Module 8

Case Studies in Longitudinal Data Analysis

Benjamin French, PhD
Radiation Effects Research Foundation
University of Pennsylvania

SISCR 2017
July 25, 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
- Case studies will be used to discuss analysis strategies, the application of appropriate analysis methods, and the interpretation of results, with 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

Review: Longitudinal data analysis

Case study: Longitudinal depression scores

Case study: Indonesian Children's Health Study

Case study: Carpal tunnel syndrome

Summary and resources

Overview

Review: Longitudinal data analysis

Case study: Longitudinal depression scores

Case study: Indonesian Children's Health Study

Case study: Carpal tunnel syndrome

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

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
- Use specialized methods that account for longitudinal correlation

Motivating example

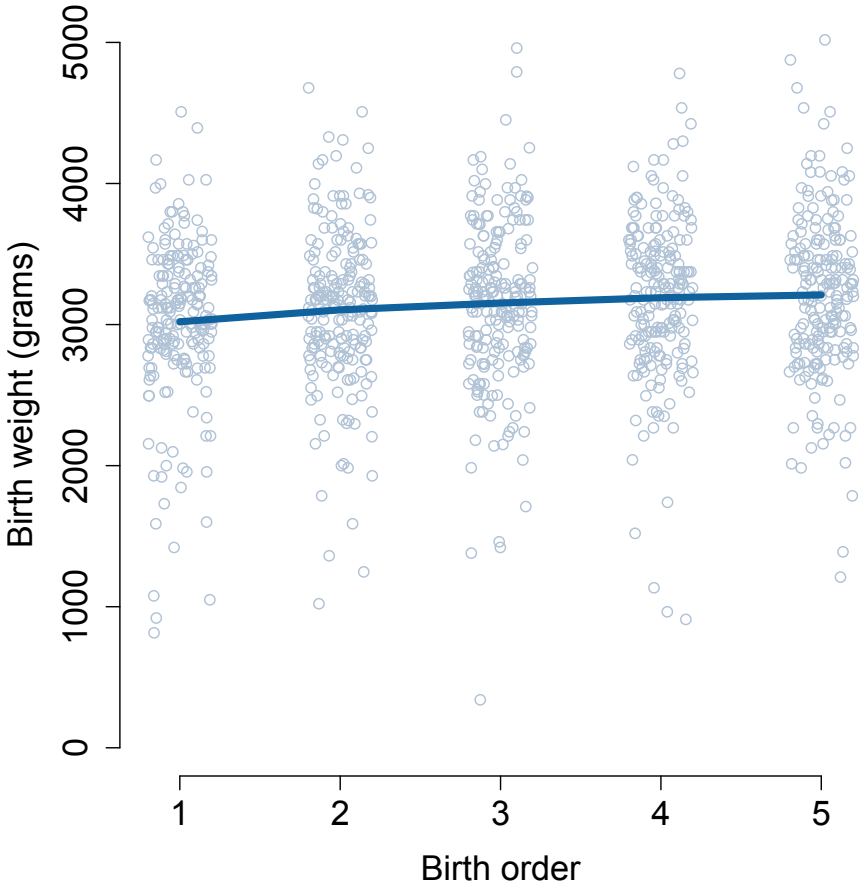
Georgian infant birth weight

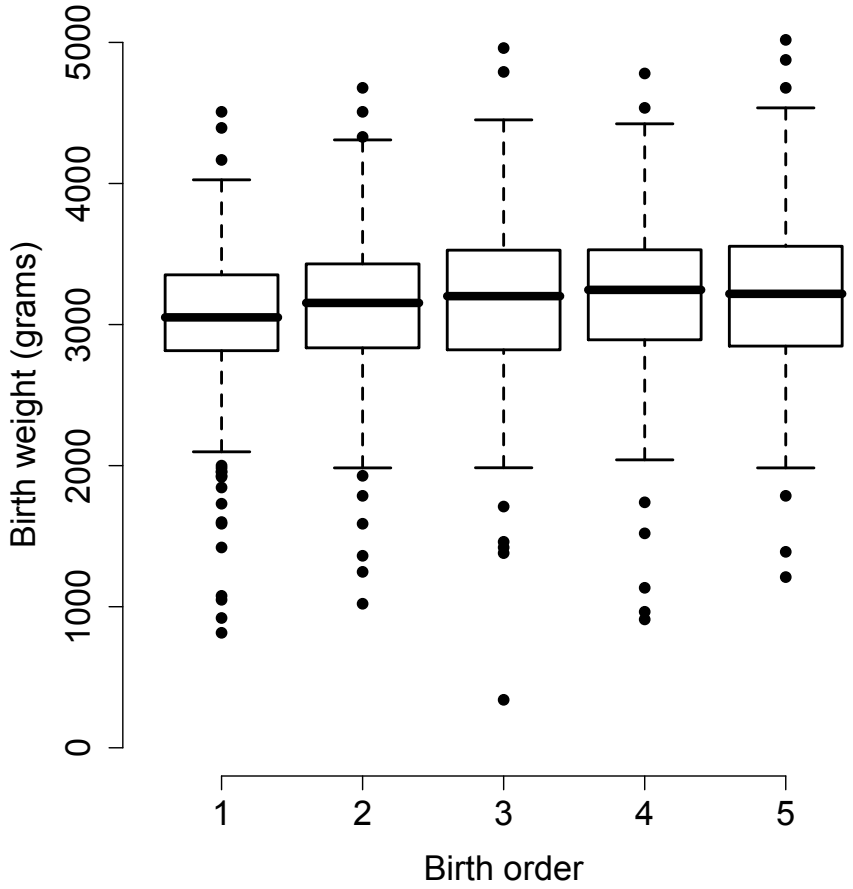
- Birth weight measured for each of $m = 5$ children of $n = 200$ mothers
- Birth weight for infants j comprise repeated measures on mothers i
- Interested in the association between birth order and birth weight
  - Estimate the average time course among all mothers
  - Estimate the time course for individual mothers
  - Quantify the degree of heterogeneity across mothers
- Consider adjustment for mother's initial age (at first birth)

