WORKSHOP ON

APPLIED LONGITUDINAL DATA ANALYSIS

O'Brien Institute for Public Health
University of Calgary
December 3, 2016

GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics
McLean Hospital

Department of Biostatistics
Harvard School of Public Health

ORGANIZATION OF LECTURES

Introduction and Overview

Longitudinal Data: Basic Concepts

Review of Linear Mixed Effects Models

More Flexible Linear Mixed Effects Models:
Smoothing and Semiparametric Regression

Extensions of Generalized Linear Models to Longitudinal Data

- Marginal Models and Generalized Estimating Equations (GEE)
- Generalized Linear Mixed Models (GLMMs)

COURSE AGENDA

9:00am−10:15am	Session I
10:15am−10:30am	*Break*
10:30am−12:00pm	Session II
12:00pm − 1:15pm	*Lunch*
1:15pm − 2:30pm	Session III
2:30pm − 2:45pm	*Break*
2:45pm − 4:15pm	Session IV

APPLIED LONGITUDINAL DATA ANALYSIS

SESSION 1

GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics
McLean Hospital

Department of Biostatistics
Harvard School of Public Health

INTRODUCTION

In recent years, there have been remarkable advances in methods for analyzing longitudinal data.

When the response variable is continuous, familiar linear regression models can be extended to handle the correlated outcomes.

For linear models the correlation among repeated measures can be modelled *explicitly* (e.g., via unrestricted or covariance pattern models) or *implicitly* (e.g., via introduction of random effects).

The latter approach yields a versatile class of regression models for longitudinal data known as *linear mixed effects models* (Session 2).

LONGITUDINAL DATA: BASIC CONCEPTS

Defining feature of longitudinal studies is that measurements of the same individuals are taken repeatedly through time.

Longitudinal studies allow direct study of change over time.

Objective: primary goal is to characterize the change in response over time and the factors that influence change.

With repeated measures, we can capture *within-individual* change.

Complications: (i) repeated measures on individuals are correlated,
(ii) variability is often heterogeneous over time.

Terminology

Individuals/Subjects: Participants in a longitudinal study are referred to as *individuals* or *subjects*.

Occasions: In a longitudinal study individuals are measured repeatedly at different *occasions* or *times*.

The number of repeated observations, and their timing, can vary widely from one longitudinal study to another.

When number and timing of the repeated measurements are the same for all individuals, study design is said to be "**balanced**" over time.

Example 1: *Treatment of Lead-Exposed Children Trial*

- Exposure to lead during infancy is associated with substantial deficits in tests of cognitive ability
- Chelation treatment of children with high lead levels usually requires injections and hospitalization
- A new agent, *Succimer*, can be given orally
- Randomized placebo-controlled trial examining changes in blood lead level during course of treatment
- 100 children randomized to placebo or Succimer
- Measures of blood lead level at baseline, 1, 4 and 6 weeks

Table 1: Mean blood lead levels (and standard deviation) at baseline, week 1, week 4, and week 6.

Group	Baseline	Week 1	Week 4	Week 6
Succimer	26.5	13.5	15.5	20.8
	(5.0)	(7.7)	(7.8)	(9.2)
Placebo	26.3	24.7	24.1	23.2
	(5.0)	(5.5)	(5.7)	(6.2)

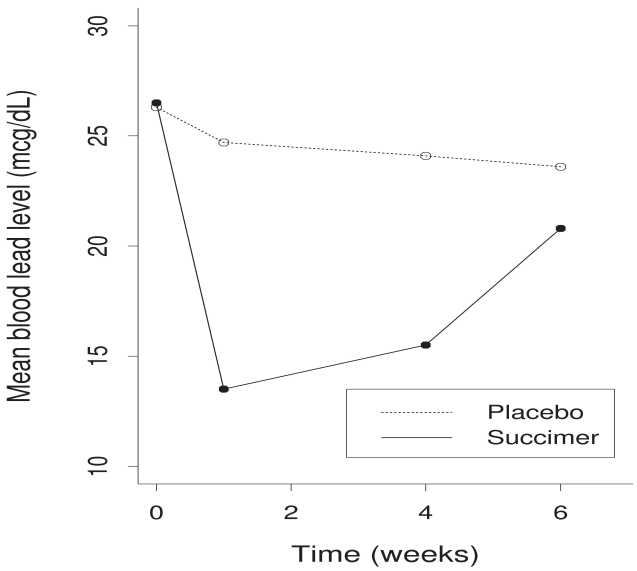


Figure 1: Plot of mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

Example 2: *Influence of Menarche on Changes in Body Fat*

- Prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study.
- At start of study, all the girls were pre-menarcheal and non-obese
- All girls were followed over time according to a schedule of annual measurements until four years after menarche.
- The final measurement was scheduled on the fourth anniversary of their reported date of menarche.
- At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis.

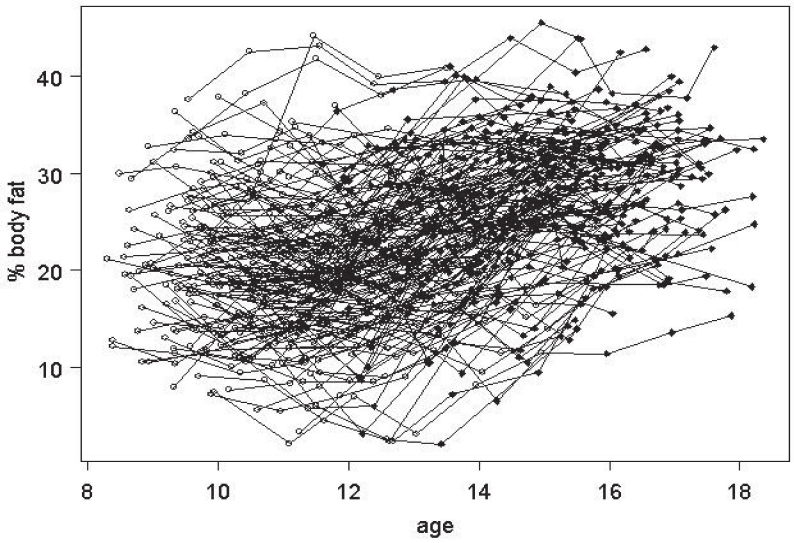


Figure 2: Timeplot of percent body fat against age (in years).

Consider an analysis of the changes in percent body fat before and after menarche.

For the purposes of these analyses "time" is coded as time since menarche and can be positive or negative.

Note: measurement protocol is the same for all girls.

Study design is almost "balanced" if timing of measurement is defined as time since baseline measurement.

It is inherently unbalanced when timing of measurements is defined as time since a girl experienced menarche.

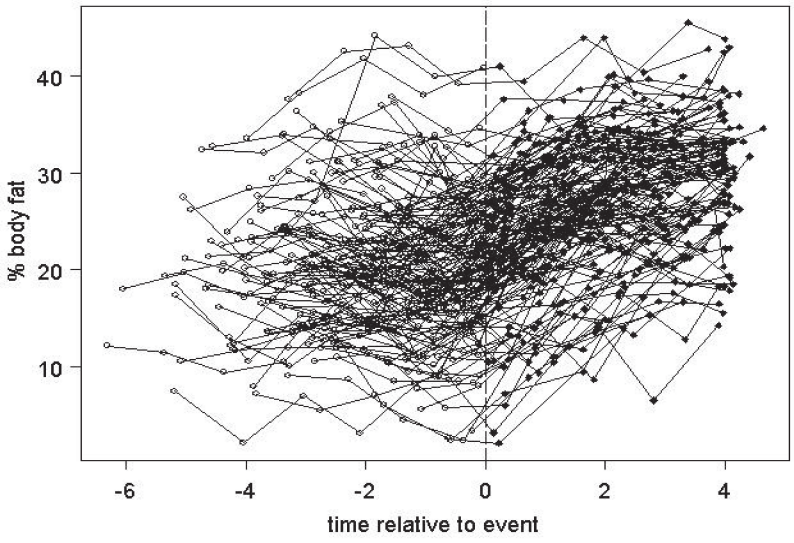


Figure 3: Timeplot of percent body fat against time, relative to age of menarche (in years).

LONGITUDINAL DATA: BASIC CONCEPTS

The primary goal is to characterize the change in response over time and the factors that influence change.

A longitudinal study can estimate change with great precision because each individual acts as his/her own control.

By comparing each individual's responses at two or more occasions, a longitudinal analysis can remove extraneous, but unavoidable, sources of variability among individuals.

This eliminates major sources of variability or "noise" from the estimation of within-individual change.

Notation

Let Y_{ij} denote response variable for i^{th} subject on j^{th} occasion.

Y_{ij} is assumed to be continuous; later (Session 4) we consider cases where Y_{ij} is binary or a count.

We assume there are n_i repeated measurements on the i^{th} subject ($i = 1, ..., N$) and each Y_{ij} is observed at time t_{ij}.

Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates, X_{ij}.

Covariates can be time-invariant (e.g., gender) or time-varying (e.g., time since baseline).

Can group Y_{ij}'s into a $n_i \times 1$ vector Y_i, and X_{ij}'s into a $n_i \times p$ matrix X_i.

Covariance and Correlation

An aspect of longitudinal data that complicates their statistical analysis is that repeated measures on the same individual are usually positively correlated.

This violates the fundamental assumption of independence that is the cornerstone of many statistical techniques.

Next, we define covariance and correlation matrices for longitudinal data.

The *covariance* between responses at two occasions, say Y_{ij} and Y_{ik},

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

where $\mu_j = E(Y_{ij}|X_{ij})$, is a measure of the *linear* dependence between Y_{ij} and Y_{ik}.

The *correlation* between Y_{ij} and Y_{ik} is denoted by

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

where σ_j and σ_k are the standard deviations of Y_{ij} and Y_{ik}.

For the vector of repeated measures, Y_i, we define the variance-covariance matrix, $\Sigma = \text{Cov}(Y_i)$,

$$\text{Cov}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \cdots & \text{Cov}(Y_{i1}, Y_{in}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \cdots & \text{Cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_{in}, Y_{i1}) & \text{Cov}(Y_{in}, Y_{i2}) & \cdots & \text{Var}(Y_{in}) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix},$$

where $\text{Cov}\left(Y_{ij}, Y_{ik}\right) = \sigma_{jk} = \sigma_{kj} = \text{Cov}(Y_{ik}, Y_{ij})$.

