BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 1

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Introduction

Longitudinal Studies: Studies in which individuals are measured repeatedly through time.

This course will cover the analysis and interpretation of results from longitudinal studies.

Emphasis will be on model development, use of statistical software, and interpretation of results.

Theoretical basis for results mentioned but not developed.

No calculus or matrix algebra is assumed.

Features of Longitudinal Data

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 on individuals, we can capture *within-individual* change.

Note: measurements in a longitudinal study are commensurate, i.e., the same variable is measured repeatedly.

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.

Complications:

(i) repeated measures on individuals are correlated
(ii) variability is often heterogeneous across measurement occasions

Longitudinal data require somewhat more sophisticated statistical techniques because the repeated observations are usually (positively) correlated.

Correlation arises due to repeated measures on the same individuals.

Sequential nature of the measures implies that certain types of correlation structures are likely to arise.

Correlation must be accounted for in order to obtain valid inferences.

Heterogeneous variability must also be accounted for in order to obtain valid inferences.

Relation to Correlated Data

Correlated data commonly arise in many applications.

Longitudinal Studies: designs in which the outcome variable is measured repeatedly over time.

Repeated Measures Studies: somewhat older terminology applied to special set of longitudinal designs characterized by measurement at a common set of occasions (usually in an experimental setting under different conditions or treatments).

This course will emphasize methods for analyzing and interpreting the results from longitudinal studies.

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 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: Blood lead levels (μg/dL) at baseline, week 1, week 4, and week 6 for 8 randomly selected children.

ID	Group[a]	Baseline	Week 1	Week 4	Week 6
046	P	30.8	26.9	25.8	23.8
149	A	26.5	14.8	19.5	21.0
096	A	25.8	23.0	19.1	23.2
064	P	24.7	24.5	22.0	22.5
050	A	20.4	2.8	3.2	9.4
210	A	20.4	5.4	4.5	11.9
082	P	28.6	20.8	19.2	18.4
121	P	33.7	31.6	28.5	25.1

[a] **P = Placebo; A = Succimer**.

Table 2: 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)

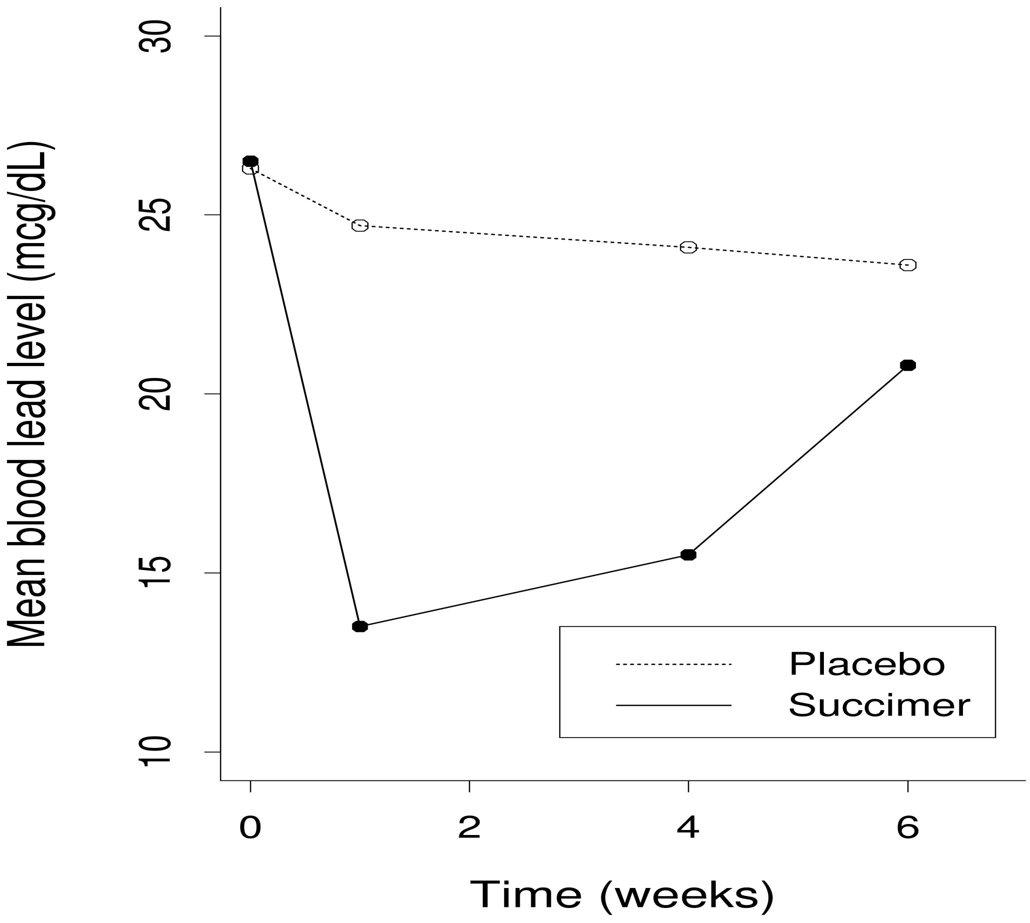


Figure 1: Plot of mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

Example 2: *Six Cities Study of Air Pollution and Health*

- Longitudinal study designed to characterize lung function growth in children and adolescents.

- Most children were enrolled between the ages of six and seven and measurements were obtained annually until graduation from high school.

- Focus on a randomly selected subset of the 300 female participants living in Topeka, Kansas.

- Response variable: Volume of air exhaled in the first second of spirometry manoeuvre, FEV_1.

Table 3: Data on age, height, and FEV_1 for a randomly selected girl from the Topeka data set.

Subject ID	Age	Height	Time	FEV_1
159	6.58	1.13	0.00	1.36
159	7.65	1.19	1.06	1.42
159	12.74	1.49	6.15	2.13
159	13.77	1.53	7.19	2.38
159	14.69	1.55	8.11	2.85
159	15.82	1.56	9.23	3.17
159	16.67	1.57	10.08	2.52
159	17.63	1.57	11.04	3.11

Note: Time represents time since entry to study.

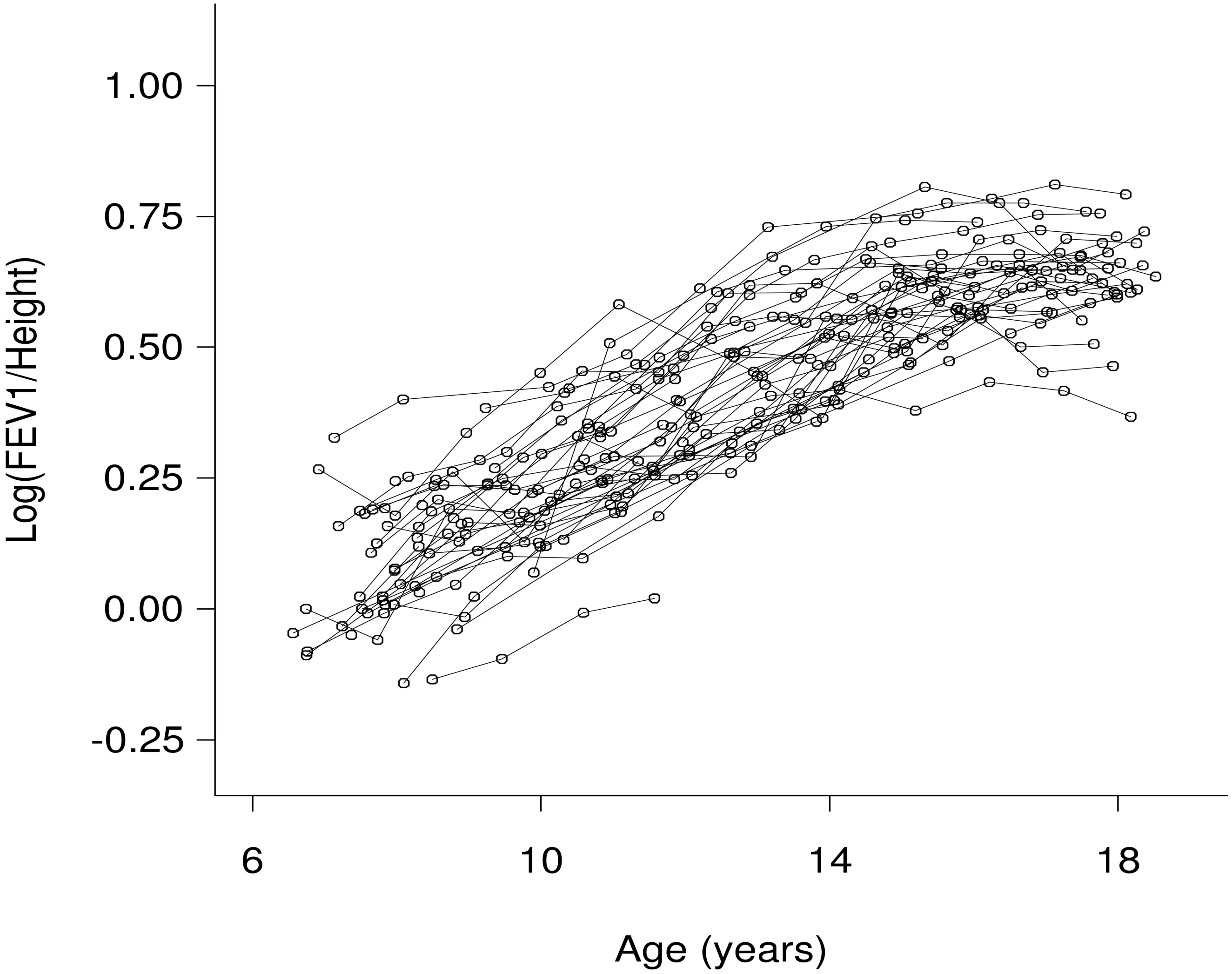


Figure 2: Timeplot of log(FEV$_1$/height) versus age for 50 randomly selected girls from the Topeka data set.