Motivating example

	momid	birthord	bweight	lowbrth	initage
[1]	39	1	3720	0	15
[2]	39	2	3260	0	15
[3]	39	3	3910	0	15
[4]	39	4	3320	0	15
[5]	39	5	2480	1	15
[6]	62	1	2381	1	17
[7]	62	2	2835	1	17
[8]	62	3	2381	1	17
[9]	62	4	2268	1	17
[10]	62	5	2211	1	17

Motivating example





Strategies for analysis of longitudinal data

- **Derived variable**: Collapse the longitudinal series for each subject into a summary statistic, such as a difference (a.k.a. "change score") or regression coefficient, and use methods for independent data

- **Repeated measures**: Include all data in a regression model for the mean response and account for longitudinal and/or cluster correlation

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

Change and randomized studies

- **Key assumption**: groups equivalent at baseline
- Methods that 'adjust' for baseline are generally preferable due to greater precision
  - $\boxed{\rho > 1/2}$ POST $\prec$ CHANGE $\prec$ ANCOVA
  - $\boxed{\rho < 1/2}$ CHANGE $\prec$ POST $\prec$ ANCOVA
  - 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

Change and non-randomized studies

- 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?
- CHANGE typically most relevant; multivariable methods to come later characterize CHANGE across multiple timepoints

Strategies for analysis of longitudinal data

- **Derived variable**: Collapse the longitudinal series for each subject into a summary statistic, such as a difference (a.k.a. "change score") or regression coefficient, and use methods for independent data
  - **Example**: birth weight of 2nd child – birth weight of 1st child
  - Might be adequate for two time points and no missing data
- **Repeated measures**: Include all data in a regression model for the mean response and account for longitudinal and/or cluster correlation
  - **Generalized estimating equations** (GEE)
  - **Generalized linear mixed-effects models** (GLMM)

Notation

Define

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

Stacks of data for each subject:

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

Dependence and correlation

Issue Response variables measured on the same subject are correlated

- Observations are **dependent** or **correlated** when one variable predicts the value of another variable
  - The birth weight for a first child is predictive of the birth weight for a second child born to the same mother

- **Variance:** measures average distance that an observation falls away from the mean
- **Covariance:** measures whether, on average, departures in one variable $Y_{ij} - \mu_j$ 'go together with' departures in another variable $Y_{ik} - \mu_k$
- **Correlation:** measure of dependence that takes values from -1 to $+1$

Covariance: Something new to model

$$\text{Cov}(Y_i) = \begin{bmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \dots & \text{Cov}(Y_{i1}, Y_{im_i}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \dots & \text{Cov}(Y_{i2}, Y_{im_i}) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(Y_{im_i}, Y_{i1}) & \text{Cov}(Y_{im_i}, Y_{i2}) & \dots & \text{Var}(Y_{im_i}) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \dots & \sigma_1\sigma_{m_i}\rho_{1m_i} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \dots & \sigma_2\sigma_{m_i}\rho_{2m_i} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{m_i}\sigma_1\rho_{m_i1} & \sigma_{m_i}\sigma_2\rho_{m_i2} & \dots & \sigma_{m_i}^2 \end{bmatrix}$$

Note: $\rho =$ correlation

GEE (Liang and Zeger, 1986)

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

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

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

  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 could be correlated within subjects

Mean model: Primary focus of the analysis

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

- Corresponds 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

Covariance model

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

1. Assume a form for **variance** that could 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 could also include a scale or dispersion parameter $\phi > 0$

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

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

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 subjects who contribute more information
- $\boxed{3}$ Simply a "change of scale" from the scale of the mean, μ_i, to the scale of the regression coefficients (covariates)

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
- $\hat{\beta}$ is a consistent estimator for β provided that the mean model is correctly specified, 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
- GEE computes a sandwich variance estimator (aka empirical, robust, or Huber-White variance estimator)
- Empirical variance estimator provides valid standard errors for $\hat{\beta}$ even if the working correlation model is incorrect, but requires $n \geq 40$ (Mancl and DeRouen, 2001)

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	$-/+$	$+$

GEE commands

- Stata: `xtset`, then use `xtgee`
- R: `geeglm` in `geepack` library, using `geese` fitter function
- SAS: PROC GENMOD
- **NB**: Order might be important for analysis in software
  - Requires sorting the data by unique subject identifier and time
  - Important for exchangeable and auto-regressive correlation structures

Motivating example

Interested in the association between birth order and birth weight

$$\mathrm{E}[Y_{ij} \mid x_{ij}] = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}$$

for $i = 1, \ldots, 200$ and $j = 1, \ldots, 5$ with

- Y_{ij}: Infant birth weight (continuous)
- x_{ij1}: Infant birth order
- x_{ij2}: Mother's initial age

Motivating example: Stata commands

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

* Fit a linear model with independence correlation
xtgee bweight birthord initage, corr(ind) robust

* Fit a linear model with exchangeable correlation
xtgee bweight birthord initage, corr(exc) robust
```

Motivating example: Stata output

```
GEE population-averaged model                   Number of obs      =      1000
Group variable:                        momid    Number of groups   =       200
Link:                               identity    Obs per group: min =         5
Family:                             Gaussian                   avg =       5.0
Correlation:                     independent                   max =         5
                                               Wald chi2(2)       =     27.95
Scale parameter:                    324458.3    Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on momid)
------------------------------------------------------------------------------
             |               Semi-robust
     bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord |      46.608   10.02134      4.65   0.000     26.96653    66.24947
     initage |    26.73226    10.1111      2.64   0.008     6.914877    46.54965
       _cons |    2526.622   177.2781     14.25   0.000     2179.164    2874.081
------------------------------------------------------------------------------
```

Motivating example: Stata output