We can also define the correlation matrix, denoted by $\text{Corr}(Y_i)$,

$$\text{Corr}(Y_i) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}.$$

This matrix is also symmetric, $\text{Corr}\left(Y_{ij}, Y_{ik}\right) = \rho_{jk} = \rho_{kj} = \text{Corr}(Y_{ik}, Y_{ij})$.

Note: Covariance and correlation matrices are commonly assumed to be homogeneous across individuals.

Example: Treatment of Lead-Exposed Children Trial

We restrict attention to the data from placebo group

Data consist of 4 repeated measurements of blood lead levels obtained at baseline (or week 0), weeks 1, 4, and 6.

The inter-dependence (or time-dependence) among the four repeated measures of blood lead level can be examined by constructing a scatter-plot of each pair of repeated measures.

Examination of the correlations confirms that they are all positive and tend to decrease with increasing time separation.

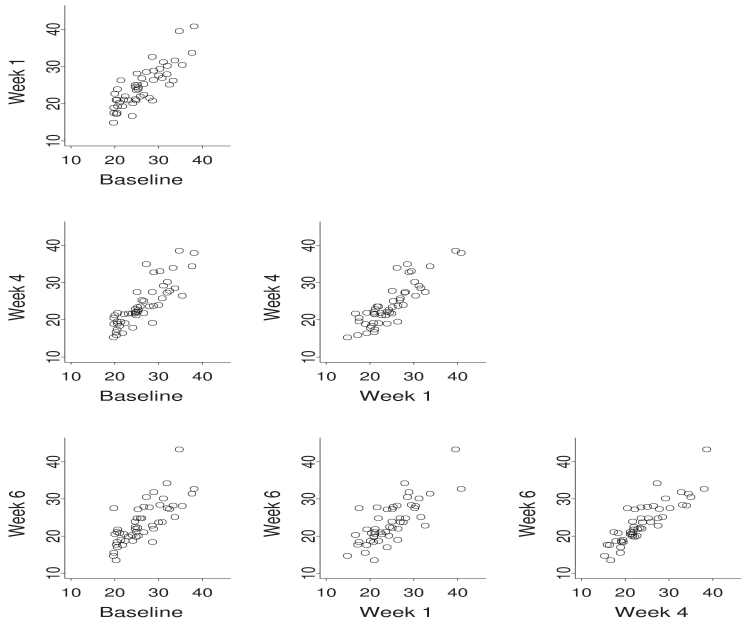


Figure 4: Pairwise scatter-plots of blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group.

Table 2: Estimated covariance matrix for the blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group of the TLC trial.

Covariance Matrix			
25.2	22.8	24.2	18.4
22.8	29.8	27.0	20.5
24.2	27.0	33.0	26.6
18.4	20.5	26.6	38.7

Table 3: Estimated correlation matrix for the blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group of the TLC trial.

Correlation Matrix			
1.00	0.83	0.84	0.59
0.83	1.00	0.86	0.60
0.84	0.86	1.00	0.74
0.59	0.60	0.74	1.00

Some Observations about Correlation in Longitudinal Data

Empirical observations about the nature of the correlation among repeated measures in longitudinal studies:

(i) correlations are positive,

(ii) correlations decrease with increasing time separation,

(iii) correlations between repeated measures rarely ever approach zero, and

(iv) correlation between a pair of repeated measures taken very closely together in time rarely approaches one.

MODELLING LONGITUDINAL DATA

Longitudinal data present two aspects of the data that require modelling:

(i) mean response over time

(ii) covariance

Models for longitudinal data must jointly specify models for the mean and covariance.

Modelling the Mean

Linear (later generalized linear) models widely used to model change in the mean response over time:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + e_{ij}, \quad j = 1, ..., n_i.$$

Modelling the Covariance

Two broad approaches can be distinguished for modelling $\Sigma = \text{Cov}(e_i) = \text{Cov}(Y_i|X_i)$:

(1) Leave Σ unrestricted or assume covariance pattern model (e.g., AR-1)

(2) Random effects covariance structure (Session 2)

Choices for Σ may depend on design: balance, timing and number of repeated occasions, etc.

When design is balanced and number of occasions is small, unrestrictive covariance model is often adopted (no assumptions on how variances and covariances change over time).

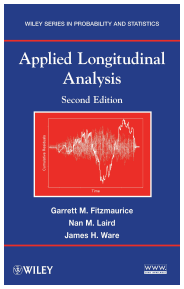
With unbalanced data and/or large number of repeated measures, unrestricted approach is not satisfactory.

Summary of Key Points

- Longitudinal Studies: Individuals are measured repeatedly through time.
- Primary goal of longitudinal study is to measure *change* in response
- Two features of longitudinal data complicate their analysis:
  - repeated measures are positively correlated
  - variability is often heterogeneous over time
- These two features violate fundamental assumptions of linear regression $\Longrightarrow$ Need regression techniqes that can handle correlated data with heterogeneous variability.

FURTHER READING

Fitzmaurice GM, Laird NM & Ware JH (2011). *Applied Longitudinal Analysis, 2nd Ed.* Hoboken, NJ: Wiley. [See Chapters 1, 2]



APPLIED LONGITUDINAL DATA ANALYSIS

SESSION 2

GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics
McLean Hospital

Department of Biostatistics
Harvard School of Public Health

LINEAR MIXED EFFECTS MODELS

Basic idea: Individuals in population are assumed to have their own subject-specific mean response trajectories over time.

Allow subset of the regression parameters to vary randomly from one individual to another, thereby accounting for sources of natural heterogeneity in the population.

Distinctive feature: mean response modelled as a combination of population characteristics (*fixed effects*) assumed to be shared by all individuals, and subject-specific effects (*random effects*) that are unique to a particular individual.

The term *mixed* denotes that model contains both fixed and random effects.

Example: Random Intercept Model

One traditional approach for handling correlation among repeated measures is to assume it arises from a random subject effect,

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij}.$$

In this model response for i^{th} subject at j^{th} occasion is assumed to differ from the population mean,

$$E(Y_{ij}|X_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$$

by a subject effect, b_i, and a within-subject error, ϵ_{ij}.

It is assumed that $b_i \sim N(0, \sigma_b^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, and $b_i \perp\!\!\!\perp \epsilon_{ij}$.

Figure 5 provides graphical representation of linear trend model:

$$Y_{ij} = (\beta_0 + b_i) + \beta_1 t_{ij} + \epsilon_{ij}$$

Overall mean response over time in the population changes linearly with time (denoted by the solid line).

Subject-specific mean responses for two specific individuals, subjects A and B, deviate from the population trend (denoted by the broken lines).

Individual A responds "higher" than the population average and thus has a positive b_i.

Individual B responds "lower" than the population average and has a negative b_i.

Inclusion of errors, ϵ_{ij}, allows response at any occasion to vary randomly above/below subject-specific trajectories (see Figure 6).

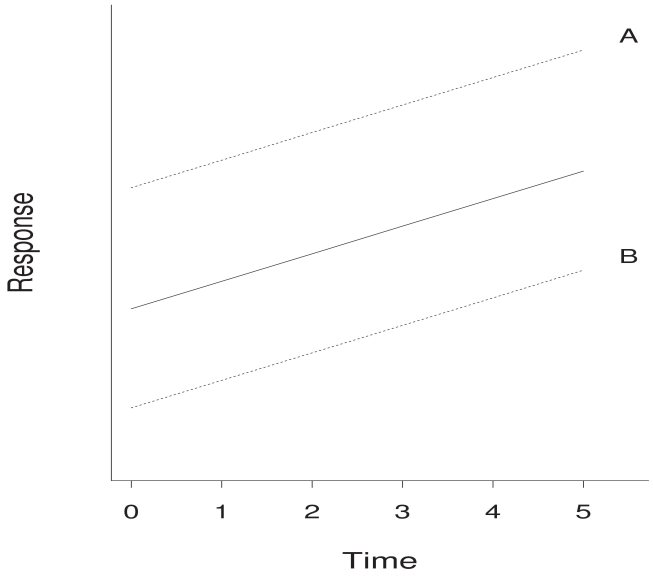


Figure 5: Graphical representation of the overall and subject-specific mean responses over time.

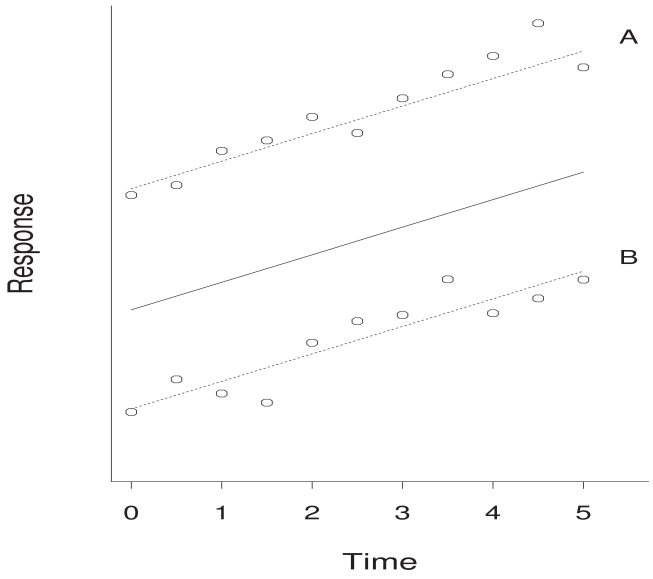


Figure 6: Graphical representation of the overall and subject-specific mean responses over time, plus errors.

Covariance/Correlation Structure

The introduction of a random subject effect induces correlation among the repeated measures.

The following "compound symmetry" covariance structure results:

$$\text{Var}(Y_{ij}|X_{ij}) = \sigma_b^2 + \sigma_\epsilon^2$$

$$\text{Cov}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \sigma_b^2 \Longrightarrow \text{Corr}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}$$

This is the correlation among pairs of observations on the same individual.

Potential Drawback: Variances and correlations are assumed to be constant.

Solution: Allow for heterogeneity in trends over time $\Longrightarrow$ random intercepts and slopes.

Extension: Random Intercept and Slope Model

Consider a model with intercepts and slopes that vary randomly among individuals,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \quad j = 1, ..., n_i,$$

where t_{ij} denotes the timing of the j^{th} response on the i^{th} subject.

This model posits that individuals vary not only in their baseline level of response (when $t_{i1} = 0$), but also in terms of their changes in the response over time (see Figure 7).

The effects of covariates (e.g., due to treatments, exposures) can be incorporated by allowing mean of intercepts and slopes to depend on covariates.