Example 3: *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.

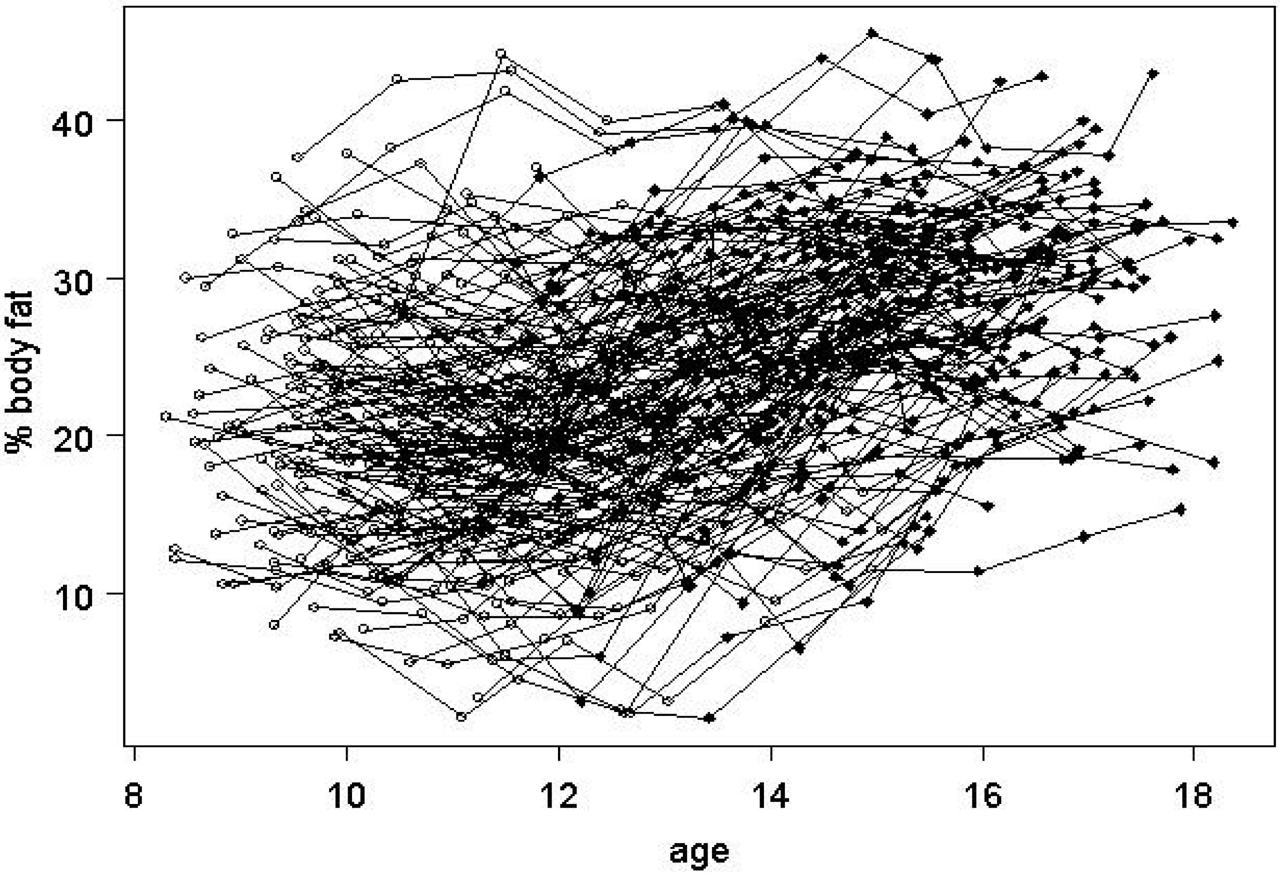


Figure 3: Timeplot of percent body fat against age (in years).

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.

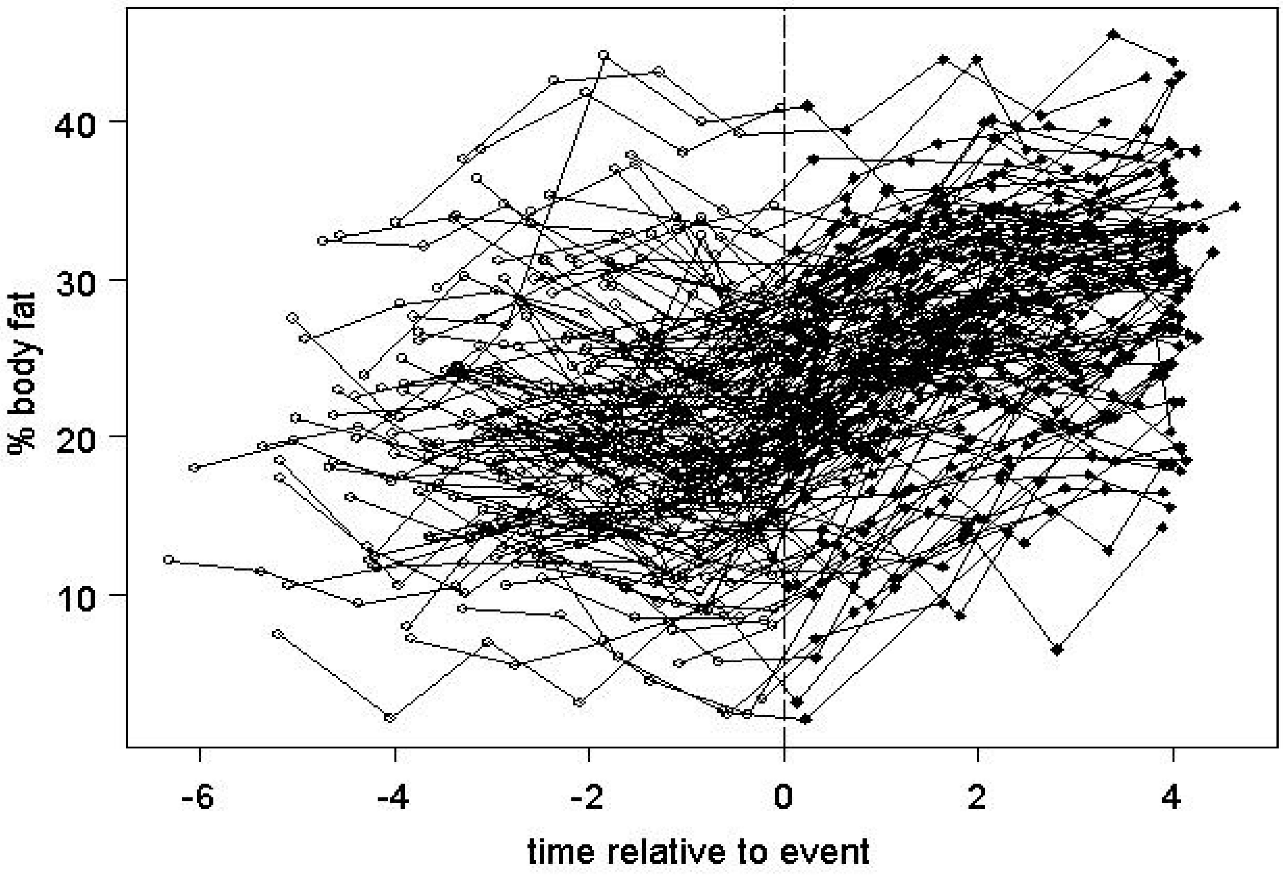


Figure 4: Timeplot of percent body fat against time, relative to age of menarche (in years).