```
GEE population-averaged model                   Number of obs      =      1000
Group variable:                        momid    Number of groups   =       200
Link:                               identity    Obs per group: min =         5
Family:                             Gaussian                   avg =       5.0
Correlation:                   exchangeable                    max =         5
                                                Wald chi2(2)       =     27.95
Scale parameter:                    324458.3    Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on momid)
------------------------------------------------------------------------------
             |               Semi-robust
     bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord |      46.608   10.02134     4.65   0.000     26.96653    66.24947
     initage |    26.73226    10.1111     2.64   0.008     6.914877    46.54965
       _cons |    2526.622   177.2781    14.25   0.000     2179.164    2874.081
------------------------------------------------------------------------------
```

Motivating example: Comments

- Difference in mean birth weight between two populations of infants whose birth order differs by one is 46.6 grams, 95% CI: (27.0, 66.2)
- Model-based standard errors are smaller for exchangeable structure, indicating efficiency gain from assuming a correct correlation structure
- In practice, i.e. with real-world data, it's often difficult to tell what the correct correlation structure is from exploratory analyses
- *A priori* scientific knowledge should ultimately guide the decision
- I tend to use working independence with empirical standard errors unless I have a good reason to do otherwise, e.g. large efficiency gain
- Try not to select the structure that gives you the smallest p-value
- Stata labels the standard errors "semi-robust" because the empirical variance estimator protects against mis-specification of the correlation model, but requires correct specification of the mean model

