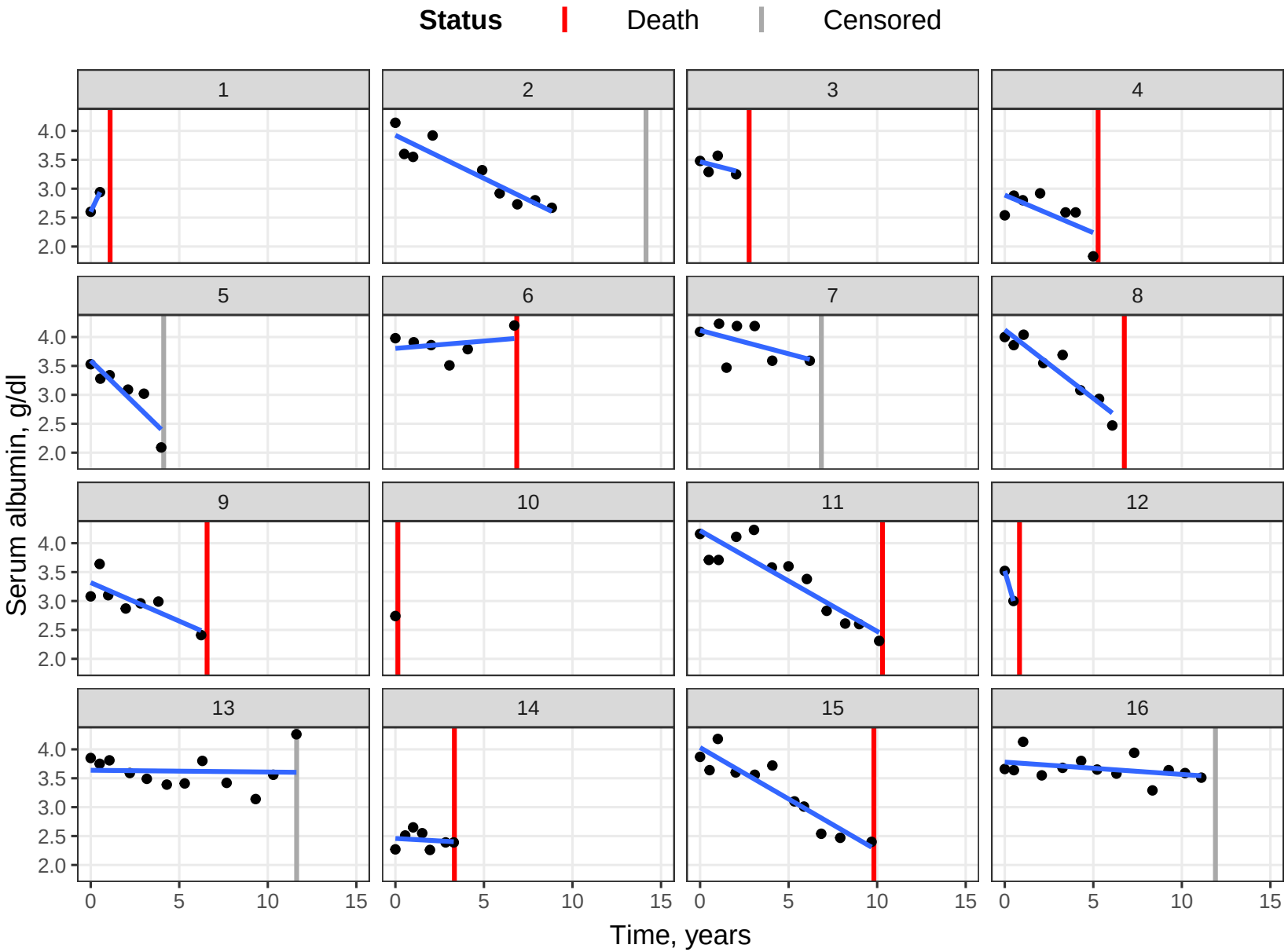

Status
Death
Censored
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
0.00
0.25
0.50
0.75
30
40
50
60
70
80
90
Monocyte count, ×10^9/l
Age, years