Example 4: *Oral Treatment of Toenail Infection*

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

Example 5: *Clinical Trial of Anti-Epileptic Drug Progabide*

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.

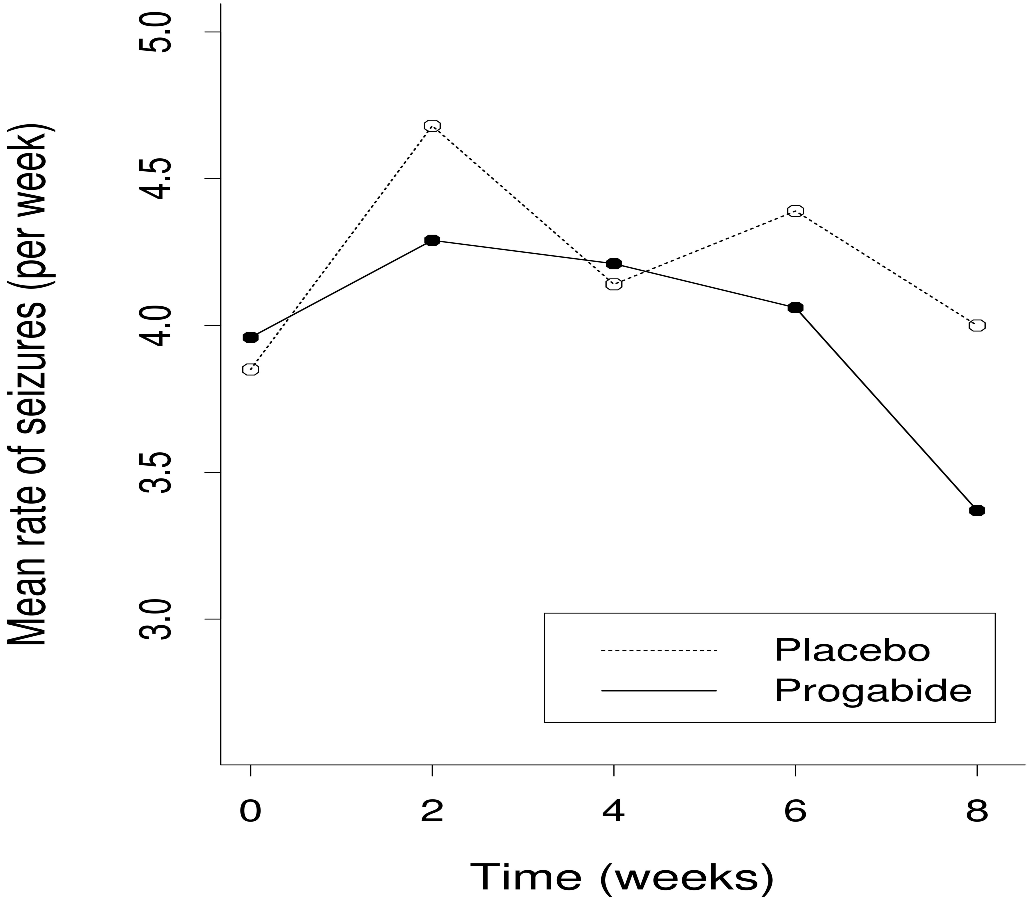


Figure 5: Mean rate of seizures (per week) at baseline, week 2, week 4, week 6, and week 8 in the progabide and placebo groups.

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.

Note: Designs can be balanced, although studies may have incompleteness in data collection.

Features of Longitudinal Data

In longitudinal studies the outcome variable can be:

- continuous (e.g., blood lead levels)
- binary (e.g., presence/absence of onycholysis)
- count (e.g., number of epileptic seizures)

The data set can be incomplete (missing data/dropout).

Subjects may be measured at different occasions (e.g., due to mistimed measurements).

In this course we will develop a set of statistical tools that can handle all of these cases.

Emphasis on concepts, model building, software, and interpretation.

Organization of Course

1) *Introduction to Repeated Measures Analysis*
   Review of Regression/One-Way ANOVA
   Simple Repeated Measures Analysis
      Outcome: Continuous
      Balanced and complete data
      Software: PROC GLM/MIXED in SAS

2) *Linear Models for Longitudinal Data*
   More general approach for fitting linear models to unbalanced, incomplete longitudinal data.
      Outcome: Continuous
      Unbalanced and incomplete data
      Class of models: Linear models
      Software: PROC MIXED in SAS

Organization of Course (cont.)

3) *Generalized Linear Models for Longitudinal Data*
Generalizations and extensions to allow fitting of nonlinear models to discrete longitudinal data.
Outcome: Continuous, binary, count
Class of models: Generalized linear models (e.g. logistic regression)
Software: PROC GENMOD/NLMIXED in SAS

4) *Multilevel Models*
Methods for fitting mixed linear models to multilevel data
Outcomes: Continuous
Unbalanced two, three, and higher-level data
Software: PROC MIXED in SAS, using multiple
RANDOM statements

Background Assumed

1) Samples and populations

2) Population values: parameters (Greek)
Sample values: estimates

3) Variables:

Y: Outcome, response, dependent variable
X: Covariates, independent variables

4) Regression Models

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$

5) Inference

Estimation, testing, and confidence intervals

6) Multiple linear regression/ANOVA
Multiple logistic regression

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 2

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Linear Regression and Analysis of Variance

As background for our discussion of repeated measures and longitudinal analysis, we review the standard linear regression model for independent observations.

We discuss maximum likelihood (ML) and least squares estimation.

We also gently introduce some vector and matrix notation.

Finally, we consider the close connection between analysis of variance (ANOVA) and linear regression.

Consider a cross-sectional data set of 300 measurements of the logarithm of FEV_1, age, and logarithm of height of children living in Topeka, Kansas.

We will fit a model describing how the value of $\log(\mathrm{FEV}_1)$ varies linearly with age and log(height).

The children varied in age from 6 to 18 years.

We will fit a multiple linear regression model, estimate the regression coefficients for age and log(height), and test the hypothesis that these coefficients are not significantly different from 0.

(Note: See the chapter on multiple regression in the folder **Articles** on the course website)

Structure of Six Cities Data

Subject	Height	Age	Log(FEV1)
48	1.45	11.2991	0.62058
17	1.17	6.6639	0.28518
166	1.19	8.1396	0.14842
81	1.48	15.3347	0.57661
3	1.60	16.0164	1.08519
218	1.35	9.8015	0.50078
80	1.66	18.5270	0.91629
14	1.27	7.4251	0.37156

Multiple Linear Regression

Multiple linear regression describes how the expected value (mean) of a measured variable depends on a set of measured or categorical covariates, that is, characteristics of the individuals.

Suppose that we have observations on N individuals.

Each individual has a measured outcome,

$$Y_i; \;\; i = 1, ..., N$$

Each observation, Y_i, has an associated set of covariates

$$X_{i1}, X_{i2}, ..., X_{ip}$$

The linear regression model for Y_i can be written as

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i$$

Typically, $X_{i1} = 1$ for all individuals, and then β_1 is the *intercept*.

This model says that the expected value of Y (the average for all individuals with the specified covariate values) varies linearly with the values of the covariates.

$$E(Y_i|X_{i1}, \ldots, X_{ip}) = \mu_{y_i|x_{i1},\ldots,x_{ip}} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Specifically, an increase of one unit in X_j (while holding the remaining covariates fixed/constant) produces an increase/decrease of β_j in the mean of Y.

Assumptions of Multiple Linear Regression

1. Individuals represent a random sample from the population of interest.

2. Independence: $Y_1, \ldots, Y_N$ are independent random variables.

3. Linearity: $E(Y|X_1, \ldots, X_p)$ is a linear function of each of the X's.

4. Normality: Given X's, individual observations of the dependent variable, Y_i, have a normal distribution, with means

$$\mu_{y|x_1,\ldots,x_p} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

5. Homoscedasticity: $\text{Var}(e_i)$ is constant, σ^2

   $\Rightarrow$ constant variation about regression line (or "plane")

Estimation

Basic Idea: Among all possible estimates $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ of $(\beta_1, \ldots, \beta_p)$ choose the estimates such that the fitted regression model "deviates" the least from the data.

$\Rightarrow$ Least Squares Estimation

"Deviation" of fitted model from the data is defined as:

$$\sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{N} \widehat{e}_i^2$$

where $\widehat{Y}_i = \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_p X_{ip}$.

Least Squares (LS) estimates are those values that minimize these deviations, i.e., minimize the residual sums of squares.