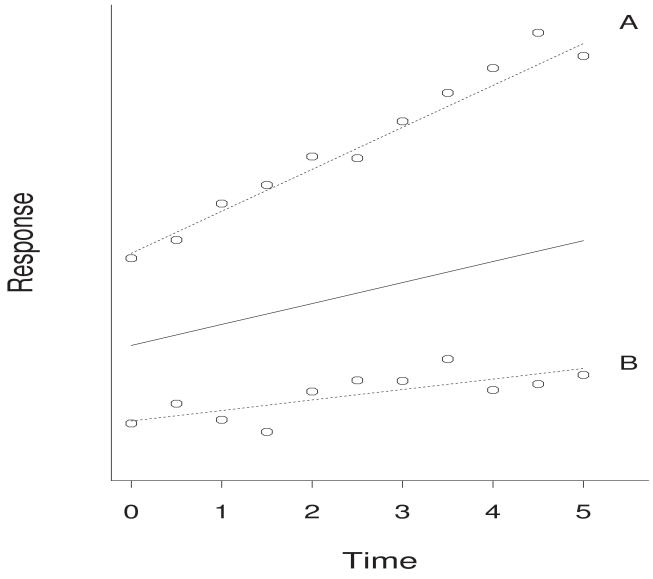


Figure 7: Graphical representation of the overall and subject-specific mean responses over time, plus errors.

For example, consider two-group study comparing a *treatment* and a *control* group:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{trt}_i + \beta_3 t_{ij} \times \text{trt}_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij},$$

where $\text{trt}_i = 1$ if the i^{th} individual assigned to treatment group, and $\text{trt}_i = 0$ otherwise.

The model can be re-expressed as follows for the *control* group and *treatment* group respectively:

trt = 0: $Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij} + \epsilon_{ij},$

trt = 1: $Y_{ij} = (\beta_0 + \beta_2 + b_{0i}) + (\beta_1 + \beta_3 + b_{1i}) t_{ij} + \epsilon_{ij},$

Finally, consider the covariance induced by the introduction of random intercepts and slopes.

Assuming $b_{0i} \sim N(0, \sigma^2_{b_0})$, $b_{1i} \sim N(0, \sigma^2_{b_1})$ (with $\text{Cov}(b_{0i}, b_{1i}) = \sigma_{b_0,b_1}$) and $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$, then

$$\begin{aligned}
\text{Var}\,(Y_{ij}|X_{ij}) &= \text{Var}\,(b_{0i} + b_{1i}t_{ij} + \epsilon_{ij}) \\
&= \text{Var}(b_{0i}) + 2t_{ij}\text{Cov}(b_{0i}, b_{1i}) + t^2_{ij}\text{Var}(b_{1i}) + \text{Var}(\epsilon_{ij}) \\
&= \sigma^2_{b_0} + 2t_{ij}\sigma_{b_0,b_1} + t^2_{ij}\sigma^2_{b_1} + \sigma^2_\epsilon.
\end{aligned}$$

Similarly, it can be shown that

$$\text{Cov}\,(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \sigma^2_{b_0} + (t_{ij} + t_{ik})\,\sigma_{b_0,b_1} + t_{ij}t_{ik}\sigma^2_{b_1}.$$

Thus, in this mixed effects model for longitudinal data the variances and correlations (covariance) are expressed as an explicit function of time, t_{ij}.

Linear Mixed Effects Model

Can allow any subset of the regression parameters to vary randomly.

Using vector notation, the linear mixed effects model can be expressed as

$$Y_{ij} = X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij},$$

where b_i is a $(q \times 1)$ vector of random effects and Z_{ij} is the vector of covariates linking the random effects to Y_{ij}.

Note: Components of Z_{ij} are a subset of the covariates in X_{ij}, e.g., in random intercepts and slopes model $X_{ij} = [1 \;\; t_{ij} \;\; \text{trt}_i \;\; t_{ij} * \text{trt}_{ij}]$ and $Z_{ij} = [1 \;\; t_{ij}]$.

Specifically, any component of β can be allowed to vary randomly by simply including corresponding covariate in Z_{ij}.

The random effects, b_i, are assumed to have a multivariate normal distribution with mean zero and covariance matrix denoted by G.

Prediction of Random Effects

In many applications, inference is focused on fixed effects, $\beta_0, \beta_1, ..., \beta_p$.

However, we can also "estimate" or predict subject-specific effects, b_i (or subject-specific response trajectories over time):

$$\widehat{b_i} = E(b_i|Y_i, X_i; \widehat{\beta}, \widehat{G}, \widehat{\sigma}^2_\epsilon).$$

This is known as "best linear unbiased predictor" (or BLUP).

In general, BLUP "shrinks" predictions towards population-averaged mean.

For example, consider the random intercept model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij},$$

where $\text{Var}\,(b_i) = \sigma^2_b$ and $\text{Var}\,(\epsilon_{ij}) = \sigma^2_\epsilon$.

It can be shown that the BLUP for b_i is:

$$\widehat{b_i} = w \times \left(\frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \mu_{ij}) \right) + (1 - w) \times 0, \text{ where } w = \frac{n_i \sigma^2_b}{n_i \sigma^2_b + \sigma^2_\epsilon}.$$

That is, a weighted-average of zero (mean of b_i) and the mean "residual" for the i^{th} subject.

Note: Less shrinkage (toward zero) when n_i is large and when σ^2_b is large relative to σ^2_ϵ.

Estimation: Maximum Likelihood

ML estimator of $\beta_0, \beta_1, ..., \beta_p$ is the *generalized least squares* (GLS) estimator and depends on covariance among the repeated measures,

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} \left(X_i' \Sigma_i^{-1} X_i \right) \right\}^{-1} \sum_{i=1}^{N} \left(X_i' \Sigma_i^{-1} y_i \right),$$

where $\Sigma_i = \text{Cov}(Y_i)$.

This is a generalization of the ordinary least squares (OLS) estimator used in standard linear regression.

In general, there is no simple expression for ML estimator of the covariance - requires iterative techniques.

Because ML estimation of covariance is known to be biased in small samples, use *restricted* ML (REML) estimation instead.

Case Study: *Influence of Menarche on Changes in Body Fat*

- Prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study.
- At start of study, all the girls were pre-menarcheal and non-obese
- All girls were followed over time according to a schedule of annual measurements until four years after menarche.
- The final measurement was scheduled on the fourth anniversary of their reported date of menarche.
- At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis.

Consider an analysis of the changes in percent body fat before and after menarche.

For the purposes of these analyses "time" is coded as time since menarche and can be positive or negative.

Note: measurement protocol is the same for all girls.

Study design is almost "balanced" if timing of measurement is defined as time since baseline measurement.

It is inherently unbalanced when timing of measurements is defined as time since a girl experienced menarche.

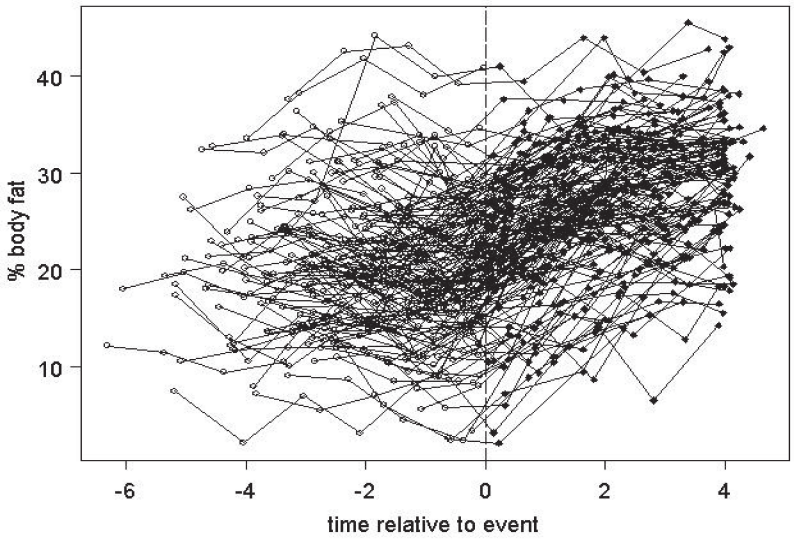


Figure 8: Timeplot of percent body fat against time, relative to age of menarche (in years).

Consider hypothesis that %body fat accretion increases linearly with age, but with different slopes before/after menarche.

We assume that each girl has a piecewise linear spline growth curve with a knot at the time of menarche (see Figure 9).

Each girl's growth curve can be described with an intercept and two slopes, one slope for changes in response before menarche, another slope for changes in response after menarche.

Note: the knot is not a fixed age for all subjects.

Let t_{ij} denote time of the j^{th} measurement on i^{th} subject before or after menarche (i.e., $t_{ij} = 0$ at menarche).

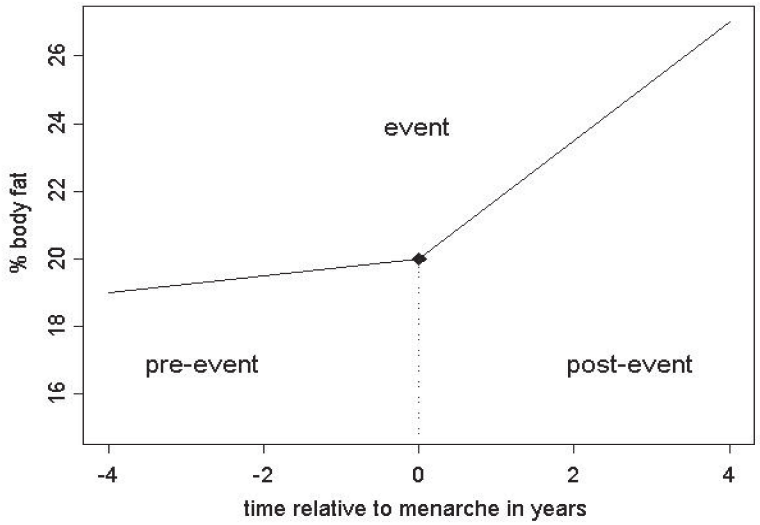


Figure 9: Graphical representation of piecewise linear trajectory.

We consider the following linear mixed effects model

$$E(Y_{ij}|b_i) = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij})_+ + b_{0i} + b_{1i} t_{ij} + b_{2i} (t_{ij})_+,$$

where $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$.