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
Serum albumin, g/dl
Time, years

Analysis approaches

Must account for correlation due to repeated measurements over time

- Failure to account for correlation $\Rightarrow$ 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

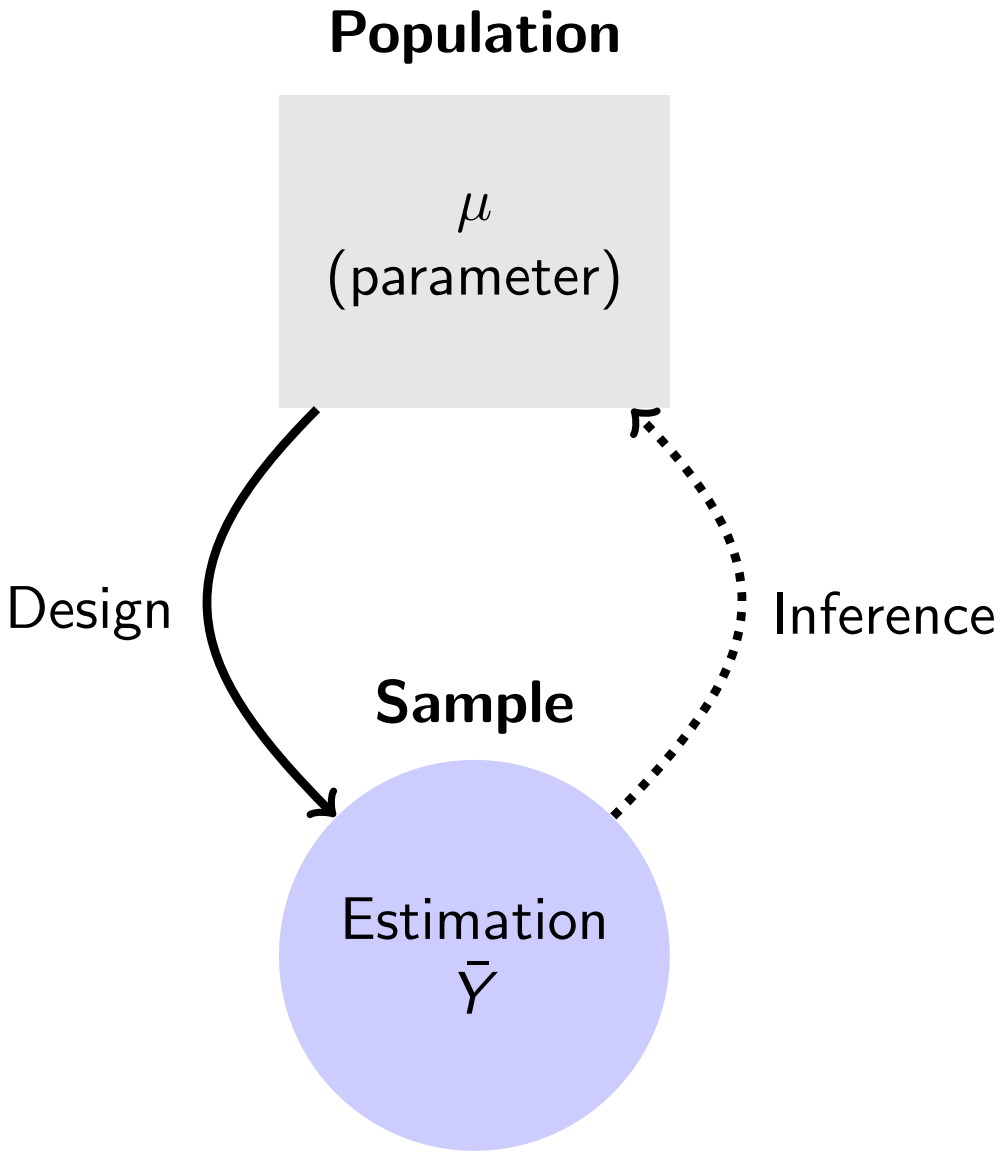
$$g(\mathrm{E}[Y_{ij} \mid x_{ij}]) = x_{ij}\beta \quad \text{and} \quad \mathrm{Corr}[Y_{ij}, Y_{ij'}] = \rho(\alpha),\ j \neq j'$$