Maximum Likelihood Estimation

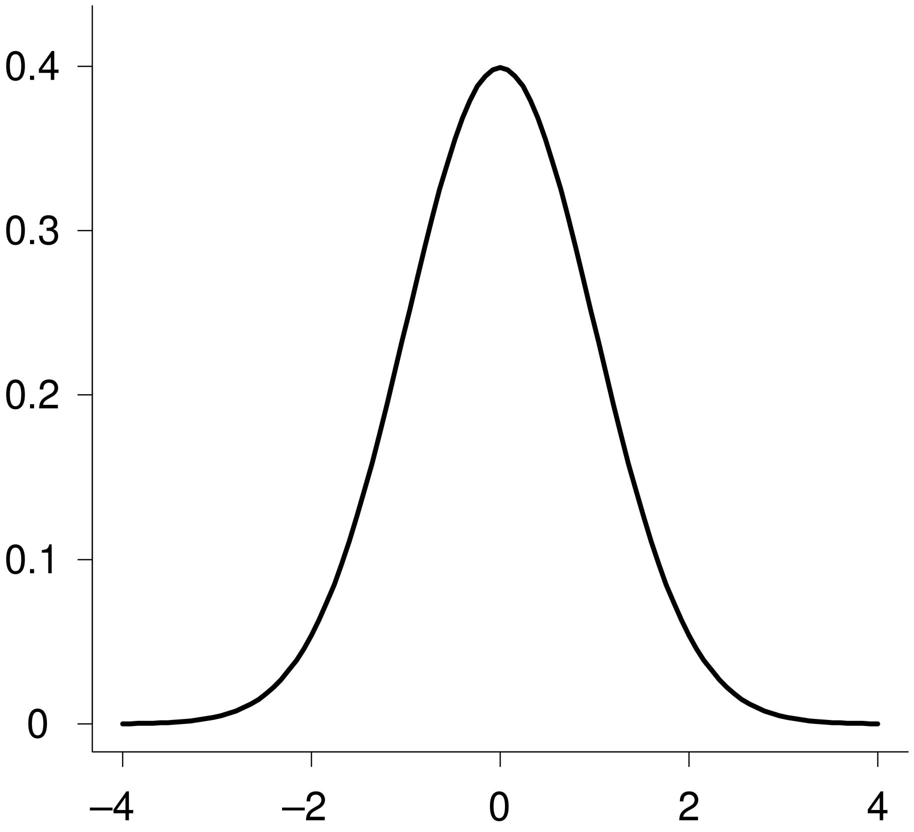
Recall: In the multiple regression model, the values of the covariates are assumed to be fixed.

Only the values of Y_i are random.

The probability distribution corresponding to the linear regression model is given by:

$$f(y_i|X_{i1}, ..., X_{ip}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

Equivalently, we assume that $e_i \sim N(0, \sigma^2)$



Notable Features of the Normal Distribution

$f(Y_i)$ is completely determined by (μ_i, σ^2)

$f(Y_i)$ depends to a very large extent on

$$(Y_i - \mu_i)^2/\sigma^2$$

The latter can be interpreted as a standardized distance of Y_i from μ_i, relative to σ^2, a measure of the spread of values around μ_i.

Maximum Likelihood Estimation

The main idea behind the method of maximum likelihood (ML) is really quite simple and conveyed by its name:

Use as estimates of $\beta_1, ..., \beta_p$ (and σ^2) the values that are most probable (or "likely") for the data that we have observed.

That is, choose values of $\beta_1, ..., \beta_p$ (and σ^2) that maximize the probability of the response variables evaluated at their observed values.

The resulting values are called the maximum likelihood estimates (MLEs) of $\beta_1, ..., \beta_p$ (and σ^2).

For a single observation, we can be at the "most likely" point on the probability curve

$$f(y_i|X_{i1}, ..., X_{ip}) = \frac{1}{\sqrt{2\pi\sigma^2}}\text{exp}\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

by choosing $\beta_1 X_{i1} + \cdots + \beta_p X_{ip} = y_i$.

However, there is more than one observation.

With N subjects, the likelihood is given by $L(\beta_1, ..., \beta_p) =$

$$\prod_{i=1}^{N} f(y_i|X_{i1}, ..., X_{ip}) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\text{exp}\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

In general, it is not possible to choose $\beta_1, \ldots, \beta_p$ that will match every y_i to every $\beta_1 X_{i1} + \cdots + \beta_p X_{ip}$.

Instead, choose $\beta_1, \ldots, \beta_p$ to make the match as close as possible for all subjects.

$\Longrightarrow$ Choose $\beta_1, \ldots, \beta_p$ to maximize $L(\beta_1, \ldots, \beta_p)$.

It happens that ML Estimates = Least Squares Estimates.

These estimates can be written in closed-form using vector and matrix notation.

Linear Regression in Vector Notation

The linear regression model for Y_i

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i$$

can also be written using vector notation

$$\begin{aligned} Y_i &= \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i \\ &= X_i'\beta + e_i \end{aligned}$$

where X_i is a $(p \times 1)$ vector representing the covariates, $X_i' = (X_{i1}, X_{i2}, ..., X_{ip})$, and $\beta' = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of p regression parameters.

Aside: Matrix Addition and Multiplication

Matrices are like spreadsheets.

Importantly, they allow us to perform several arithmetic operations simultaneously.

They also provide a convenient shorthand notation.

Consider the following two simple examples.

$$\text{Let } A = \begin{pmatrix} 2 & 4 \\ 3 & 5 \end{pmatrix}, \quad \text{and} \quad B = \begin{pmatrix} 1 & 6 \\ 8 & 7 \end{pmatrix}.$$

Then,

$$A + B = \begin{pmatrix} 2+1 & 4+6 \\ 3+8 & 5+7 \end{pmatrix} = \begin{pmatrix} 3 & 10 \\ 11 & 12 \end{pmatrix}$$

$$A * B = \begin{pmatrix} 2*1+4*8 & 2*6+4*7 \\ 3*1+5*8 & 3*6+5*7 \end{pmatrix} = \begin{pmatrix} 34 & 40 \\ 43 & 53 \end{pmatrix}$$

Vectors

Vectors are special cases of matrices with one row or one column.

They follow the rules for matrices.

By convention, when we write a vector as X, we understand it to be a column vector of dimension, say $p \times 1$.

When we want to indicate a row vector, we write X'.

In linear regression we have the product of the following two vectors:

$$X_i'\beta = (X_{i1}, X_{i2}, ..., X_{ip}) \left(\begin{array}{c} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{array} \right) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Maximum Likelihood Estimation

With independent observations, the joint density is simply the product of the individual univariate normal densities for Y_i.

Hence, we wish to maximize

$$\begin{aligned}\prod_{i=1}^{N} f(Y_i|X_{i1}, ..., X_{ip}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left\{-\frac{\left(Y_i - X_i'\beta\right)^2}{2\sigma^2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\sum_{i=1}^{N}\frac{\left(Y_i - X_i'\beta\right)^2}{2\sigma^2}\right\},\end{aligned}$$

evaluated at the observed values of the data, with respect to the regression parameters, β.

This is called maximizing the likelihood function.

Note that maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood.

Hence, we can maximize

$$-\sum_{i=1}^{N} \left(Y_i - X_i'\beta\right)^2 / 2\sigma^2$$

by minimizing

$$\sum_{i=1}^{N} \left(Y_i - X_i'\beta\right)^2 / 2\sigma^2$$

Note: This is equivalent to finding the least squares estimates of β, i.e., the values that minimize the sum of the squares of the residuals.

The least squares solution can be written as

$$\widehat{\beta} = \left[\sum_{i=1}^{N} (X_i X_i') \right]^{-1} \sum_{i=1}^{N} (X_i Y_i)$$

This least squares estimate is the value that PROC GLM or PROC REG in SAS or any least squares regression program will produce.

Properties of Least Square Estimator

1. For any choice of σ^2, the least squares estimate of β is unbiased, that is

$$E(\widehat{\beta}) = \beta$$

2. The sampling distribution is given by:

$$\text{Cov}(\widehat{\beta}) = \sigma^2 \left[\sum_{i=1}^{N} (X_i X_i') \right]^{-1}$$

$$\widehat{\beta} \sim N\left(\beta,\ \sigma^2 \left[\sum_{i=1}^{N} (X_i X_i') \right]^{-1} \right)$$

Regression using PROC GLM in SAS

```
DATA topeka;
     INFILE ’g\shared\bio226\topeka.txt’;
     INPUT id height age logfev;
        loght = log(height);
RUN;

PROC GLM DATA=topeka;
     MODEL logfev = age loght;
RUN;
```

```
                SAS The GLM Procedure

Dependent Variable:  logfev

Solution for Fixed Effects

                  Sum of
Source     DF     Squares    Mean Square    F Value     Pr > F
Model       2     29.149       14.5746        876.6    <0.0001
Error     297      4.938        0.0166
Total     299     34.087

                   Estimates

                         Standard
Parameter     Estimate     Error     t value        Pr > |t|
Intercept    -0.355       0.0319     -11.14         <0.0001
Age           0.020       0.0045       4.42         <0.0001
LogHt         2.295       0.1640      13.99         <0.0001
```

Interpretation

The fitted model is

$$\log(\mathrm{FEV}_1) = -0.355 + 0.020 * \mathrm{age} + 2.295 * \log(\mathrm{ht})$$

So, a 1 year increase in age is associated with a 0.020 increase in $\log(\mathrm{FEV}_1)$ (while holding height constant).