⋆ See `help xtgee` for detailed syntax, other options, and saved results

GEE 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
- The correlation model does not need to be correctly specified to obtain a consistent estimator for β or valid standard errors for $\hat{\beta}$
- 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

Strategies for analysis of longitudinal data

- **Derived variable**: Collapse the longitudinal series for each subject into a summary statistic, such as a difference (a.k.a. "change score") or regression coefficient, and use methods for independent data
  - **Example**: birth weight of 2nd child – birth weight of 1st child
  - Might be adequate for two time points and no missing data
- **Repeated measures**: Include all data in a regression model for the mean response and account for longitudinal and/or cluster correlation
  - **Generalized estimating equations** (GEE): A marginal model for the mean response and a model for longitudinal or cluster correlation

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

  - **Generalized linear mixed-effects models** (GLMM)

Mixed-effects models (Laird and Ware, 1982)

⋆ 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

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

with

  - 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

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

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

Choices for random effects

Consider the linear mixed-effects models that include

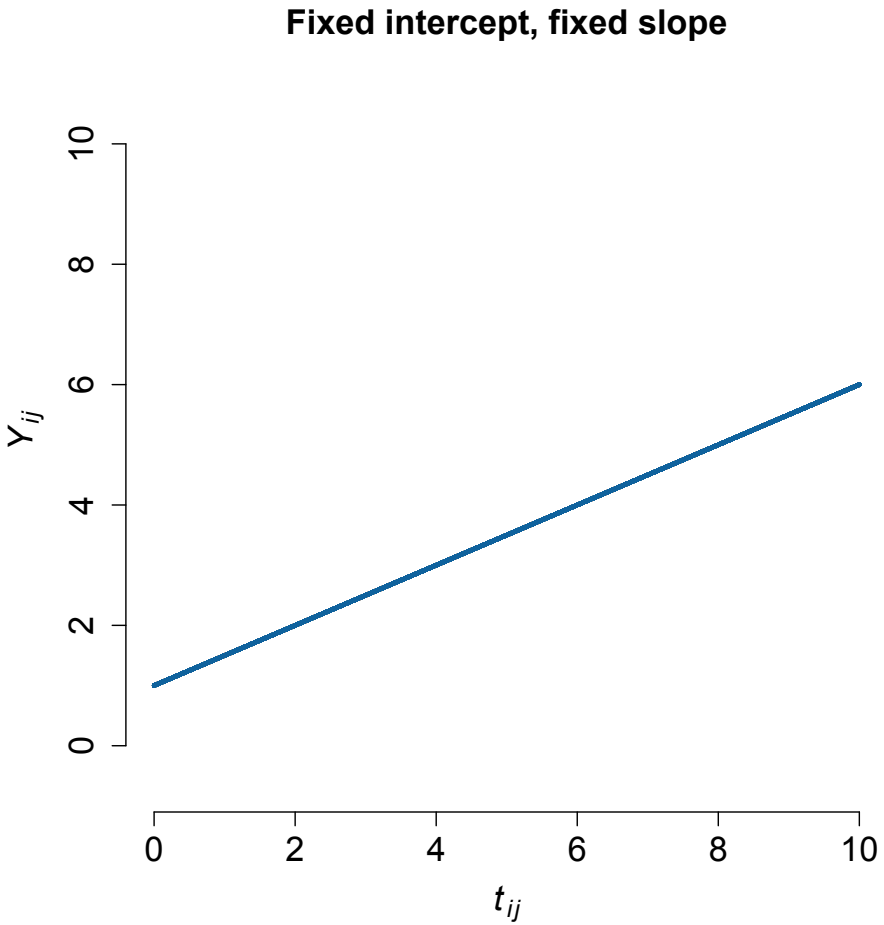
- **Random intercepts**

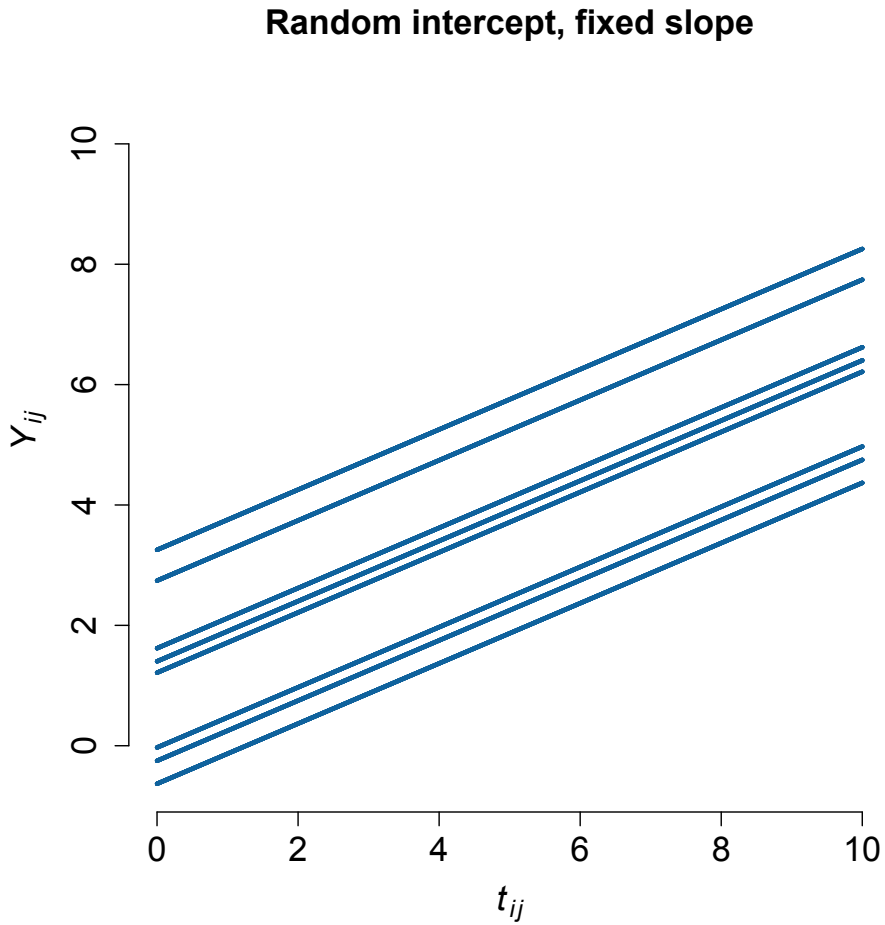
$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \gamma_{0i} + \epsilon_{ij} \\ &= (\beta_0 + \gamma_{0i}) + \beta_1 t_{ij} + \epsilon_{ij} \end{aligned}$$

- **Random intercepts and slopes**

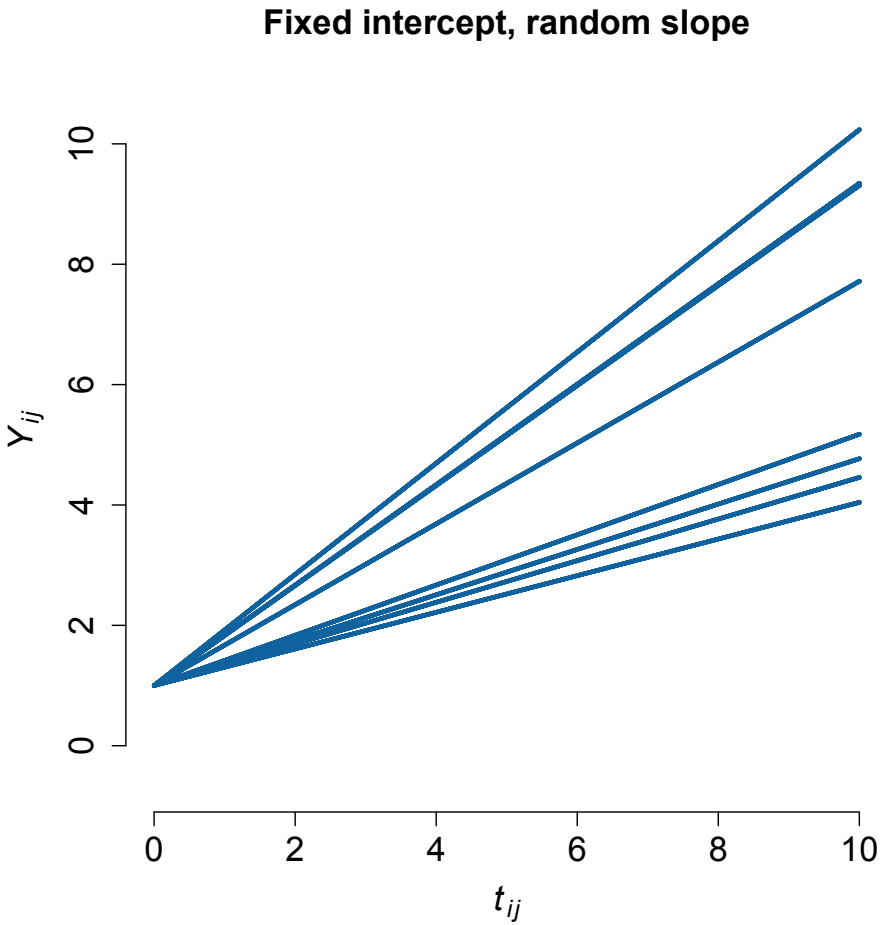