  - **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

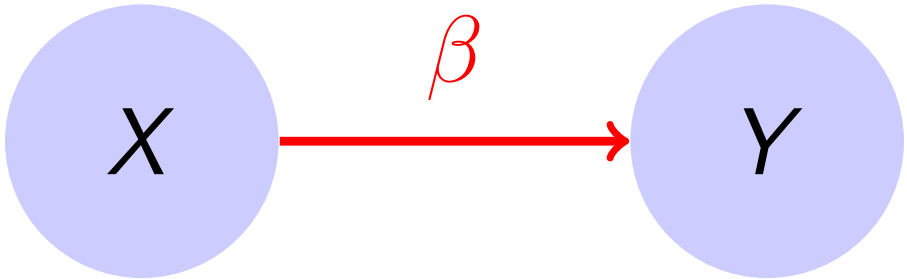
$$g(\mathrm{E}[Y_{ij} \mid x_{ij},\ b_i]) = x_{ij}\beta + z_{ij}b_i \quad \text{with} \quad b_i \sim N(0, D)$$

NB: Differences in interpretation of β between GEE and GLMM

Statistics


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

Regression



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

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

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

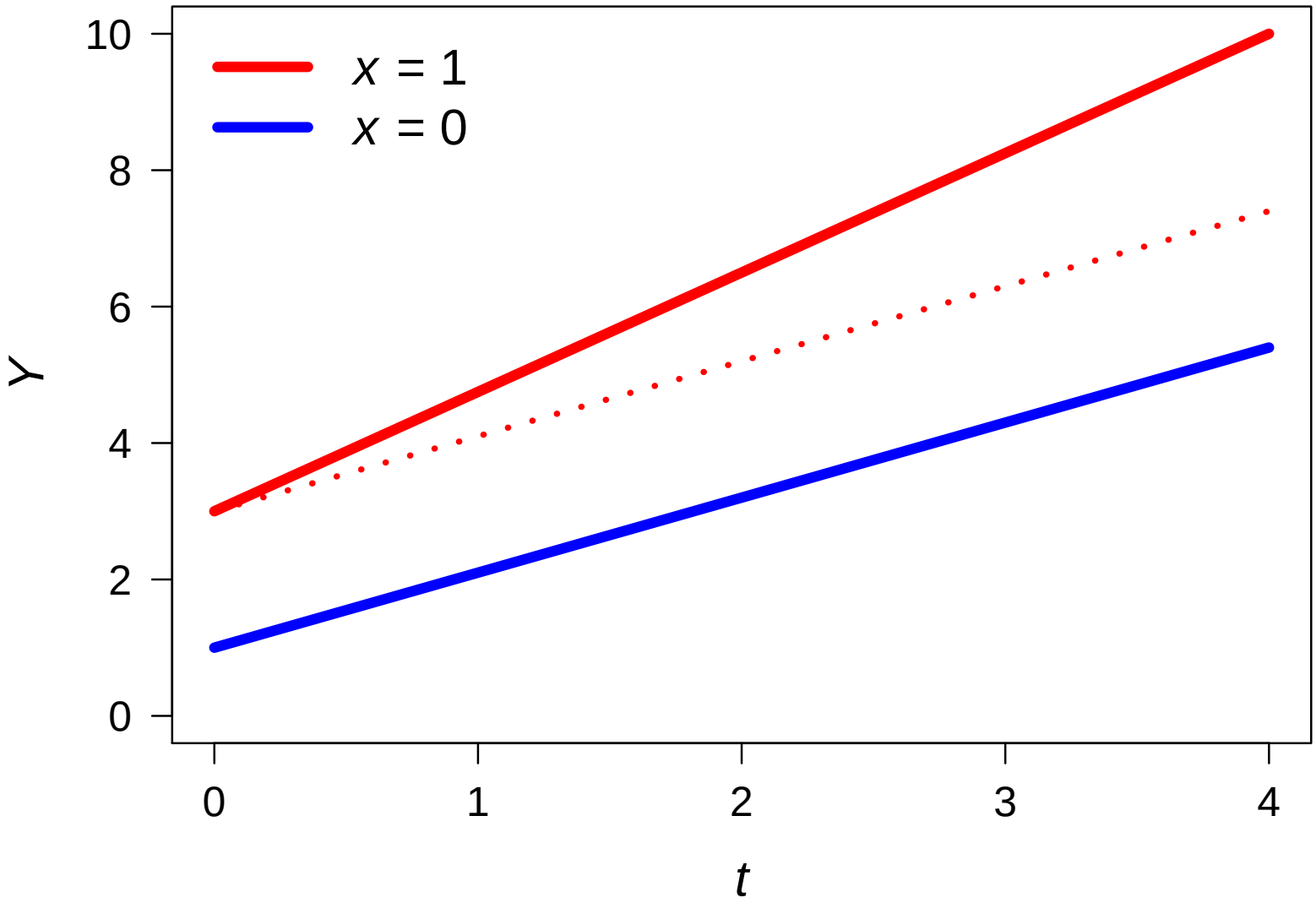
with

  - 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)

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

- 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 estimating equations

Advanced topics
- Missing data
- Time-dependent exposures

Summary

GEE

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

- Focus on a generalized linear model with regression parameters β, which characterize the systemic variation in Y across covariates X

$$
\begin{aligned}
y_i &= \{y_{i1}, y_{i2}, \ldots, y_{im_i}\}^{\mathsf{T}} && \text{Outcomes} \\
x_{ij} &= \{1, x_{ij1}, x_{ij2}, \ldots, x_{ijp}\} && \text{Covariates} \\
X_i &= \{x_{i1}, x_{i2}, \ldots, x_{im_i}\}^{\mathsf{T}} && \text{Design matrix} \\
\beta &= \{\beta_0, \beta_1, \beta_2, \ldots, \beta_p\}^{\mathsf{T}} && \text{Regression parameters}
\end{aligned}
$$

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

- Longitudinal correlation structure is a nuisance feature of the data

(Liang and Zeger, 1986)

Mean model

Assumptions

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

Mean model: Primary focus of the analysis

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

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

Covariance model

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

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

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

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

2. Select a model for longitudinal correlation with parameters α

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

Covariance model

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

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

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

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

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

Correlation models

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

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

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

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

Correlation models

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

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

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

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

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**
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
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

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

- 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 model?

Answer: Selecting a non-independence or weighted correlation model

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

Variance estimators

- **Independence estimating equation**: An estimation equation with a working independence correlation model
  - 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 model
  - 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

Consider testing one or more parameters in nested 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., $H\colon\ \beta_2 = 0$

- Wald test (based on coefficient and standard error) is generally valid
  - Requires computation under the alternative hypothesis K
- Likelihood ratio test not available; not relied on a likelihood function

Summary

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

Issues

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

Overview

Introduction to longitudinal studies

Generalized estimating equations

Advanced topics
- Missing data
- Time-dependent exposures

Summary

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 $\Rightarrow$ decreased efficiency (power)
  - Missingness can depend on outcome values $\Rightarrow$ 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