Similarly, for the first child on slide 29, the fitted value of $\log(\mathrm{FEV}_1)$ is

$$\log(\mathrm{FEV}_1) = -0.355 + 0.020 * 11.3 + 2.295 * \log(1.45) = 0.724$$

so that the predicted $\mathrm{FEV}_1 = \exp(0.724) = 2.06$ liters.

Analysis of Variance

ANOVA: Describes how the mean of a continuous dependent variable depends on a nominal (categorical, class) independent variable.

Analyzing samples from each of the p populations, we ask:

- Are there any differences in the p population means?
- If so, which of the means differ?

$\Longrightarrow$ One-Way Analysis of Variance (ANOVA)

Objective: To estimate and test hypotheses about the population group means, $\mu_1, \mu_2, \ldots, \mu_p$.

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$
$H_A : \mu_j$'s not all equal

Note: Some of the μ_j's could be equal under H_A

Example: Dosages of four cardiotoxic drugs at death of infused guinea pigs

- Evaluating potencies of four cardiac treatments
- Observe dosage at which animals (guinea pigs) die for each treatment
- 10 guinea pigs per treatment (40 observations in all)
- Assess any differences in toxicity of four treatments, i.e., differences in mean dosage required to kill animal

$$\bar{y}_1 = 25.9, \;\; \bar{y}_2 = 22.2, \;\; \bar{y}_3 = 20.0, \;\; \bar{y}_4 = 19.6$$

Goal of One-Way ANOVA

Assess whether a factor has a significant "effect" on a continuous outcome variable (Y)

Two complementary ways to consider this:

- Does the mean of Y differ among levels of a factor?

- Do differences among levels of a factor explain some of the variation in Y?

ANOVA: Analyzing variances? Although interested in comparing means, we do so by comparing variances.

Sources of Variation:

- quantify variability between sample means
  $\longrightarrow$ Between groups variability ("mean square between")

- quantify error variability or variability of observations in the same group
  $\longrightarrow$ Error or within groups variability ("mean square error")

The ANOVA table provides a summary and comparison of these sources of variation.

$$\begin{aligned} \text{Between} >> \text{Within (Error)} \quad & \Longrightarrow \mu_j\text{'s vary} \\ & \Longrightarrow \text{reject } H_0 \end{aligned}$$

Otherwise cannot reject H_0.

Example: Dosages of four cardiotoxic drugs at death of infused guinea pigs

- Evaluating potencies of four cardiac treatments
- Observe dosage at which animals (guinea pigs) die for each treatment
- 10 guinea pigs per treatment (40 observations in all)
- Assess any differences in toxicity of four treatments, i.e. differences in mean dosage required to kill animal

$$\bar{y}_1 = 25.9, \ \ \bar{y}_2 = 22.2, \ \ \bar{y}_3 = 20.0, \ \ \bar{y}_4 = 19.6$$

SAS Syntax for One-Way ANOVA

```
DATA toxic;
      INFILE ’g:\shared\bio226\tox.txt’;
      INPUT y drug;
RUN;
PROC GLM DATA=toxic;
      CLASS drug;
      MODEL y=drug;
RUN;
```

ANOVA Table

Source	**DF**	**SS**	**MS**	**F**
Between (Drug)	3	249.9	83.3	8.5
Within (Error)	36	350.9	9.7	
Total	39	600.8		

$F = 83.3/9.7 = 8.5, \quad (p < 0.001) \Longrightarrow \text{reject } H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$

Note:

- This is a “global” test.
- It does not specify which μ_j’s differ.

Relationship between ANOVA & Regression

Essentially identical, although often obscured by differences in terminology.

The ANOVA model can be represented as a multiple regression model with dummy (or indicator) variables.

$\Longrightarrow$ A multiple regression analysis with dummy-variable coded factors will yield the same results as an ANOVA.

Dummy or Indicator Variable Coding

Consider a factor with p levels:

Define $X_2 = 1$ if subject belongs to level 2, and 0 otherwise; define $X_3 = 1$ if subject belongs to level 3, and 0 otherwise; and define $X_4, ..., X_p$ similarly.

Note 1: By omission, the first level of the factor is selected as a "reference".
Note 2: Default option in many procedures in SAS is to use last level as a "reference".

Level	X_2	X_3	X_4	$\ldots$	X_p
1	0	0	0	$\ldots$	0
2	1	0	0	$\ldots$	0
3	0	1	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots\vdots$	$\vdots$
p	0	0	0	$\ldots$	1

This leads to a simple way of expressing the ANOVA model:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + e_i$$

Note:

$$\begin{aligned} \mu_1 &= \beta_1 \\ \mu_2 &= \beta_1 + \beta_2 \\ \mu_3 &= \beta_1 + \beta_3 \\ &\vdots \\ \mu_p &= \beta_1 + \beta_p \end{aligned}$$

Equivalently:

$$\begin{array}{lcl} \text{Group 2 versus (minus) Group 1} & = & \beta_2 \\ \text{Group 3 versus (minus) Group 1} & = & \beta_3 \\ \quad\vdots & & \vdots \\ \text{Group p versus (minus) Group 1} & = & \beta_p \end{array}$$

Choice of Reference Level

The usual choice of reference group:

(i) A natural baseline or comparison group, and/or

(ii) group with largest sample size

Summary

The regression representation of ANOVA is more attractive because:

- It can handle balanced (i.e., equal cell sizes) and unbalanced data in a seamless fashion.
- In addition to the usual ANOVA table summaries, it provides other useful and interpretable results, e.g., estimates of effects and standard errors.
- Generalizations of ANOVA to include continuous predictors (and interactions among nominal and continuous predictors) are straightforward.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 3

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Longitudinal Data - Basic Concepts

Features of Longitudinal Data:

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 on individuals, we can capture *within-individual* change.

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.

Note: Designs can be balanced, although studies may have incompleteness in data collection.

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.

Why are longitudinal data correlated?

What are the potential consequences of not accounting for correlation among longitudinal data in the analysis?

Variability

An additional, although often overlooked, aspect of longitudinal data that complicates their statistical analysis is heterogeneous variability.

That is, the variability of the outcome at the end of the study is often discernibly different than the variability at the start of the study.

This violates the assumption of homoscedasticity that is the basis for standard linear regression techniques.

Thus, there are two aspects of longitudinal data that complicate their statistical analysis: (i) repeated measures on the same individual are usually positively correlated, and (ii) variability is often heterogeneous across measurement occasions.

Notation

Let Y_{ij} denote the response variable for the i^{th} individual $(i = 1, ..., N)$ at the j^{th} occasion $(j = 1, ..., n)$.

If the repeated measures are assumed to be equally-separated in time, this notation will be sufficient.

Later, we refine notation to handle the case where repeated measures are unequally-separated and unbalanced over time.

We can represent the n observations on the N individuals in a two-dimensional array, with rows corresponding to individuals and columns corresponding to the responses at each occasion.

Table 4: Tabular representation of longitudinal data, with n repeated observations on N individuals.

			Occasion		
Individual	1	2	3	$\ldots$	n
1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1n}
2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2n}
.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.
.	.	.	.	$\cdots$	.
N	y_{N1}	y_{N2}	y_{N3}	$\cdots$	y_{Nn}

Vector Notation

We can group the n repeated measures on the same individual into a $n \times 1$ response vector:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix}.$$

Alternatively, we can denote the response vectors Y_i as

$$Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})'.$$

Correlation

Before we can give a formal definition of correlation we need to introduce the notion of *expectation*.

We denote the expectation or mean of Y_{ij} by

$$\mu_j = E(Y_{ij}),$$

where $E(\cdot)$ can be thought of as a long-run average (over individuals).

The mean, μ_j, provides a measure of the location of the center of the distribution of Y_{ij}.

The *variance* provides a measure of the spread or dispersion of the values of Y_{ij} around its respective mean:

$$\sigma_j^2 = E[Y_{ij} - E(Y_{ij})]^2 = E(Y_{ij} - \mu_j)^2.$$

The positive square-root of the variance, σ_j, is known as the *standard deviation*.

The *covariance* between two variables, say Y_{ij} and Y_{ik},

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

is a measure of the *linear* dependence between Y_{ij} and Y_{ik}.

When the covariance is zero, there is no linear dependence between the responses at the two occasions.

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}.