Interpretation of model parameters:

The intercept β_0 is the average %body fat at menarche (when $t_{ij} = 0$).

The slope β_1 is the average rate of change in %body fat (per year) during the pre-menarcheal period.

The average rate of change in %body fat (per year) during the post-menarcheal period is given by $(\beta_1 + \beta_2)$.

Scientific Goal: Assess whether population slopes differ before and after menarche, i.e., $H_0 : \beta_2 = 0$.

Similarly, $(\beta_0 + b_{0i})$ is intercept for i^{th} subject and is the "true" (net of ϵ_{ij}) %body fat at menarche (when $t_{ij} = 0$).

$(\beta_1 + b_{1i})$ is i^{th} subject's slope, or rate of change in %body fat during the pre-menarcheal period.

Finally, the i^{th} subject's slope during the post-menarcheal period is given by $[(\beta_1 + \beta_2) + (b_{1i} + b_{2i})]$.

Interpretation of variance components:

Recall that the subject-specific slopes, $(\beta_1 + b_{1i})$, have mean β_1 and variance $\sigma^2_{b_1}$.

Furthermore, since $b_{1i} \sim N(0, \sigma^2_{b_1})$ this implies that $(\beta_1 + b_{1i}) \sim N(\beta_1, \sigma^2_{b_1})$.

Under the assumption of normality, we expect 95% of the subject-specific slopes, $(\beta_1 + b_{1i})$, to lie between: $\beta_1 \pm 1.96 \times \sigma_{b_1}$.

Variance components for b_{2i} (and b_{0i}) can be interpreted in similar fashion.

Table 4: Estimated regression coefficients (fixed effects) and standard errors for the percent body fat data.

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	21.3614	0.5646	37.84
time	0.4171	0.1572	2.65
$(\text{time})_+$	2.0471	0.2280	8.98

Table 5: Estimated covariance of the random effects and standard errors for the percent body fat data.

PARAMETER	ESTIMATE	SE
$\text{Var}(b_{0i})$	45.9413	5.7393
$\text{Var}(b_{1i})$	1.6311	0.4331
$\text{Var}(b_{2i})$	2.7497	0.9635
$\text{Cov}(b_{0i}, b_{1i})$	2.5263	1.2185
$\text{Cov}(b_{0i}, b_{2i})$	-6.1096	1.8730
$\text{Cov}(b_{1i}, b_{2i})$	-1.7505	0.5980
$\text{Var}(\epsilon_{ij}) = \sigma^2_\epsilon$	9.4732	0.5443

Results

Estimated intercept, $\widehat{\beta}_0 = 21.36$, has interpretation as the average percent body fat at merarche (when $t_{ij} = 0$).

Of note, actual percent body fat at menarche is not observed.

The estimate of the population mean pre-menarcheal slope, β_1, is 0.42, which is statistically significant at the 0.05 level.

This estimated slope is rather shallow and indicates that the annual rate of body fat accretion is less that 0.5%.

The estimate of the population mean post-menarcheal slope, $\beta_1 + \beta_2$, is 2.46 (with SE = 0.12), which is statistically significant at the 0.05 level.

This indicates that annual rate of body fat accretion is approximately 2.5%, almost six times higher than in the pre-menarcheal period.

Based on magnitude of $\widehat{\beta}_2$, relative to its standard error, slopes before and after menarche differ (at the 0.05 level).

Thus, there is evidence that body fat accretion differs before and after menarche.

Estimated variance of b_{1i} is 1.6, indicating substantial variability from girl to girl in rates of fat accretion during the pre-menarcheal period.

For example, approximately 95% of girls have changes in percent body fat between -2.09% and 2.92% $\left(\text{i.e., } 0.42 \pm 1.96 \times \sqrt{1.63}\right)$.

Estimated variance of slopes post-menarche, $\text{Var}(b_{1i} + b_{2i})$, is 0.88 (or [1.63 + 2.75 -2×1.75]), indicating less variability in slopes after menarche.

For example, approximately 95% of girls have changes in percent body fat between 0.62% and 4.30% $\left(\text{i.e., } 2.46 \pm 1.96 \times \sqrt{0.88}\right)$.

Almost all girls expected to have increases post-menarche; substantially fewer (approx. 63%) have increases pre-menarche.

Finally, there is strong positive correlation (approximately 0.8) between annual measurements of percent body fat.

The estimated marginal correlations among annual measurements of percent body fat can be derived from the estimated variances and covariances among the random effects in Table 5.

Strength of correlation declines over time, but does not decay to zero even when measurements are taken 8 years apart (see Table 6).

Table 6: Marginal correlations (below diagonal) among repeated measures of percent body fat between 4 years pre- and post-menarche, with estimated variances along main diagonal.

-4	-3	-2	-1	0	1	2	3	4
61.3								
0.82	54.9							
0.78	0.81	51.8						
0.71	0.76	0.80	52.0					
0.61	0.70	0.76	0.81	55.4				
0.60	0.68	0.74	0.79	0.81	49.1			
0.57	0.64	0.71	0.76	0.78	0.79	44.6		
0.52	0.60	0.66	0.71	0.73	0.76	0.77	41.8	
0.47	0.54	0.60	0.64	0.66	0.70	0.74	0.76	40.8

The mixed effects model can be used to obtain estimates of each girl's growth trajectory over time.

Figure 10 displays estimated population mean growth curve and predicted (empirical BLUP) growth curves for two girls.

Note: two girls differ in the number of measurements obtained (6 and 10 respectively).

A noticeable feature of the predicted growth curves is that there is more shrinkage towards the population mean curve when fewer data points are available.

This becomes more apparent when BLUPs are compared to ordinary least squares (OLS) estimates based only on data from each girl (see Figure 11).

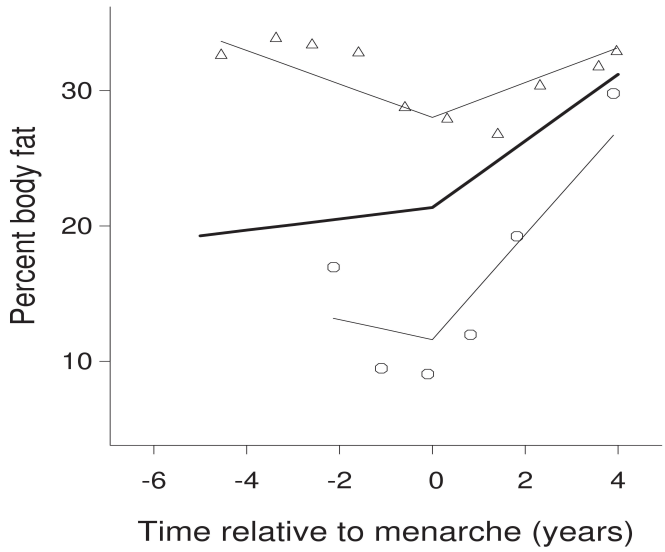


Figure 10: Population average curve and empirical BLUPs for two randomly selected girls.

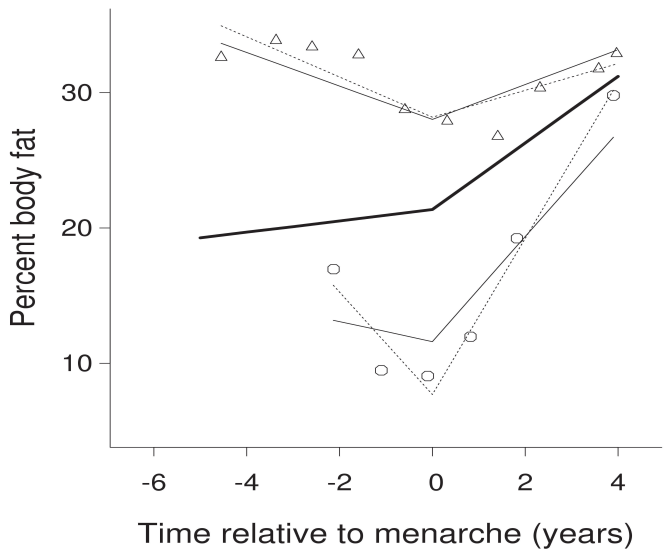


Figure 11: Population average curve, empirical BLUPs, and OLS predictions for two randomly selected girls.

Summary of Key Points

Linear mixed effects models are increasingly used for the analysis of longitudinal data.

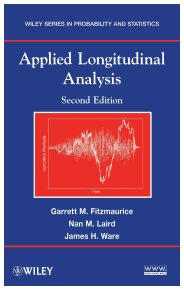
Introduction of random effects accounts for the correlation among repeated measures and allows for heterogeneity of the variance over time.

In general, the random effects covariance structure is relatively parsimonious (e.g., random intercepts and slopes model has only four parameters, $\sigma^2_{b_0}, \sigma^2_{b_1}, \sigma_{b_0,b_1}$, and σ^2_ϵ).

Models can also be used to "estimate" (predict) subject-specific effects.

FURTHER READING

Fitzmaurice GM, Laird NM & Ware JH (2011). *Applied Longitudinal Analysis, 2nd Ed.* Hoboken, NJ: Wiley. [See Chapter 8]



Cnaan A, Slasor P & Laird NM (1997). Tutorial in Biostatistics: Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, **16**, 2349–80.

Naumova EN, Must A & Laird NM (2001). Evaluating the impact of "critical periods" in longitudinal studies of growth using piecewise mixed effects models. *International Journal of Epidemiology*, **30**, 1332–41.

APPLIED LONGITUDINAL DATA ANALYSIS

SESSION 3

GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics
McLean Hospital

Department of Biostatistics
Harvard School of Public Health

Outline

Smoothing Longitudinal Data: Semiparametric Regression

Review of Penalized Splines & Mixed Model Representation

Penalized Splines for Cross-Sectional Data

Extensions of Penalized Splines to Longitudinal Data

Illustration

Linear Mixed Models: Potential Limitations

In earlier session, we discussed linear mixed effects models for longitudinal data.

Models assume shape of the functional relationship between the mean response and covariates is known.

Next, we discuss a simple extension that allows greater flexibility for the form of the relationship.

This extension exploits connection between penalized splines and linear mixed effects models.

Penalized Splines (Cross-Sectional Setting)

To fix ideas, suppose we have a continuous response variable, Y_i, and a single covariate, x_i obtained on N individuals ($i = 1, ..., N$).

Interested in underlying relationship between Y_i and x_i.

Consider the following model,

$$Y_i = \theta(x_i) + e_i,$$

where $\theta(x)$ is an unknown smooth regression function.