$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \gamma_{0i} + \gamma_{1i} t_{ij} + \epsilon_{ij} \\ &= (\beta_0 + \gamma_{0i}) + (\beta_1 + \gamma_{1i}) t_{ij} + \epsilon_{ij} \end{aligned}$$

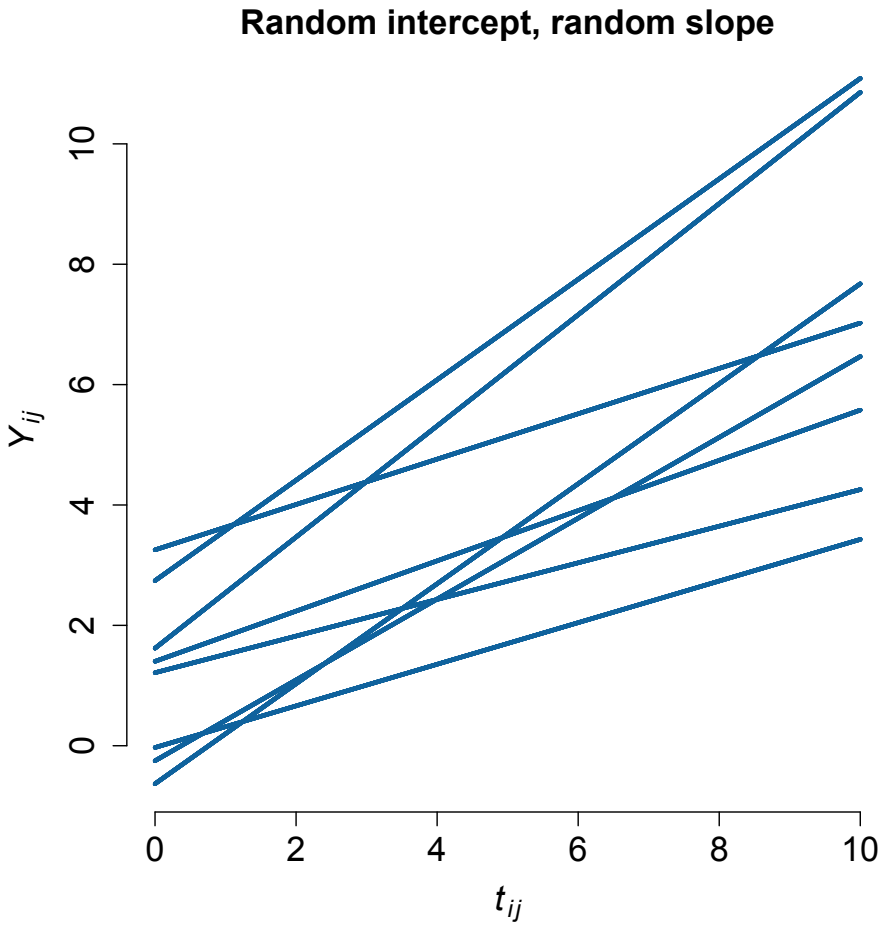
Choices for random effects





Choices for random effects





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

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

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**: $\mu_{ij}^{\star}$ 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 $g(\cdot)$

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

  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
- Could eliminate potentially large sources of bias by controlling for all stable characteristics of the individuals under study $(+)$
- Variation across subjects is ignored, which could 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 $(+)$

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

$\star$ These assumptions must be verified to evaluate any fitted model

Motivating example

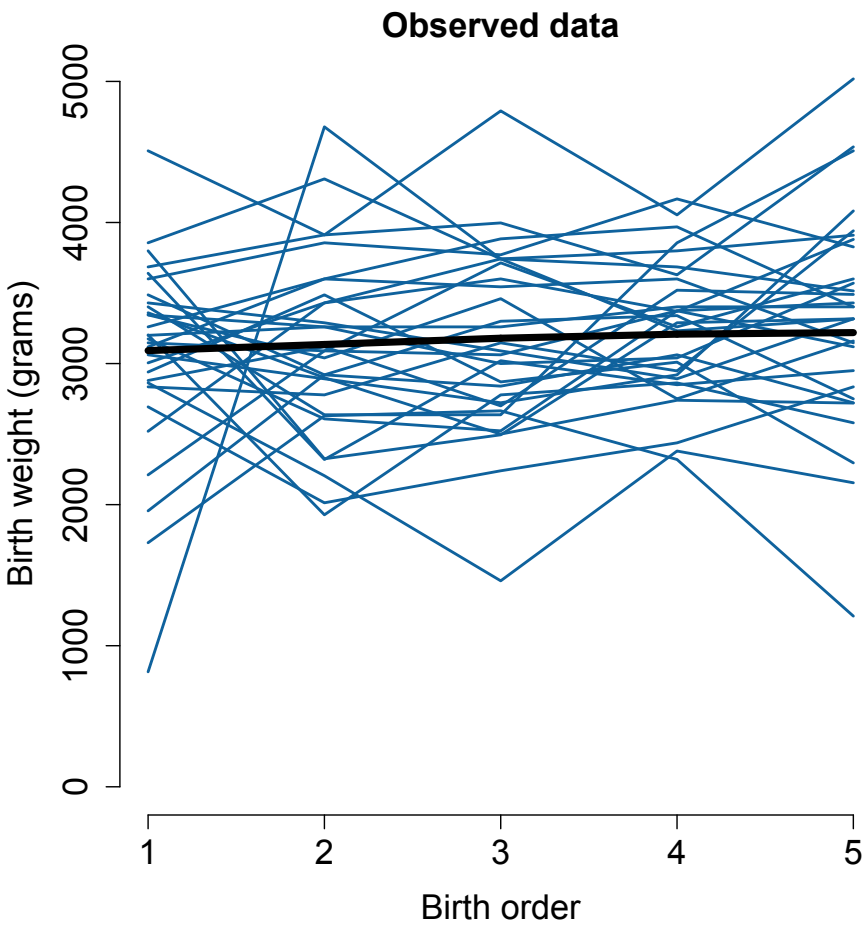
Interested in the association between birth order and birth weight

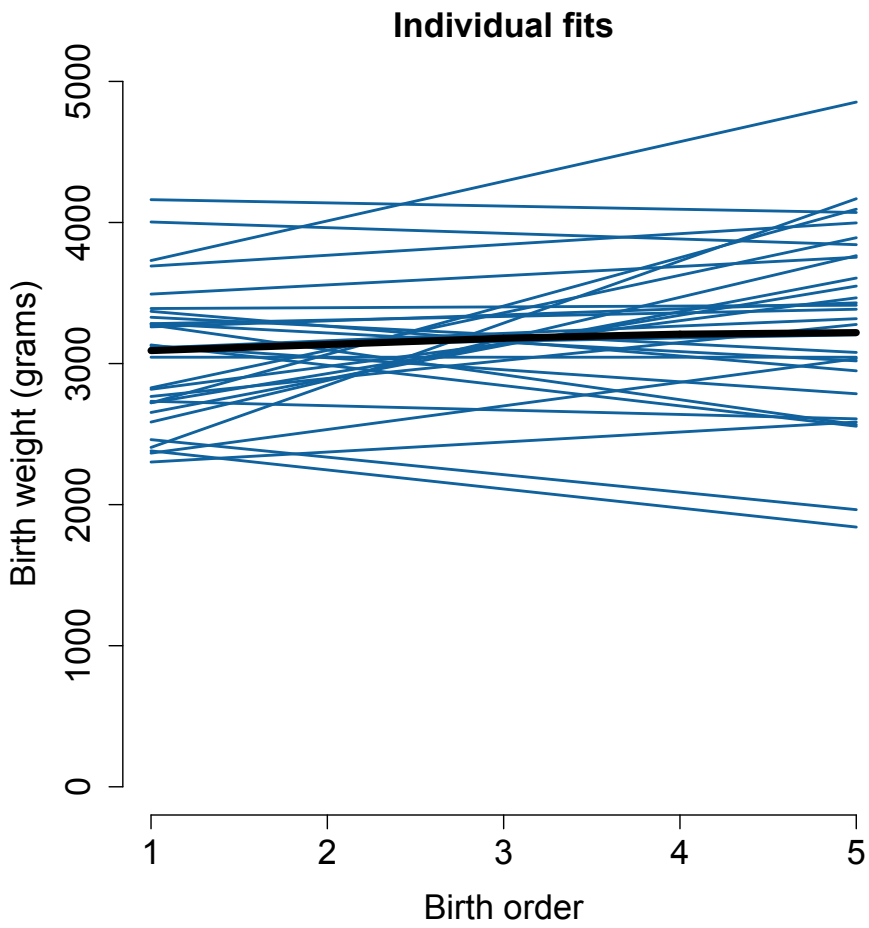
$$\begin{aligned} \mathrm{E}[Y_{ij} \mid x_{ij}, \gamma_i] &= \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \gamma_{0i} \\ \text{or} \quad & \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \gamma_{0i} + \gamma_{1i} x_{ij1} \end{aligned}$$

for $i = 1, \ldots, 200$ and $j = 1, \ldots, 5$ with

- Y_{ij}: Infant birth weight (continuous)
- x_{ij1}: Infant birth order
- x_{ij2}: Mother's initial age

Motivating example





Motivating example: Stata commands

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

* Fit a linear model with random intercepts
xtmixed bweight birthord initage || momid:

* Fit a linear model with random intercepts and slopes
xtmixed bweight birthord initage || momid: birthord
```