The correlation, unlike covariance, is a measure of dependence free of scales of measurement of Y_{ij} and Y_{ik}.

By definition, correlation must take values between -1 and 1.

A correlation of 1 or -1 is obtained when there is a perfect linear relationship between the two variables.

For the vector of repeated measures, $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in})'$, we define the variance-covariance matrix, $\text{Cov}(Y_i)$,

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

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

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

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

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

This matrix is also symmetric in the sense that $\mathrm{Corr}\,(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = \mathrm{Corr}(Y_{ik}, Y_{ij})$.

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 scatterplot of each pair of repeated measures.

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

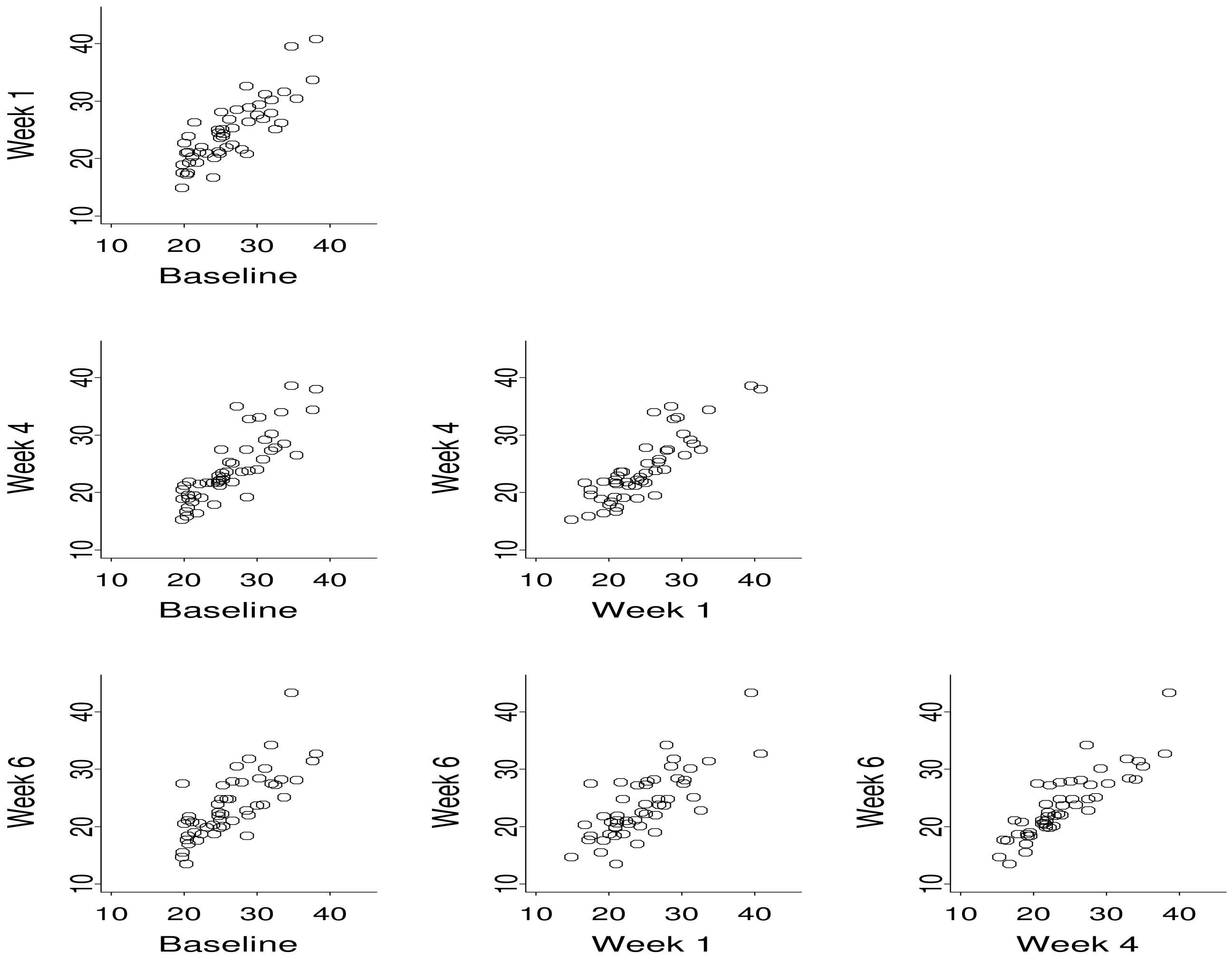


Figure 6: Pairwise scatter-plots of blood lead levels at baseline, week 1, week 4, and week 6 for children in the placebo group.

Table 5: 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 6: 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

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.

Consequences of Ignoring Correlation

Potential impact of ignoring correlation can be illustrated using data from the *Treatment of Lead-Exposed Children Trial.*

For simplicity, consider only the first two repeated measures, taken at week 0 and week 1.

It is of interest to determine the change in the mean response over time.

An estimate of change is given by

$$\widehat{\delta} = \widehat{\mu}_2 - \widehat{\mu}_1,$$

where $\widehat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} Y_{ij}$.

In the TLC trial, the estimate of change in the succimer group is -13.0 (or $13.5 - 26.5$).

For inferences, we also need a standard error.

Variance of $\widehat{\delta}$ is

$$\text{Var}(\hat{\delta}) = \text{Var}\left\{\frac{1}{N}\sum_{i=1}^{N}(Y_{i2} - Y_{i1})\right\} = \frac{1}{N}\left(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\right).$$

Note: Last term, $-2\sigma_{12}$, accounts for the correlation among the repeated measures.

Substituting estimates of the variances and covariance into this expression:

$$\widehat{\text{Var}}(\hat{\delta}) = \frac{1}{50}\{25.2 + 58.9 - 2(15.5)\} = 1.06.$$

What if we had ignored that the data are correlated and proceeded with an analysis assuming all observations are independent?

Independence $\Longrightarrow$ zero covariance.

Leading to (incorrect) estimate of the variance of $\widehat{\delta}$

$$\frac{1}{50}(25.2 + 58.9) = 1.68,$$

which is approximately 1.6 times larger.

In this illustration, ignoring the correlation results in:

- standard errors that are too large
- confidence intervals that are too wide
- p-values for the test of H_0: $\delta = 0$ that are too large

In general, failure to take account of the correlation (covariance) among the repeated measures will result in incorrect estimates of the sampling variability and can lead to quite misleading scientific inferences.

Summary

Primary goal of a longitudinal study is to characterize the change in response over time and the factors that influence change.

Longitudinal data require somewhat more sophisticated statistical techniques because: (i) repeated measures on the same individual are usually positively correlated, and (ii) variability is often heterogeneous across measurement occasions.

Correlation and heterogeneous variability must be accounted for in order to obtain valid inferences about change in response over time.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 4

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Statistical Basis of Longitudinal Analysis (Part 1)

Overview:

In this part of the course we focus on linear models for longitudinal data.

Response variable is continuous and has distribution that is approximately symmetric (without excessive skewness or outliers).

We introduce some additional vector and matrix notation.

We present a general linear regression model for longitudinal data.

Single-Group Repeated Measures Design

Initially, we consider methods for analyzing longitudinal data collected in the simplest design: single-group repeated measures design.

In this design, we have n repeated measures of the response on each of N subjects.

Note: In certain repeated measures designs (e.g., cross-over designs), subjects receive n different treatments at the n occasions.

In cross-over designs, goal is to compare treatments assigned at different occasions.

Listing each observation at the n occasions:

	Occasions					
Subject	1	2	.	.	.	n
1	Y_{11}	Y_{12}	.	.	.	Y_{1n}
2	Y_{21}	Y_{22}	.	.	.	Y_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
N	Y_{N1}	Y_{N2}	.	.	.	Y_{Nn}

If observations satisfied assumptions of one-way ANOVA, we could order them from 1 to Nn in a vector with elements Y_i, and write the model as

$$Y_i \quad = \quad \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_n X_{in} + e_i$$

where

$$\begin{aligned} X_{ij} \quad = \quad & 1, \text{ if observation } i \text{ was obtained} \\ & \quad \text{at } j^{th} \text{ occasion;} \quad (j = 2, ..., n) \\ & 0, \text{ otherwise.} \end{aligned}$$

However, this model needs to be modified to account for the statistical dependence among repeated observations obtained on the same subject.

Example: Treatment of Lead-Exposed Children Trial

For illustrative purposes, consider the data on the 50 children randomized to Succimer.

Subject	Week 0	Week 1	Week 4	Week 6
1	26.5	14.8	19.5	21.0
2	25.8	23.0	19.1	23.2
3	20.4	2.8	3.2	9.4
4	20.4	5.4	4.5	11.9
5	24.8	23.1	24.6	30.9
6	27.9	6.3	18.5	16.3
7	35.3	25.5	26.3	30.3
8	28.6	15.8	22.9	25.9
⋮	⋮	⋮	⋮	⋮
49	21.9	7.6	10.8	13.0
50	20.7	8.1	25.7	12.3

Denote the population means at the n occasions by $\mu_1,\ \mu_2,\ \ldots,\ \mu_n$.