The errors, e_i, are assumed to be independent with common variance σ_e^2.

Goal: Estimate the regression function, $\theta(x)$, from the data at hand.

Can estimate $\theta(x)$ based on a piecewise linear function with M knots,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - \kappa_1)_+ + \beta_{12}(x_i - \kappa_2)_+ + \cdots + \beta_{1M}(x_i - \kappa_M)_+ + e_i,$$

where truncated line function $(x_i - \kappa_m)_+ = (x_i - \kappa_m)$ if $(x_i - \kappa_m) > 0$ and is equal to zero otherwise.

Can estimate regression parameters via OLS by minimizing Residual SS,

$$\sum_{i=1}^{N} [Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\beta}_{11}(x_i - \kappa_1)_+ + \widehat{\beta}_{12}(x_i - \kappa_2)_+ + \cdots + \widehat{\beta}_{1M}(x_i - \kappa_M)_+)]^2.$$

Challenge: Choice of number and location of knots.

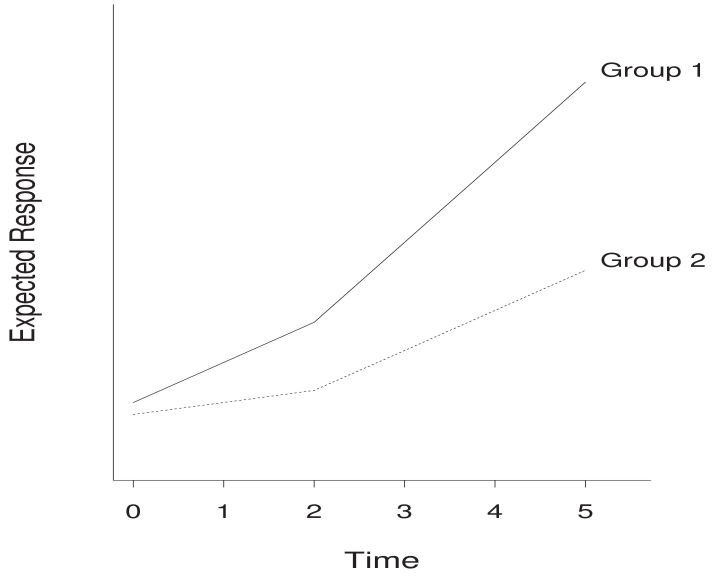


Figure 12: Graphical representation of piecewise linear model for two groups, with single knot at Time = 2.

Penalized Splines: Essential idea is to retain a relatively large number of knots, but *constrain* their influence.

Specifically, estimate the regression parameters by minimizing:

$$\sum_{i=1}^{N}[Y_i - (\beta_0 + \beta_1 x_i + \beta_{11}(x_i - \kappa_1)_+ + \cdots + \beta_{1M}(x_i - \kappa_M)_+)]^2 + \lambda \sum_{m=1}^{M} \beta_{1m}^2.$$

The **roughness penalty**, $\lambda \sum_{m=1}^{M} \beta_{1m}^2$ (for $\lambda \geq 0$), yields a smoother fit to the data depending on the magnitude of λ.

Larger values of λ produce smoother curve.

As $\lambda \longrightarrow 0$, corresponds to no smoothing.

Challenge: How to choose λ?

Let the data determine the degree of smoothing.

Interestingly, if $e_i \sim N(0, \sigma_e^2)$, there is close connection between penalized spline estimator of $\theta(x_i)$ and linear mixed effects models.

Specifically, penalized spline estimator corresponds to REML estimator in an equivalent linear mixed effects model.

Expressing regression function in terms of fixed and random effects,

$$\theta(x_i) = \beta_0 + \beta_1 x_i + \sum_{m=1}^{M} a_m (x_i - \kappa_m)_+,$$

where random effects $a_m \sim N(0, \sigma_a^2)$.

In this mixed model representation,

$$Y_i = \beta_0 + \beta_1 x_i + \sum_{m=1}^{M} a_m (x_i - \kappa_m)_+ + e_i,$$

the coefficients for the truncated line functions $(x_i - \kappa_m)_+$ are the random effects, and can be shown that $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$.

This mixed model representation is useful because it suggests natural ways to estimate λ and to extend penalized splines to the longitudinal setting.

Aside

Note we have assumed a *linear* spline model for $\theta(x)$.

We can consider *polynomial* spline models of any order. For example, a *cubic* spline model with knots at $\kappa_1, ..., \kappa_M$ is given by

$$\theta(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{m=1}^{M} \beta_{3m}(x_i - \kappa_m)_+^3,$$

where $(x_i - \kappa_m)_+^3 = [(x_i - \kappa)_+]^3$.

Cubic spine models are a common choice among polynomial splines.
(Note: B-splines provide more numerically stable basis for cubic splines).

For ease of exposition, focus only on *linear* spline models.

Penalized Splines for Longitudinal Data

Basic Idea: Use mixed model representation of penalized splines, but include additional random subject effects to account for the correlation among the repeated measures.

Consider piecewise linear function of time with M knots, $\kappa_1, ..., \kappa_M$, and randomly varying subject effect,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_{11}(t_{ij} - \kappa_1)_+ + \cdots + \beta_{1M}(t_{ij} - \kappa_M)_+ \; + \; b_i + \epsilon_{ij},$$

where $b_i \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

Adding "roughness penalty", $\lambda \sum_{m=1}^{M} \beta_{1m}^2$, mixed model representation is

$$\theta(t_{ij}) + b_i = \beta_0 + \beta_1 t_{ij} + \sum_{m=1}^{M} a_m(t_{ij} - \kappa_m)_+ + b_i, \text{ where } a_m \sim N(0, \sigma_a^2).$$

Random effects, b_i and a_m, have different roles.

The a_m are coefficients for the truncated line functions, $(t_{ij} - \kappa_m)_+$, and produce a smooth regression function, $\theta(t_{ij})$.

Inclusion of b_i allows each individual to have her own piecewise linear curve; marginally, induces correlation among the repeated measures.

Can allow for more general patterns of covariance through inclusion of additional random effects, e.g.,

$$\theta(t_{ij}) + b_{0i} + b_{1i}t_{ij} = \beta_0 + \beta_1 t_{ij} + \sum_{m=1}^{M} a_m(t_{ij} - \kappa_m)_+ + b_{0i} + b_{1i}t_{ij},$$

allows for heterogeneous variances and correlations that depend on t_{ij}.

Linear mixed model representation allows great flexibility.

In many longitudinal analyses, functional form of the mean time trend cannot be settled beforehand.

In two group setting, can incorporate group effect in a parametric fashion and time trends in a highly non-linear, but not predetermined, way.

Consider the following model,

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \beta_2 \text{Group}_i + \beta_3 \text{Group}_i \times t_{ij} \\ &\quad + \textstyle\sum_{m=1}^{M} a_m(t_{ij} - \kappa_m)_+ + b_{0i} + b_{1i}t_{ij} + \epsilon_{ij}, \end{aligned}$$

where Group = 1 for active treatment (or exposure) and 0 otherwise.

This model allows a very general spline curve for Group = 0,

$$E(Y_{ij}|\text{Group}_i = 0) = \beta_0 + \beta_1 t_{ij} + \sum_{m=1}^{M} a_m (t_{ij} - \kappa_m)_+,$$

but constrains the *differences* between the smooth curves for the two groups to be a simple linear function of time,

$$E(Y_{ij}|\text{Group}_i = 1) - E(Y_{ij}|\text{Group}_i = 0) = \beta_2 + \beta_3 t_{ij}.$$

Here, β_3 is the constant rate of change (over time) in the *differences* between the smooth curves for the mean response in the two groups.

Assumed *linearity* refers to the hypothesized pattern of *differences* between the group means, not to the time trends.

Case Study: *Repeated measures of progesterone metabolite concentration during menstrual cycle*

- Longitudinal data on PdG measured daily in urine from day −8 to day 15 in the menstrual cycle (day 0 denotes ovulation day)
- A sample of 22 conceptive cycles from 22 women and 29 non-conceptive cycles from another 29 women
- Goal is to describe and compare the mean hormone profiles in the conceptive and non-conceptive groups
- Figures 13(a) and 13(b) display time plots for the conceptive and non-conceptive groups

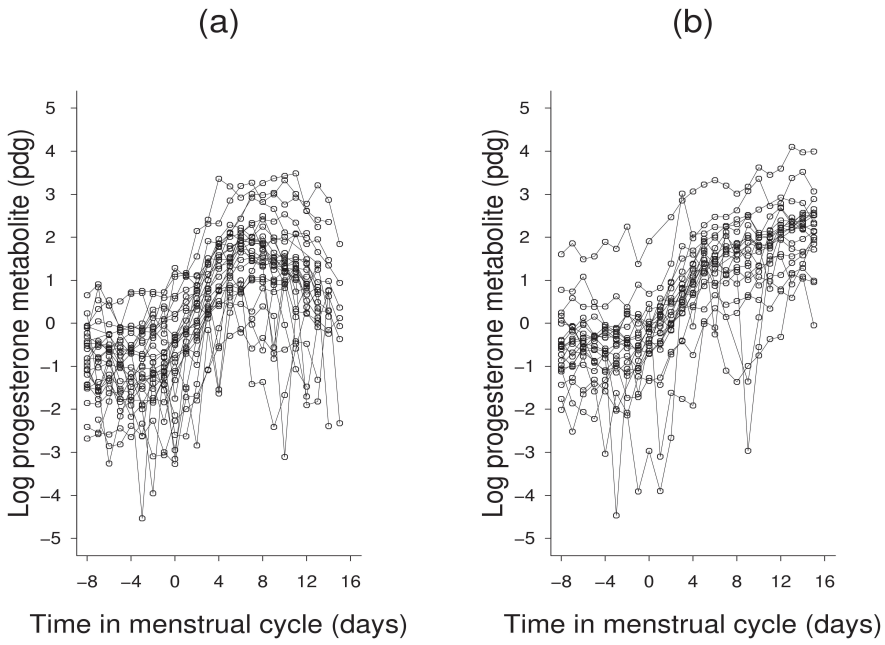


Figure 13: Time plots, with joined line segments, of log progesterone concentration versus days of menstrual cycle for (a) women in non-conceptive group, and (b) women in conceptive group.

Fit separate penalized splines to data for conceptive and non-conceptive groups.