Motivating example: Stata output

```
Mixed-effects REML regression                   Number of obs      =      1000
Group variable: momid                           Number of groups   =       200

                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5


                                                Wald chi2(2)       =     30.75
Log restricted-likelihood = -7649.3763          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord |      46.608   9.951014     4.68   0.000     27.10437    66.11163
     initage |   26.73226   9.002678     2.97   0.003      9.08734    44.37719
       _cons |   2526.622   163.3387    15.47   0.000     2206.484     2846.76
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Identity              |
                   sd(_cons) |   358.1759   23.71804      314.5797    407.8139
-----------------------------+------------------------------------------------
                sd(Residual) |   445.0229   11.13253      423.7298     467.386
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   209.20 Prob >= chibar2 = 0.0000
```

Motivating example: Stata output

```
Mixed-effects REML regression                   Number of obs      =      1000
Group variable: momid                           Number of groups   =       200

                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5


                                                Wald chi2(2)       =     29.29
Log restricted-likelihood = -7647.4511          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthord |      46.608   10.41108     4.48   0.000     26.20265    67.01335
     initage |    27.06415   8.899505     3.04   0.002     9.621441    44.50686
       _cons |    2520.799   161.1498    15.64   0.000     2204.952    2836.647
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
momid: Independent           |
                sd(birthord) |   49.35226   13.57685      28.78331    84.62007
                   sd(_cons) |   325.7759    29.6488      272.5532    389.3916
-----------------------------+------------------------------------------------
                sd(Residual) |   438.6626   11.43016      416.8224    461.6472
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =   213.05   Prob > chi2 = 0.0000
```