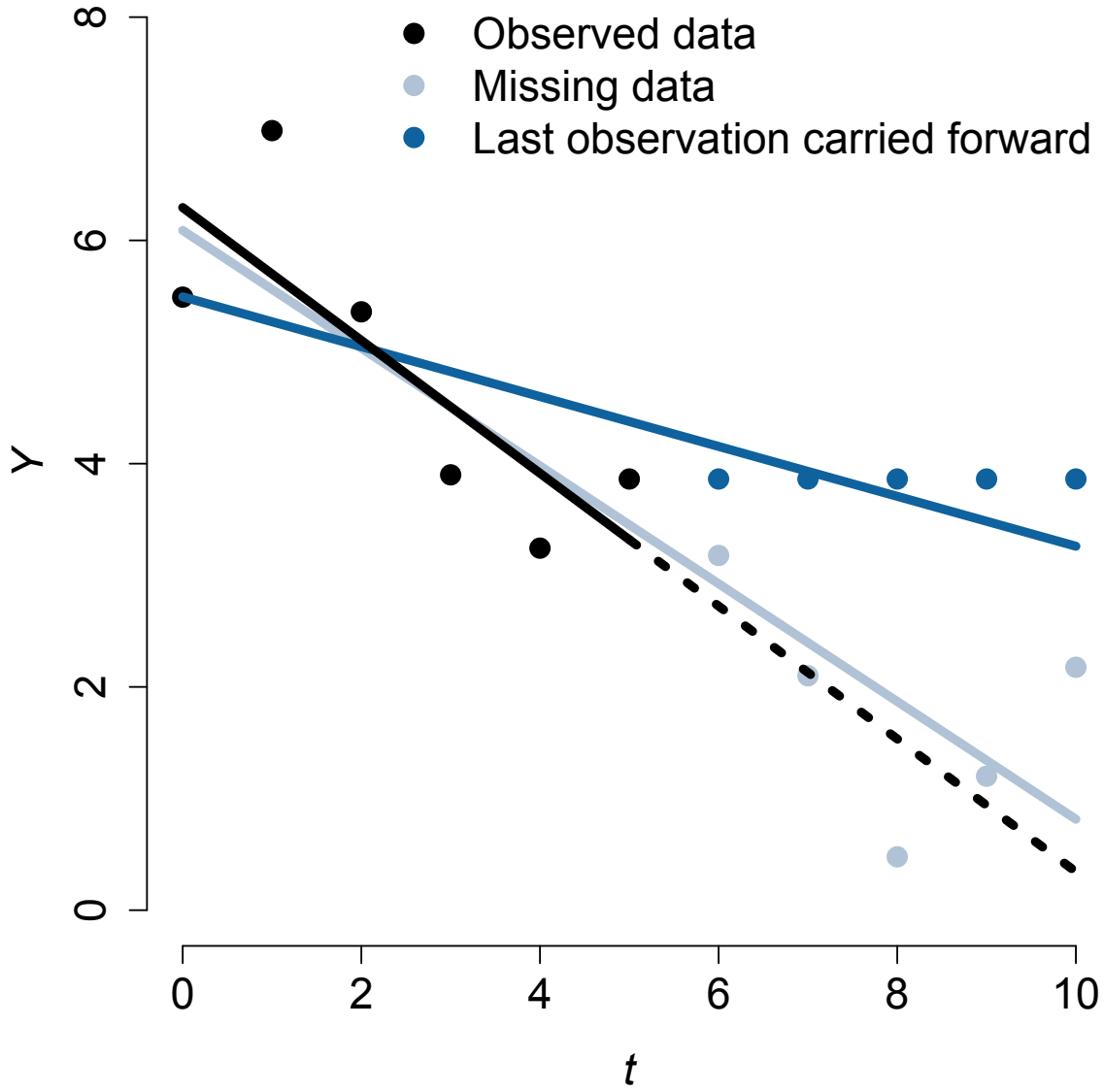
Last observation carried forward

- 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	2.0	2.0
2	4.1	3.5	3.8	2.4	2.8
3	2.7	2.4	2.9	3.5	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 $y_i(t-1)$

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

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

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

⋆ 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 estimating equations

Advanced topics
- Missing data
- Time-dependent exposures

Summary

Key points

- Marginal mean regression model
- Model for longitudinal correlation
- Only one source of positive or negative 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: $n \geq 40$
- Efficiency of non-independence correlation models
- Testing with Wald tests
- Marginal or population-averaged inference
- Missing completely at random (MCAR)
- Time-dependent covariates and endogeneity
- R package `geepack`; Stata command `xtgee`

Big picture

- Provide valid estimates and standard errors for regression parameters of interest even if the correlation model is incorrectly specified $(+)$
- Empirical variance estimator requires large sample size $(-)$
- 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 $(-)$

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*, 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.