Using mixed model representation, and with 22 knots located consecutively from days -7 through 14,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \sum_{m=1}^{22} a_m (t_{ij} - \kappa_m)_+ + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij},$$

$a_m \sim N(0, \sigma_a^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, and

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b0}^2 & \sigma_{b0,b1} \\ \sigma_{b0,b1} & \sigma_{b1}^2 \end{bmatrix} \right).$$

Fitted functions for the two groups obtained by combining REML estimates of β and BLUP predictions of a_m.

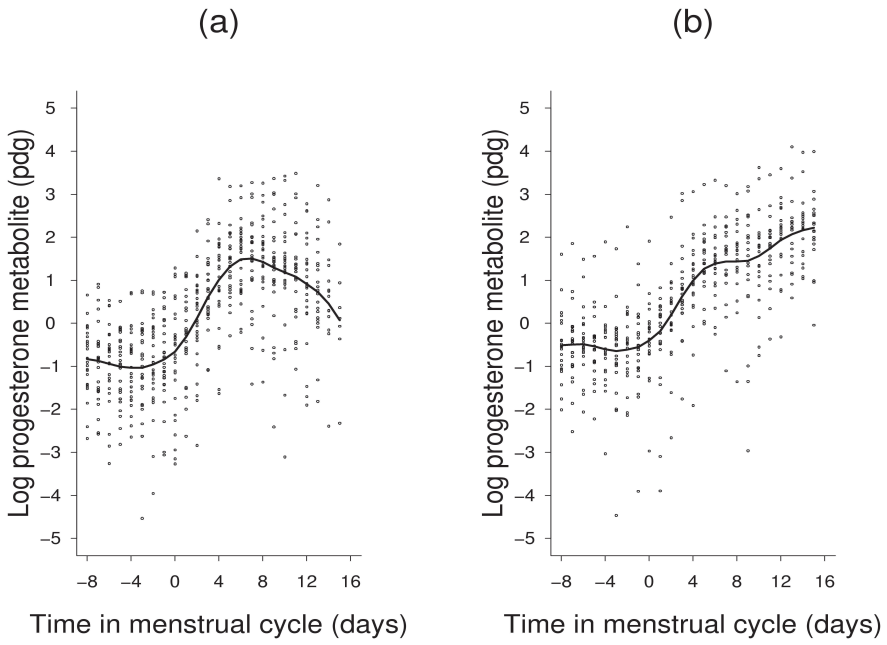


Figure 14: Time plots, with fitted penalized spline superimposed, of log progesterone concentration versus days of menstrual cycle for (a) women in non-conceptive group, and (b) women in conceptive group.

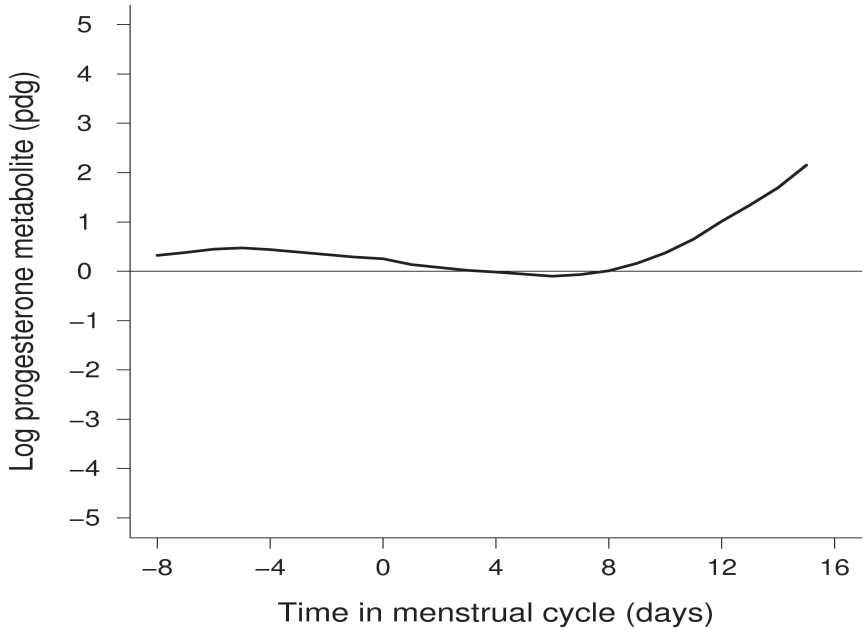


Figure 15: Plot of group differences (conceptive versus non-conceptive group) in fitted penalized splines for log progesterone concentration during menstrual cycle.

Figures 14(a) and 14(b) suggest discernible differences between groups following implantation (approx. day 7).

Figure 15 displays group differences (conceptive minus non-conceptive) in mean log progesterone concentration over time.

Plot suggests large differences between the groups after implantation.

Plot suggests *differences* can be represented by a piecewise linear trend with single knot at day 7.

Fit following semi-parametric regression model for combined data,

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \beta_2 \text{Group} + \beta_3 \text{Group} \times t_{ij} + \beta_4 \text{Group} \times (t_{ij} - 7)_+ \\ &\quad + \textstyle\sum_{m=1}^{22} a_m (t_{ij} - \kappa_m)_+ + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \end{aligned}$$

where Group = 1 for the conceptive group and 0 otherwise.

Table 7: REML estimates and SEs from mixed model representation of semiparametric regression model for log progesterone concentration.

Variable	Estimate	SE	Z
Intercept	−0.8132	0.5656	−1.44
Time	0.0162	0.0746	0.22
Group	0.1625	0.2281	0.71
Group × Time	−0.0479	0.0118	−4.05
Group × $(\text{Time} - 7)_+$	0.2961	0.0232	12.75
$\text{Var}(a_m)$	0.0155	0.0069	
$\text{Var}(b_{0i})$	0.6337	0.1287	
$\text{Var}(b_{1i})$	0.0010	0.0003	
$\text{Cov}(b_{0i}, b_{1i})$	0.0039	0.0042	
$\text{Var}(\epsilon_{ij})$	0.2856	0.0127	

Results indicate that initial mean difference between two groups has declined significantly prior to implantation (with estimated slope, $\widehat{\beta}_3 = -0.0479, Z = -4.05, p < 0.0001$).

By day 7, there is no significant difference between the two groups in mean log progesterone concentration ($\widehat{\beta}_2 + 7 * \widehat{\beta}_3 = 0.1625 - 7 * 0.0479 = -0.1728, SE = 0.2514, Z = -0.69, p > 0.45$).

Thereafter, the trajectories of mean log progesterone concentration for the two groups depart significantly (with estimated slope, $\widehat{\beta}_3 + \widehat{\beta}_4 = -0.0479 + 0.2961 = 0.2482, SE = 0.0203, Z = 12.3, p < 0.0001$).

Figure 16 shows plot of estimated *differences* between two groups (conceptive versus non-conceptive group), and 95% pointwise confidence limits.

Strongest evidence for differences between groups is during days 10-15.

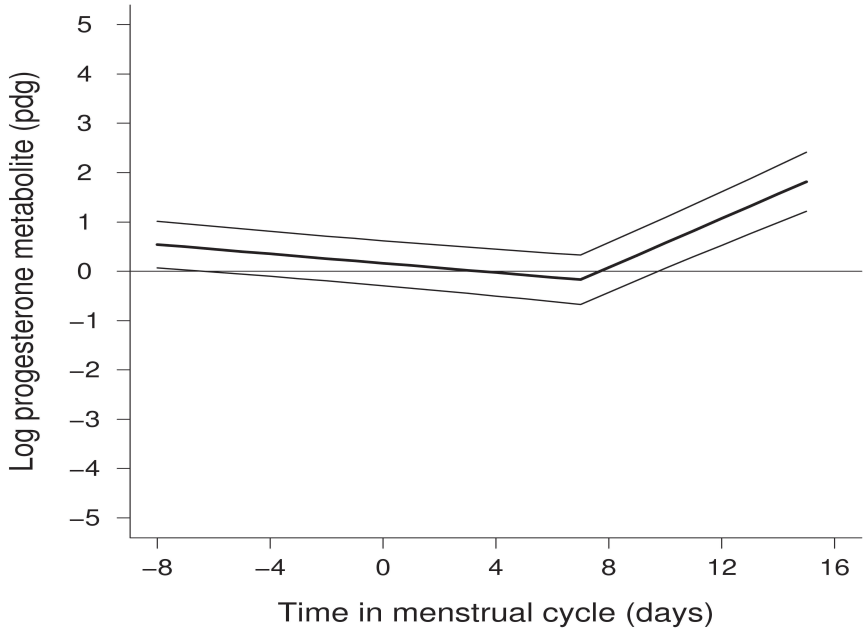


Figure 16: Plot of fitted group differences (non-conceptive versus conceptive group), and 95% confidence limits, from semiparametric regression model.

Summary of Key Points

Linear mixed effects models have become established methods for longitudinal analyses.

Assume shape of the functional relationship between the mean response and covariates is known.

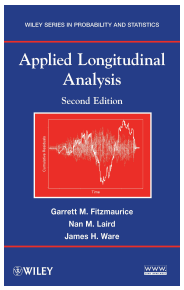
Penalized splines allow greater flexibility for the form of the relationship.

Mixed model representation of penalized splines makes this extension straightforward.

Further extensions: Subject-specific smooth curves for longitudinal data, extensions to generalized linear mixed models for longitudinal data...

Further Reading

Fitzmaurice GM, Laird NM & Ware JH (2011). *Applied Longitudinal Analysis, 2nd Ed.* Hoboken, NJ: Wiley. [See Chapter 19]



Ruppert D, Wand MP & Carroll RJ (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

Gurrin LC, Scurrah KJ & Hazelton ML (2005). Tutorial in biostatistics: Spline smoothing with linear mixed models. *Statistics in Medicine*, 24, 3361-3381.

APPLIED LONGITUDINAL DATA ANALYSIS

SESSION 4

GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics
McLean Hospital

Department of Biostatistics
Harvard School of Public Health

Extensions of Generalized Linear Models to Longitudinal Data (Part 1)

When the response variable is categorical (e.g., binary, ordinal and count data), generalized linear models (e.g., logistic regression) can be extended to handle the correlated outcomes.

However, non-linear transformations of the mean response (e.g., logit) raise additional issues concerning the interpretation of the regression coefficients.

Different approaches for accounting for the correlation lead to models having regression coefficients with distinct interpretations.

As we will see, different models for discrete longitudinal data have somewhat different targets of inference.