Then the null hypothesis of interest is

$$H_0:\ \mu_1 = \mu_2 =\ \ldots\ = \mu_n$$

How can we test this hypothesis?

We could choose pairs of occasions and perform a series of paired $t-$tests $\Rightarrow\ n\,(n-1)\,/2$ tests.

This approach allows only pairwise comparisons.

Instead, we need to address the problem of correlation (covariance) among repeated measures and extend the one-way ANOVA model.

One approach to analyzing such data is to consider extensions of the one-way ANOVA model that account for the covariance.

That is, rather than assume that repeated observations of the same subject are independent, with homogeneous variance, allow the repeated measurements to have an unknown covariance structure.

To do this, we can use the SAS procedure, PROC MIXED, an extension of PROC GLM which allows clusters of correlated observations.

We will illustrate the use of PROC MIXED using the data from the TLC trial.

Later we will consider the statistical basis for this analysis.

Note: PROC MIXED in SAS requires the data to be in a univariate (or “long”) form.

As a first step, often it will be necessary to transform the data from a “multivariate” (or “wide”) format to a “univariate” (or “long”) format.

PROC MIXED in SAS

```
DATA tlc;
        INFILE ’g:\shared\bio226\lead.txt’;
        INPUT id y1 y2 y3 y4;
            y=y1; time=0; OUTPUT;
            y=y2; time=1; OUTPUT;
            y=y3; time=4; OUTPUT;
            y=y4; time=6; OUTPUT;
        DROP y1-y4;
RUN;

PROC MIXED DATA=tlc;
        CLASS id time;
        MODEL y = time /S CHISQ;
        REPEATED time /TYPE=UN SUBJECT=id R;
        CONTRAST ’Week 6 - Week 0’
            time -1 0 0 1 / CHISQ;
```

Multivariate (or Wide) Form of Succimer Data

ID	Y1	Y2	Y3	Y4
1	26.5	14.8	19.5	21.0
2	25.8	23.0	19.1	23.2
3	20.4	2.8	3.2	9.4
4	20.4	5.4	4.5	11.9
5	24.8	23.1	24.6	30.9
6	27.9	6.3	18.5	16.3
7	35.3	25.5	26.3	30.3
8	28.6	15.8	22.9	25.9
⋮	⋮	⋮	⋮	⋮
49	21.9	7.6	10.8	13.0
50	20.7	8.1	25.7	12.3

Univariate (or Long) Form of Succimer Data
(1st 3 subjects only)

OBS	ID	Y	TIME
1	1	26.5	0
2	1	14.8	1
3	1	19.5	4
4	1	21.0	6
5	2	25.8	0
6	2	23.0	1
7	2	19.1	4
8	2	23.2	6
9	3	20.4	0
10	3	2.8	1
11	3	3.2	4
12	3	9.4	6

Selected Output from PROC MIXED

```
                 The Mixed Procedure

             Estimated R Matrix for id 1

Row          Col1          Col2          Col3          Col4

  1       25.2098       15.4654       15.1380       22.9854
  2       15.4654       58.8671       44.0291       35.9660
  3       15.1380       44.0291       61.6571       33.0220
  4       22.9854       35.9660       33.0220       85.4946
```

```
Covariance Parameter Estimates

Cov Parm     Subject     Estimate


UN(1,1)      id           25.2098
UN(2,1)      id           15.4654
UN(2,2)      id           58.8671
UN(3,1)      id           15.1380
UN(3,2)      id           44.0291
UN(3,3)      id           61.6571
UN(4,1)      id           22.9854
UN(4,2)      id           35.9660
UN(4,3)      id           33.0220
UN(4,4)      id           85.4946
```

```
           Fit Statistics

-2 Res Log Likelihood          1280.3
AIC (smaller is better)        1300.3
AICC (smaller is better)       1301.5
BIC (smaller is better)        1319.5


  Null Model Likelihood Ratio Test

    DF    Chi-Square      Pr > ChiSq

     9         86.73          <.0001
```

```
                      The Mixed Procedure

                  Solution for Fixed Effects

                                   Standard
Effect      time   Estimate           Error     DF   t Value   Pr > |t|

Intercept           20.7620          1.3076     49     15.88     <.0001
time        0        5.7780          1.1378     49      5.08     <.0001
time        1       -7.2400          1.2036     49     -6.02     <.0001
time        4       -5.2480          1.2736     49     -4.12     0.0001
time        6             0               .      .         .          .
```

```
                          The Mixed Procedure

                    Type 3 Tests of Fixed Effects

             Num    Den
Effect        DF     DF   Chi-Square   F Value     Pr > ChiSq   Pr > F

time           3     49       163.72     54.57         <.0001   <.0001


                              Contrasts

                  Num   Den
Label              DF    DF  Chi-Square  F Value    Pr > ChiSq  Pr > F

Week 6 - Week 0     1    49       25.79    25.79        <.0001  <.0001
```

Covariance Structure

When we estimate the covariance matrix without making any particular assumption about the covariance structure, we say that we are using an unrestricted or unstructured covariance matrix.

As we shall see later, it is sometimes advantageous to model the covariance structure more parsimoniously.

How important is it to take account of the covariance among repeated measures?

We can address that question by re-analyzing the blood lead level data under the assumption of independence and homogeneity of variance.

PROC GLM versus PROC MIXED in SAS

```
DATA tlc;
    INFILE 'g:\shared\bio226\lead.txt';
    INPUT id y1 y2 y3 y4;
        y=y1; time=0; OUTPUT;
        y=y2; time=1; OUTPUT;
        y=y3; time=4; OUTPUT;
        y=y4; time=6; OUTPUT;
    DROP y1-y4;
RUN;
PROC GLM DATA=tlc;
    CLASS time;
    MODEL y = time /SOLUTION;
    ESTIMATE 'Week 6 - Week 0' time -1 0 0 1;
RUN;
PROC MIXED DATA=tlc;
    CLASS id time;
    MODEL y = time /S CHISQ;
    REPEATED time /TYPE=UN SUBJECT=id R;
    ESTIMATE 'Week 6 - Week 0' time -1 0 0 1;
RUN;
```

Selected Output from PROC GLM

```
                             The GLM Procedure

Dependent Variable: y

                                        Sum of
 Source                      DF        Squares   Mean Square  F Value  Pr > F

 Model                        3     5104.41815    1701.47272    29.43  <.0001
 Error                      196    11330.20380      57.80716
 Corrected Total            199    16434.62195



 Source                      DF    Type III SS   Mean Square  F Value  Pr > F

 time                         3    5104.418150   1701.472717    29.43  <.0001
```

```
                                         Standard
   Parameter           Estimate             Error    t Value    Pr > |t|

   Intercept        20.76200000        1.07524102      19.31      <.0001
   time      0       5.77800000        1.52062043       3.80      0.0002
   time      1      -7.24000000        1.52062043      -4.76      <.0001
   time      4      -5.24800000        1.52062043      -3.45      0.0007
   time      6       0.00000000         .                .         .


                                            Standard
Parameter                   Estimate           Error    t Value    Pr > |t|

Week 6 - Week 0          -5.77800000      1.52062043      -3.80      0.0002
```

Selected Output from PROC MIXED

```
                          Solution for Fixed Effects

                                     Standard
Effect        time     Estimate         Error       DF    t Value     Pr > |t|

Intercept               20.7620        1.3076       49      15.88       <.0001
time          0          5.7780        1.1378       49       5.08       <.0001
time          1         -7.2400        1.2036       49      -6.02       <.0001
time          4         -5.2480        1.2736       49      -4.12       0.0001
time          6               0             .        .          .            .
```

```
                  The Mixed Procedure

              Type 3 Tests of Fixed Effects

             Num    Den
Effect        DF     DF   Chi-Square   F Value     Pr > ChiSq   Pr > F

time           3     49       163.72     54.57         <.0001   <.0001



                        Estimates

                                Standard
Label              Estimate        Error      DF    t Value    Pr > |t|

Week 6 - Week 0     -5.7780       1.1378      49      -5.08      <.0001
```

Note that the estimates of the change in mean from baseline (week 0) to week 6 are the same in both analyses, i.e., -5.778; but the standard errors are discernibly different.

The standard error yielded by PROC GLM, 1.52, is not valid since the procedure has incorrectly assumed that all of the observations are independent and with homogeneous variance.

The standard error yielded by PROC MIXED, 1.14, is valid since the procedure has accounted for the covariance among repeated measures in the analysis.