Motivating example: Comments

- Difference in mean birth weight between two populations of infants whose birth order differs by one is 46.6 grams, 95% CI: (27.0, 66.2)
- Agrees well with GEE point estimate and confidence interval
- $\sqrt{\hat{G}_{00}} = 323$ indicates substantial variability across mothers in the initial level of infant birth weight; $\sqrt{\hat{G}_{11}} = 49$ indicates substantial variability across mothers in the trend of birth weight over time
- **Note**: Typically can specify correlated intercepts and slopes, i.e. $G_{01} \neq 0$, but in this case the model would not converge
- There are options for formal statistical evaluation of two random-effects specifications, but I generally do not recommend an inferential procedure in which a p-value tells you to use a simpler model
- Specification for the random effects should be guided by *a priori* scientific knowledge and exploratory data analysis

GLMM summary

- A GLMM is defined by a random component describing outcomes given subject-specific effects and a systematic component describing the conditional mean given subject-specific effects
- Conditional likelihood for 'fixed effects' eliminates subject-specific effects by conditioning on their sufficient statistics
- Maximum likelihood for 'random effects' integrates over the assumed distribution of the subject-specific effects
- Interpretation of parameter estimates depends on the assumed distribution for the outcome given the subject-specific effects and the specified structure for subject-specific effects

Issues

- GLMM requires that any missing data are missing at random
- Issues arise with time-dependent exposures and covariance weighting

Strategies for analysis of longitudinal data

- **Derived variable**: Collapse the longitudinal series for each subject into a summary statistic, such as a difference (a.k.a. "change score") or regression coefficient, and use methods for independent data
  - **Example**: birth weight of 2nd child – birth weight of 1st child
  - Might be adequate for two time points and no missing data
- **Repeated measures**: Include all data in a regression model for the mean response and account for longitudinal and/or cluster correlation
  - **Generalized estimating equations** (GEE): A marginal model for the mean response and a model for longitudinal or cluster correlation

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

  - **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

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

Overview

Motivation and design

- Gregoire et al. (1996) published the results of an efficacy study on estrogen patches in treating postnatal depression
- 61 women with major depression, which began within 3 months of childbirth and persisted for up to 18 months postnatally, participated in a double-blind, placebo-controlled study
- Women were randomly assigned to active treatment ($n = 34$) or placebo ($n = 27$)
- Participants attended clinics monthly and at each visit self-ratings of depressive symptoms on the Edinburgh postnatal depression scale (EPDS) were measured
- EPDS is a standardized, validated, self-rating scale consisting of 10 items, each rated on a 4-point scale of 0–3
- **Goal**: Investigate the antidepressant efficacy of treatment with estrogen over time

Data

- Depression scores are assessed across $m = 7$ months for the $n = 61$ subjects in the study
- Depression scores for visit j are the longitudinal components measured on subject i

	subj	group	dep0	dep1	dep2	dep3	dep4	dep5	dep6
1.	1	placebo	18	17	18	15	17	14	15
2.	2	placebo	27	26	23	18	17	12	10
3.	3	placebo	16	17	14	.	.	.	.
4.	4	placebo	17	14	23	17	13	12	12
5.	5	placebo	15	12	10	8	4	5	5
6.	6	placebo	20	19	11.54	9	8	6.82	5.05
7.	7	placebo	16	13	13	9	7	8	7
8.	8	placebo	28	26	27	.	.	.	.
9.	9	placebo	28	26	24	19	13.94	11	9
10.	10	placebo	25	9	12	15	12	13	20

- "Wide" form: A row for each subject
- Note that there are some missing data due to drop-out

Exploratory analyses

1. Summarize the depression scores by visit and treatment group
2. Examine within-person correlations among depression scores, graphically and numerically
3. Graph depression scores over time, by treatment group; include a lowess line (smoother) for each group to summarize trends
4. Plot individual trajectories by treatment group

Regression analyses

5. Consider collapsing the longitudinal series for each subject into a summary statistic between the baseline and sixth depression scores. Use methods for independent data to evaluate the association between change in depression scores and estrogen treatment

6. Reshape the data into long form and evaluate longitudinal associations between depression scores and treatment using GEE
   - Use visit as a linear variable
   - Use visit as a categorical variable
   - Evaluate whether the treatment effect varies over time

Reshape the data

Recall what the data look like in wide form

	subj	group	dep0	dep1	dep2	dep3	dep4	dep5	dep6
1.	1	placebo	18	17	18	15	17	14	15
2.	2	placebo	27	26	23	18	17	12	10
3.	3	placebo	16	17	14	.	.	.	.
4.	4	placebo	17	14	23	17	13	12	12
5.	5	placebo	15	12	10	8	4	5	5

For some analyses, reshape the data from wide form to long form

```
. reshape long dep, i(subj) j(visit)
(note: j = 0 1 2 3 4 5 6)

Data                                wide   ->   long
-----------------------------------------------------
Number of obs.                        61   ->      427
Number of variables                    9   ->        4
j variable (7 values)                      ->   visit
xij variables:
                     dep0 dep1 ... dep6    ->   dep
-----------------------------------------------------
```

Reshape the data

"Long" form: A row for each observation

	subj	visit	group	dep
1.	1	0	placebo	18
2.	1	1	placebo	17
3.	1	2	placebo	18
4.	1	3	placebo	15
5.	1	4	placebo	17
6.	1	5	placebo	14
7.	1	6	placebo	15
8.	2	0	placebo	27
9.	2	1	placebo	26
10.	2	2	placebo	23

Answers

Overview

Indonesian Children's Health Study (ICHS)

- Determine the effects of vitamin A deficiency in preschool children
- $n = 275$ children examined for respiratory infection at up to 6 visits
- Xerophthalmia is an ocular manifestation of vitamin A deficiency
- **Goal**: Evaluate association between vitamin A deficiency and risk of respiratory infection

		Age (years)							
Xerophthalmia	Infection	0	1	2	3	4	5	6	7
No	No	77	229	154	196	176	143	65	5
No	Yes	8	30	30	15	9	7	1	0
Yes	No	0	1	9	10	15	8	4	1
Yes	Yes	0	0	4	3	0	0	0	0