MOTIVATING EXAMPLE

Oral Treatment of Toenail Infection

Randomized, double-blind, parallel-group, multicenter study of 294 patients comparing 2 oral treatments (denoted A and B) for toe-nail infection.

Outcome variable: Binary variable indicating presence of onycholysis (separation of the nail plate from the nail bed).

Patients evaluated for degree of onycholysis (separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4, 8, 12, 24, 36, and 48.

Interested in the rate of decline of the proportion of patients with onycholysis over time and the effects of treatment on that rate.

MOTIVATING EXAMPLE

Clinical trial of anti-epileptic drug progabide
(Thall and Vail, *Biometrics*, 1990)

Randomized, placebo-controlled study of treatment of epileptic seizures with progabide.

Patients were randomized to treatment with progabide, or to placebo in addition to standard therapy.

Outcome variable: Count of number of seizures

Measurement schedule: Baseline measurement during 8 weeks prior to randomization. Four measurements during consecutive two-week intervals.

Sample size: 28 epileptics on placebo; 31 epileptics on progabide

REVIEW OF GENERALIZED LINEAR MODELS

Generalized linear models are a class of regression models; they include the standard linear regression model but also many other important models:

- Linear regression for continuous data
- Logistic regression for binary data
- Loglinear models for count data

Generalized linear models extend the methods of regression analysis to settings where the outcome variable can be categorical.

In this short course, we consider extensions of generalized linear models to longitudinal data.

Notation for Generalized Linear Models

Assume N independent observations of a single response variable, Y_i.

Associated with each response, Y_i, there are p covariates, $X_{i1}, ..., X_{ip}$.

Goal: Primarily interested in relating the mean of Y_i,
$\mu_i = E(Y_i|X_{i1}, ..., X_{ip})$, to the covariates.

In generalized linear models:

(i) the distribution of the response is assumed to belong to a family of distributions known as the exponential family, e.g., normal, Bernoulli, binomial, and Poisson distributions.

(ii) A transformation of the mean response, μ_i, is then linearly related to the covariates, via an appropriate link function:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where link function $g(\cdot)$ is a known function, e.g., $\log(\mu_i)$ or $\text{logit}(\mu_i)$.

This implies that it is the transformed mean response that changes linearly with changes in the values of the covariates.

Mean and Variance of Exponential Family Distributions

Exponential family distributions share some common statistical properties.

The variance of Y_i can be expressed in terms of

$$\text{Var}\,(Y_i|X_{i1}, ..., X_{ip}) = \phi\, v(\mu_i),$$

where the scale parameter $\phi > 0$.

The variance function, $v(\mu_i)$, describes how the variance of the response is functionally related to μ_i, the mean of Y_i.

Canonical link and variance functions for the normal, Bernoulli, and Poisson distributions.

Distribution	**Var. Function,** $v(\mu)$	**Canonical Link**
Normal	$v(\mu) = 1$	Identity: $\mu = \eta$
Bernoulli	$v(\mu) = \mu(1-\mu)$	Logit: $\log\left[\frac{\mu}{(1-\mu)}\right] = \eta$
Poisson	$v(\mu) = \mu$	Log: $\log(\mu) = \eta$

where $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.

GENERALIZED LINEAR MODELS FOR LONGITUDINAL DATA

Next, we focus on two general approaches for analyzing longitudinal responses:

1. Marginal Models

2. Generalized Linear Mixed Models

These approaches can be considered extensions of generalized linear models to correlated data.

The main emphasis will be on discrete response data, e.g., count data or binary responses.

MARGINAL MODELS

The basic premise of marginal models is to make inferences about population averages.

The term 'marginal' is used here to emphasize that the mean response modelled is conditional only on covariates and not on other responses or random effects.

A feature of marginal models is that the models for the mean and the 'within-subject association' (e.g., covariance) are specified separately.

Notation

Let Y_{ij} denote response variable for i^{th} subject on j^{th} occasion.

Y_{ij} can be continuous, binary, or a count.

We assume there are n_i repeated measurements on the i^{th} subject and each Y_{ij} is observed at time t_{ij}.

Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates, X_{ij}.

Covariates can be time-invariant (e.g., gender) or time-varying (e.g., time since baseline).

Three-Part Specification of Marginal Models

1. The marginal expectation of the response, μ_{ij}, depends on covariates through a known link function

$$g(\mu_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij}.$$

2. The marginal variance of Y_{ij} depends on the marginal mean according to

$$\text{Var}(Y_{ij}|X_{ij}) = \phi\, \upsilon(\mu_{ij})$$

   where $\upsilon(\mu_{ij})$ is a known 'variance function' and ϕ is a scale parameter that may be fixed and known or may need to be estimated.
   Note: For continuous response, can allow $\text{Var}(Y_{ij}|X_{ij}) = \phi_j\, \upsilon(\mu_{ij})$.

3. The 'within-subject association' among the responses is a function of the means and of additional parameters, say $\boldsymbol{\alpha}$, that may also need to be estimated.

For example, when α represents pairwise correlations among responses, the covariances among the responses depend on $\mu_{ij}(\beta)$, ϕ, and α:

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) &= \text{s.d.}(Y_{ij})\,\text{Corr}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik})\,\text{s.d.}(Y_{ik}) \\ &= \sqrt{\phi\, v(\mu_{ij})}\,\text{Corr}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik})\,\sqrt{\phi\, v(\mu_{ik})}\end{aligned}$$

where s.d.(Y_{ij}) is the standard deviation of Y_{ij}.

In principle, can also specify higher-order moments.

Examples of Marginal Models

Example 1. Binary responses:

1. $\text{Logit}\,(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$.
   (i.e., logistic regression)

2. $\text{Var}\,(Y_{ij}|X_{ij}) = \mu_{ij}\,(1 - \mu_{ij})$
   (i.e., Bernoulli variance)

3. $\text{Log OR}\,(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha_{jk}$
   (i.e., unstructured log odds ratios)
   where

$$\text{OR}\,(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) =$$

$$\frac{\Pr(Y_{ij}=1, Y_{ik}=1|X_{ij}, X_{ik})\Pr(Y_{ij}=0, Y_{ik}=0|X_{ij}, X_{ik})}{\Pr(Y_{ij}=1, Y_{ik}=0|X_{ij}, X_{ik})\Pr(Y_{ij}=0, Y_{ik}=1|X_{ij}, X_{ik})}.$$

Example 2. Count data:

1. $\text{Log}\,(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$.
   (i.e., Poisson regression)

2. $\text{Var}\,(Y_{ij}|X_{ij}) = \phi\,\mu_{ij}$
   (i.e., extra-Poisson variance or "overdispersion" when $\phi > 1$)

3. $\text{Corr}\,(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha_{jk}$
   (i.e., unstructured correlation)

Interpretation of Marginal Model Parameters

The regression parameters, $\boldsymbol{\beta}$, have 'population-averaged' interpretations ('averaging' over individuals within subgroups of population).

For example, consider the following logistic model,

$$\text{logit}(\mu_{ij}) = \text{logit}\{E(Y_{ij}|X_{ij})\} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}.$$

Interpretation of any component of β, say β_k, is in terms of adjusted changes in the transformed mean (or "population-averaged") response for a unit change in the corresponding covariate, say X_{ijk}.

When X_{ijk} takes on some value x, the log odds of a positive response is,

$$\log\left\{\frac{\Pr(Y_{ij}=1|X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}{\Pr(Y_{ij}=0|X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}\right\} =$$

$$\beta_0 + \beta_1 X_{ij1} + \cdots + \beta_k x + \cdots + \beta_p X_{ijp}.$$

Similarly, when X_{ijk} now takes on some value $x+1$,

$$\log\left\{\frac{\Pr(Y_{ij}=1|X_{ij1},\ldots,X_{ijk}=x+1,\ldots,X_{ijp})}{\Pr(Y_{ij}=0|X_{ij1},\ldots,X_{ijk}=x+1,\ldots,X_{ijp})}\right\} =$$

$$\beta_0 + \beta_1 X_{ij1} + \cdots + \beta_k (x+1) + \cdots + \beta_p X_{ijp}.$$

$\longrightarrow \beta_k$ is adjusted change in log odds for unit change in X_{ijk}.

Statistical Inference for Marginal Models

Maximum Likelihood (ML):

Unfortunately, with discrete response data there is no simple analogue of the multivariate normal distribution.

In the absence of a "convenient" likelihood function for discrete data, there is no unified likelihood-based approach for marginal models.

Alternative approach to estimation - *Generalized Estimating Equations* (GEE).

GENERALIZED ESTIMATING EQUATIONS

The GEE estimator of β solves the following *generalized estimating equations*

$$\sum_{i=1}^{N} D_i' V_i^{-1} (y_i - \mu_i) = 0,$$

where $D_i = \partial \mu_i / \partial \beta$ is the "derivative" matrix and V_i is the so-called "working" covariance matrix.

By "working" covariance matrix we mean that V_i approximates the true underlying covariance matrix for the vector of responses Y_i.

That is, $V_i \approx \text{Cov}(Y_i|X_i)$, recognizing that $V_i \neq \text{Cov}(Y_i|X_i)$ unless the models for the variances and the within-subject associations are correct.

Properties of GEE estimators

1. In many cases $\widehat{\beta}$ is almost efficient when compared to MLE.
   For example, GEE has same form as likelihood equations for multivariate normal models and also certain models for discrete data

2. $\widehat{\beta}$ is consistent even if the covariance of Y_i has been misspecified

3. Standard errors for $\widehat{\beta}$ can be obtained using so-called 'sandwich' estimator, $\text{Cov}(\widehat{\beta}) = B^{-1}MB^{-1}$, where

$$B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i, \quad M = \sum_{i=1}^{N} D_i' V_i^{-1} \text{Cov}\left(Y_i | X_i\right) V_i^{-1} D_i.$$

   B and M can be estimated by replacing α, ϕ, and $\boldsymbol{\beta}$ by their estimates, and replacing $\text{Cov}\left(Y_i|X_i\right)$ by $\left(Y_i - \widehat{\mu}_i\right)\left(Y_i - \widehat{\mu}_i\right)'$.

Case Study: *Oral Treatment of Toenail Infection*

Randomized, double-blind, parallel-group, multicenter study of 294 patients comparing 2 oral treatments (denoted A and B) for toe-nail infection.

Outcome variable: Binary variable indicating presence of onycholysis (separation of the nail plate from the nail bed).

Patients evaluated for degree of onycholysis (separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4, 8, 12, 24, 36, and 48.