Data

```
. list id age time infection xerop gender hfora cost sint
```

	id	age	time	infect~n	xerop	gender	hfora	cost	sint
1.	121013	31	1	0	0	0	-3	-1	0
2.	121013	34	2	0	0	0	-3	0	-1
3.	121013	37	3	0	0	0	-2	1	0
4.	121013	40	4	0	0	0	-2	0	1
5.	121013	43	5	1	0	0	-2	-1	0
6.	121013	46	6	0	0	0	-3	0	-1
7.	121113	-9	1	0	0	1	2	-1	0
8.	121113	-6	2	0	0	1	0	0	-1
9.	121113	-3	3	0	0	1	-1	1	0
10.	121113	0	4	0	0	1	-2	0	1
11.	121113	3	5	1	0	1	-3	-1	0
12.	121113	6	6	0	0	1	-3	0	-1
13.	121114	-26	1	0	0	0	8	-1	0
14.	121114	-23	2	0	0	0	5	0	-1
15.	121114	-20	3	0	0	0	3	1	0
16.	121114	-17	4	1	0	0	0	0	1
17.	121114	-14	5	1	0	0	0	-1	0
18.	121114	-11	6	0	0	0	0	0	-1

Multiple records per person, with age in months, centered at 36 months, and time indicating visit number

Exploratory analyses

1. Plot vitamin A deficiency and infection status, by age, for a sample of individuals
2. Plot percent with respiratory infection versus age, by presence or absence of vitamin A deficiency
3. Explore correlation structure by visit number, and calculate percent with respiratory infection at each visit

Regression analyses

4. Evaluate the association between respiratory infection and vitamin A deficiency using an ordinary logistic regression model
5. Use GEE to estimate the population-averaged odds ratio for respiratory infection, comparing those with vitamin A deficiency to those without, given equivalent values of other covariates. Explore multiple specifications of working correlation
6. Use GLMM to estimate the conditional odds ratio for respiratory infection, comparing a typical individual with vitamin A deficiency to a typical individual without, given equivalent values of other covariates. Estimate the variability in the probability of respiratory infection across individuals

Answers

Overview

Carpal tunnel syndrome trial

- Jarvik et al. (2009) compared surgical versus non-surgical treatment for carpal tunnel syndrome (CTS)
- 116 participants were randomized
- Outcomes were assessed using the CTS assessment questionnaire (CTSAQ), every 3 months for 12 months
  - Primary: functional status (low values are favorable)
  - Secondary: symptom severity
- Crossover to surgery was allowed after 3 months
- **Goal**: Determine whether surgery improves functional status

Data (wide format)

```
. list ID idgroup treatassign surgical ctsaqf0 ctsaqf1 ctsaqf2 ctsaqf3 ctsaqf4
```

	ID	idgroup	treata~n	surgical	ctsaqf0	ctsaqf1	ctsaqf2	ctsaqf3	ctsaqf4
1.	11050	2	0	3	1.888889	1.666667	1.888889	1.333333	2.888889
2.	11068	2	0	0	4	4.111111	4.222222	3.777778	4
3.	11071	2	1	1	2	1.571429	1.222222	1	1
4.	11078	2	0	0	1.375	1.5	2.125	2.5	2.333333
5.	11086	2	1	1	3.222222	2.111111	1	1.777778	1
6.	11087	2	1	1	2.555556	1.333333	1.555556	1.222222	1.222222
7.	11098	2	0	4	2	1.555556	1.444444	1.333333	1
8.	11117	2	1	1	2.875	.	2.888889	.	2
9.	12001	4	1	1	3.125	2.75	3.25	2.75	2.75
10.	12004	4	0	3	3.777778	4.333333	4.555555	3.333333	1.888889
11.	12049	4	1	1	2	1	1	1	1.666667
12.	12068	4	1	0	2.444444	3.333333	2.333333	2.333333	2.444444
13.	12093	4	0	0	2.888889	4.222222	.	3.777778	4.222222
14.	12143	4	1	1	2.888889	1.444444	1	1	1
15.	12153	4	0	1	3	3.25	.	.	2.222222
16.	12177	4	1	1	4.555555	3.777778	.	.	.
17.	13001	3	1	0	2	1.222222	1.111111	1.333333	1
18.	13002	3	1	1	2.333333	1.333333	1.444444	1	1
19.	13005	3	0	1	1.888889	1.666667	1.777778	1.444444	1.555556
20.	13006	3	1	1	3.111111	2.333333	1.777778	2	2

```
--more--
```

Variables

- `ID`: unique participant ID
- `idgroup`: study site
  (1 = private, 2 = UW, 3 = VA, 4 = HMC)
- `age`: age in years
- `gender`
  (0 = male, 1 = female)
- `treatassign`: randomized intervention
  (0 = non-surgery, 1 = surgery)
- `surgreported#`: surgery reported at visit #
  (0 = no, 1 = yes)
- `ctsaqf#`: CTSAQ functional status at visit #
- `ctsaqs#`: CTSAQ symptom severity at visit #
- `surgical`: treated surgically during study
  (0 = never, 1 = 0–3 mos, 2 = 3–6 mos, 3 = 6–9 mos, 4 = 9–12 mos)

Exploratory analyses

1. Plot individual trajectories in CTSAQF over time by treatment
2. Plot average CTSAQF over time by treatment
3. Summarize means, variances, and correlations over time by treatment

Regression analyses (intention-to-treat)

4. Evaluate POST, CHANGE, and ANCOVA models for each timepoint with CTSAQF as outcome
   - POST: follow-up measurement only
   - CHANGE: difference between follow-up and baseline measurement
   - ANCOVA: follow-up measurement controlling for baseline
5. Evaluate difference in mean CTSAQF between treatments at 12 months, adjusted for baseline CTSAQF and study site
6. Using repeated measures regression models, evaluate difference in mean CTSAQF between treatments using all timepoints, adjusted for baseline CTSAQF and study site

Bonus analyses (as treated)

7. Summarize actual treatment patterns by assigned treatment group
8. Plot average CTSAQF by visit. . .
   - For those who received surgery by 3 months versus those who did not
   - For those who received surgery by 9 months versus those who did not
9. Use linear mixed-effects models to estimate differences in mean CTSAQF when exposure is surgery by 3 months and surgery by 9 months instead of assigned treatment group

Answers

Overview

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/ 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.

Thank you!