Interested in the rate of decline of the proportion of patients with onycholysis over time and the effects of treatment on that rate.

Assume that the marginal probability of onycholysis follows a logistic model,

$$\text{logit}\{E(Y_{ij}|X_{ij})\} = \beta_0 + \beta_1 \text{Month}_{ij} + \beta_2 \text{Trt}_i * \text{Month}_{ij}$$

where $Trt = 1$ if treatment group B and 0 otherwise.

Here, we assume that $\text{Var}(Y_{ij}|X_{ij}) = \mu_{ij}(1 - \mu_{ij})$.

We also assume an unstructured correlation for the within-subject association (i.e., estimate all possible pairwise correlations).

Table 8: GEE estimates and standard errors (empirical) from marginal logistic regression model for onycholysis data.

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	-0.698	0.122	-5.74
Month	-0.140	0.026	-5.36
Trt × Month	-0.081	0.042	-1.94

Results

From the output above, we would conclude that:

1. There is a suggestion of a difference in the rate of decline in the two treatment groups (P = 0.052).

2. Over 12 months, the odds of onycholysis has decreased by a factor of 0.19 [exp(-0.14*12)] in treatment group A.

3. Over 12 months, the odds of onycholysis has decreased by a factor of 0.07 [exp(-0.221*12)] in treatment group B.

4. Odds ratio comparing 12 month decreases in risk of onycholysis between treatments A and B is approx 2.6 (or $e^{12*0.081}$).

5. Overall, there is a significant decline over time in the prevalence of onycholysis for all randomized patients.

Extensions of Generalized Linear Models to Longitudinal Data (Part 2)

So far, we have discussed *marginal models* for longitudinal data.

Next, we consider a second type of extension, *generalized linear mixed models* (GLMMs).

We describe how these models extend the conceptual approach represented by the linear mixed effects model (Session 2).

The basic premise is that we assume natural heterogeneity across individuals in a subset of the regression coefficients via the introduction of random effects.

Generalized Linear Mixed Models

The generalized linear mixed model can be considered in two steps:

First Step: Assumes that the conditional distribution of each Y_{ij}, given individual-specific effects b_i, belongs to the exponential family with conditional mean,

$$g\{E(Y_{ij}|X_{ij}, b_i)\} = X'_{ij}\boldsymbol{\beta} + Z'_{ij}b_i$$

where $g(\cdot)$ is a known link function and Z_{ij} is a known design vector, a subset of X_{ij}, linking the random effects b_i to Y_{ij}.

The particular subset of the regression parameters β that vary randomly is determined by components of X_{ij} that comprise Z_{ij}.

Second-Step: The b_i are assumed to vary independently from one individual to another and $b_i \sim N(0, G)$.

Here, G is the covariance matrix for the random effects.

Note: There is an additional assumption of 'conditional independence'.

That is, given b_i, the responses $Y_{i1}, Y_{i2}, ..., Y_{in_i}$ are assumed to be mutually independent.

Example 1:

Binary logistic model with random intercepts:

$$\text{logit}\{E(Y_{ij}|X_{ij}, b_i)\} = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + b_i$$

$\text{Var}(Y_{ij}|X_{ij}, b_i) = E(Y_{ij}|X_{ij}, b_i)\{1 - E(Y_{ij}|X_{ij}, b_i)\}$ (Bernoulli variance),

and $b_i \sim N(0, \sigma_b^2)$.

Example 2:

Random coefficients (random intercepts and slopes) Poisson regression model:

$$\log\{E(Y_{ij}|X_{ij}, b_i)\} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij}$$

$\text{Var}(Y_{ij}|X_{ij}, b_i) = E(Y_{ij}|X_{ij}, b_i)$ (Poisson variance),

and $b_i \sim N(0, G)$.

Note: G is the covariance matrix for b_{0i} and b_{1i}.

Interpretation of Fixed Effects

GLMMs are most useful when the scientific objective is to make inferences about individuals rather than population averages.

For example, consider the following logistic model,

$$\text{logit}\{E(Y_{ij}|X_{ij}, b_i)\} = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + b_i$$

with $b_i \sim N(0, \sigma^2)$.

The interpretation of any component of β, say β_k, is in terms of adjusted changes in an *individual's* log odds of response for a unit change in the corresponding covariate, say X_{ijk}.

When X_{ijk} takes on some value x, the log odds of a positive response is,

$$\log\left\{\frac{\Pr(Y_{ij}=1|b_i, X_{ij1}, \ldots, X_{ijk}=x, \ldots, X_{ijp})}{\Pr(Y_{ij}=0|b_i, X_{ij1}, \ldots, X_{ijk}=x, \ldots, X_{ijp})}\right\} =$$

$$\beta_0 + b_i + \beta_1 X_{ij1} + \cdots + \beta_k x + \cdots + \beta_p X_{ijp}.$$

Similarly, when X_{ijk} now takes on some value $x + 1$,

$$\log\left\{\frac{\Pr(Y_{ij}=1|b_i, X_{ij1}, \ldots, X_{ijk}=x+1, \ldots, X_{ijp})}{\Pr(Y_{ij}=0|b_i, X_{ij1}, \ldots, X_{ijk}=x+1, \ldots, X_{ijp})}\right\} =$$

$$\beta_0 + b_i + \beta_1 X_{ij1} + \cdots + \beta_k (x+1) + \cdots + \beta_p X_{ijp}.$$

$\longrightarrow$ β_k is adjusted change in log odds for individual with propensity to respond, b_i.

Estimation

ML estimation of $\boldsymbol{\beta}$ (and possibly ϕ) and G is based on the marginal or integrated likelihood of the data

$$L(\beta, \phi, G) = \prod_{i=1}^{N} \int f(Y_i|X_i, b_i) f(b_i) db_i$$

obtained by averaging over distribution of random effects, b_i.

However, simple analytic solutions are rarely available.

In general, ML estimation requires numerical or Monte Carlo integration techniques that can be computationally quite intensive.

Case Study

Oral Treatment of Toenail Infection

Randomized, double-blind, parallel-group, multicenter study of 294 patients comparing 2 oral treatments (denoted A and B) for toe-nail infection.

Outcome variable: Binary variable indicating presence of onycholysis (separation of the nail plate from the nail bed).

Patients evaluated for degree of onycholysis (separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4, 8, 12, 24, 36, and 48.

Interested in the effect of treatment on changes in an individual's risk of onycholysis over time?

Assume that the conditional probability of onycholysis follows a logistic model,

$$\text{logit}\left\{E(Y_{ij}|X_{ij}, b_i)\right\} = \beta_0 + \beta_1 \text{Month}_{ij} + \beta_2 \text{Trt}_i * \text{Month}_{ij} + b_i$$

where $Trt = 1$ if treatment group B and 0 otherwise.

Here, we assume that $\text{Var}(Y_{ij}|X_{ij}, b_i) = E(Y_{ij}|X_{ij}, b_i)\left\{1 - E(Y_{ij}|X_{ij}, b_i)\right\}$.

We also assume $b_i \sim N(0, \sigma_b^2)$.

Table 9: ML estimates and standard errors from random effects logistic regression model for onycholysis data.

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	-1.697	0.330	-5.15
Month	-0.389	0.043	-8.97
Trt × Month	-0.142	0.065	-2.19
σ_b^2	16.034	3.039	

ML based on 100-point adaptive Gaussian quadrature.

Results

From the output above, we would conclude that:

1. There is a significant difference in the rate of decline of risk for individuals in the two treatment groups ($P < 0.05$).

2. Over 12 months, the odds of onycholysis decreases by a factor of 0.01 [or $\exp(-0.389 * 12)$] for an individual receiving treatment A.

3. Over 12 months, the odds of onycholysis decreases by a factor of 0.002 [$\exp(-0.531 * 12)$] for an individual receiving treatment B.

4. Odds ratio comparing 12 month decreases in risk between treatments A and B is approx 5.5 (or $e^{12*0.142}$).
   Note: Approx. twice as large as corresponding odds ratio from marginal model (OR = 2.6).

Wrap-Up: Summary of Key Points

Linear mixed effects models particularly well-suited to analysis of highly unbalanced longitudinal data.

Introduction of random effects accounts for correlation among repeated measures and allows for heterogeneity of variance over time.

In general, random effects covariance structure is relatively parsimonious.

Semiparametric models: Combining mixed models with penalized spline smoothing provides additional flexibility for modelling change over time.

Alternatively, can avoid introduction of random effects altogether; so that $\text{Cov}(Y_i|X_i) = \Sigma$, for some arbitrary structure on Σ.

Parallel methods available for analyses of discrete longitudinal outcomes:

(a) Marginal models bypass random effects; instead allow dependence among repeated measures via "working" covariance matrix.

(b) GLMMs extend in natural way conceptual approach of linear mixed models via introduction of vector of random effects.

With non-linear link functions, distinction is more important.

Greater care required in choice of model for discrete longitudinal data.

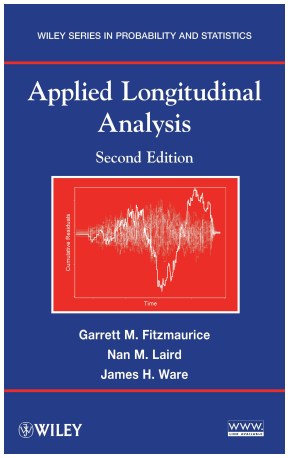
With different targets of inference, different models for categorical outcomes address subtly different questions regarding longitudinal change.

Choice among models?

- should be guided by specific scientific question of interest
- answers to different questions will usually demand that different models have to be applied
- different questions will often produce different, albeit compatible, answers
- "one size does not fit all"

FURTHER READING

Fitzmaurice GM, Laird NM & Ware JH (2011). *Applied Longitudinal Analysis, 2nd Ed.* Hoboken, NJ: Wiley. [See Chapters 12, 13, 14, 16]



Also, see web site: `www.biostat.harvard.edu/~fitzmaur/ala2e`

FURTHER READING

Agresti A (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley. (Section 12.2).

Diggle PJ, Heagerty P, Liang K-Y & Zeger SL (2002). *Analysis of Longitudinal Data*, 2nd ed. New York: Oxford University Press. (Section 7.4).

Graubard BI & Korn EL (1994). Regression analysis with clustered data. *Statistics in Medicine*, **13**, 509–522.

Neuhaus JM, Kalbfleisch JD & Hauck WW (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, **59**, 25–35.

Zeger SL, Liang K-Y & Albert PS (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.