Notation of General Linear Model

Previously, we assumed a sample of N subjects are measured repeatedly at n occasions.

Either by design or happenstance, subjects may not have same number of repeated measures or be measured at same set of occasions.

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

We can group the response variables for the i^{th} subject into a $n_i \times 1$ vector:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad i = 1, ..., N.$$

Associated with Y_{ij} there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, ..., N; \;\; j = 1, ..., n_i.$$

Note: Information about the time of observation, treatment or exposure group, and other predictor and confounding variables can be expressed through this vector of covariates.

We can group the vectors of covariates into a $n_i \times p$ matrix:

$$X_i \quad = \quad \left(\begin{array}{cccc} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{array} \right), \quad i = 1, ..., N.$$

X_i is simply an ordered collection of the values of the p covariates for the i^{th} subject at the n_i occasions.

Linear Models for Longitudinal Data

Throughout this course we consider linear regression models for changes in the mean response over time:

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

where $\beta_1, ..., \beta_p$ are unknown regression coefficients.

The e_{ij} are random errors, with mean zero, and represent deviations of the Y_{ij}'s from their means,

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

Typically, $X_{ij1} = 1$ for all i and j, and then β_1 is the intercept term in the model.

Vector and Matrix Representation

Note that the linear model

$$E(Y_{ij}|X_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}, \quad j = 1, ..., n_i;$$

describes the mean response at all n_i occasions.

For example, at the third occasion ($j = 3$),

$$E(Y_{i3}|X_{i3}) = \beta_1 X_{i31} + \beta_2 X_{i32} + \cdots + \beta_p X_{i3p}.$$

This model can also be represented in vector/matrix notation as:

$$E(Y_i|X_i) = X_i\beta,$$

where $\beta' = (\beta_1, ..., \beta_p)$.

Note that the model

$$E(Y_i|X_i) = X_i\beta,$$

is simply a shorthand representation for

$$E\left(\begin{array}{c} Y_{i1} \\ \\ Y_{i2} \\ \vdots \\ \\ Y_{in_i} \end{array}\right) = \left(\begin{array}{cccc} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{array}\right)\left(\begin{array}{c} \beta_1 \\ \\ \beta_2 \\ \vdots \\ \\ \beta_p \end{array}\right).$$

Vectors and matrices simply allow us to express regression models for longitudinal data in a very economical fashion.

Illustration: *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 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 7: Blood lead levels (μg/dL) at baseline, week 1, week 4, and week 6 for 8 randomly selected children.

ID	Group[a]	Baseline	Week 1	Week 4	Week 6
046	P	30.8	26.9	25.8	23.8
149	A	26.5	14.8	19.5	21.0
096	A	25.8	23.0	19.1	23.2
064	P	24.7	24.5	22.0	22.5
050	A	20.4	2.8	3.2	9.4
210	A	20.4	5.4	4.5	11.9
082	P	28.6	20.8	19.2	18.4
121	P	33.7	31.6	28.5	25.1

[a] **P = Placebo; A = Succimer**.

For illustrative purposes, consider model that assumes mean blood lead level changes linearly over time, but at a rate that differs by group.

Assume two treatment groups have different intercepts and slopes:

$$Y_{ij} \quad = \quad \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4} + e_{ij},$$

where $X_{ij1} = 1$ for all i and all j;
$X_{ij2} = t_j$, the week in which the blood lead level was obtained;
$X_{ij3} = 1$ if the i^{th} subject is assigned to the succimer group and $X_{ij3} = 0$ otherwise.
$X_{ij4} = t_j$ if the i^{th} subject is assigned to the succimer group and $X_{ij4} = 0$ otherwise. Alternatively, $X_{ij4} = X_{ij2} * X_{ij3}$.

Thus, for children in the placebo group

$$E(Y_{ij}|X_{ij}) = \beta_1 + \beta_2 t_j,$$

where β_1 represents the mean blood lead level at baseline (week $=$ 0) and β_2 is the constant rate of change in mean blood level.

Similarly, for children in the succimer group

$$E(Y_{ij}|X_{ij}) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)t_j,$$

where $\beta_2 + \beta_4$ is the constant rate of change in mean blood level per week.

Hypothesis that treatments are equally effective in reducing blood lead levels translated into hypothesis that $\beta_4 = 0$.

To reinforce notation, consider the responses and covariates at the 4 occasions for any individual.

For example, the responses at the 4 occasions for ID = 046:

$$\begin{pmatrix} 30.8 \\ 26.9 \\ 25.8 \\ 23.8 \end{pmatrix}.$$

The values of the covariates at the 4 occasions for ID = 046:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 1 & 6 & 0 & 0 \end{pmatrix}.$$

This individual was assigned to treatment with placebo.

On the other hand, the responses at the 4 occasions for ID = 149:

$$\begin{pmatrix} 26.5 \\ 14.8 \\ 19.5 \\ 21.0 \end{pmatrix}.$$

The values of the covariates at the 4 occasions for ID = 149:

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 4 & 1 & 4 \\ 1 & 6 & 1 & 6 \end{pmatrix}.$$

This individual was assigned to treatment with succimer.

So, using vectors and matrices, model for the mean blood lead levels can be represented as

$$E(Y_i) = X_i\beta,$$

where, for example,

$$E(Y_i) = E\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 1 & 6 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + 4\beta_2 \\ \beta_1 + 6\beta_2 \end{pmatrix}$$

for children in the placebo group.

So, model for the mean blood lead levels can be represented as

$$\left(\begin{array}{c} E(Y_{i1}) \\ E(Y_{i2}) \\ E(Y_{i3}) \\ E(Y_{i4}) \end{array}\right) = \left(\begin{array}{ccccccc} \beta_1 * 1 & + & \beta_2 * 0 & + & \beta_3 * 0 & + & \beta_4 * 0 \\ \beta_1 * 1 & + & \beta_2 * 1 & + & \beta_3 * 0 & + & \beta_4 * 0 \\ \beta_1 * 1 & + & \beta_2 * 4 & + & \beta_3 * 0 & + & \beta_4 * 0 \\ \beta_1 * 1 & + & \beta_2 * 6 & + & \beta_3 * 0 & + & \beta_4 * 0 \end{array}\right) = \left(\begin{array}{c} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + 4\beta_2 \\ \beta_1 + 6\beta_2 \end{array}\right)$$

for children in the placebo group, and

$$\left(\begin{array}{c} E(Y_{i1}) \\ E(Y_{i2}) \\ E(Y_{i3}) \\ E(Y_{i4}) \end{array}\right) = \left(\begin{array}{ccccccc} \beta_1 * 1 & + & \beta_2 * 0 & + & \beta_3 * 1 & + & \beta_4 * 0 \\ \beta_1 * 1 & + & \beta_2 * 1 & + & \beta_3 * 1 & + & \beta_4 * 1 \\ \beta_1 * 1 & + & \beta_2 * 4 & + & \beta_3 * 1 & + & \beta_4 * 4 \\ \beta_1 * 1 & + & \beta_2 * 6 & + & \beta_3 * 1 & + & \beta_4 * 6 \end{array}\right) = \left(\begin{array}{c} (\beta_1 + \beta_3) \\ (\beta_1 + \beta_3) + (\beta_2 + \beta_4) \\ (\beta_1 + \beta_3) + 4(\beta_2 + \beta_4) \\ (\beta_1 + \beta_3) + 6(\beta_2 + \beta_4) \end{array}\right)$$

for children in the succimer group.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 5

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Statistical Basis of Longitudinal Analysis (Part 2)

Overview:

Previously, we introduced some additional vector and matrix notation.

We also presented a general linear regression model for longitudinal data:

$$E(Y_{ij}|X_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}, \;\; j = 1, ..., n_i.$$

Next, we consider distributional assumptions and discuss inference based on maximum likelihood (ML).

General Linear Model for Longitudinal Data

We assume a general linear regression model,

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

where $\beta_1, ..., \beta_p$ are unknown regression coefficients.

The e_{ij} are random errors, with mean zero, and are expected to be correlated within individuals.

That is, $\text{Cov}(e_{ij}, e_{ij'}) \neq 0 \quad (j \neq j')$.

To simplify notation, in the following we assume that $n_i = n$.

Assumptions

(1) The individuals represent a random sample from the population of interest.

(2) The elements of the vector of repeated measures $Y_{i1}, \ldots, Y_{in}$, have a Multivariate Normal (MVN) distribution, with means

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

(3) Observations from different individuals are independent, while repeated measurements of the same individual are not assumed to be independent.

The covariance matrix of the vector of observations, $Y_{i1}, \ldots, Y_{in}$, is denoted Σ and its elements are $\sigma_{jj'}$ (typically, we denote variances, σ_{jj}, by σ_j^2).

Probability Models

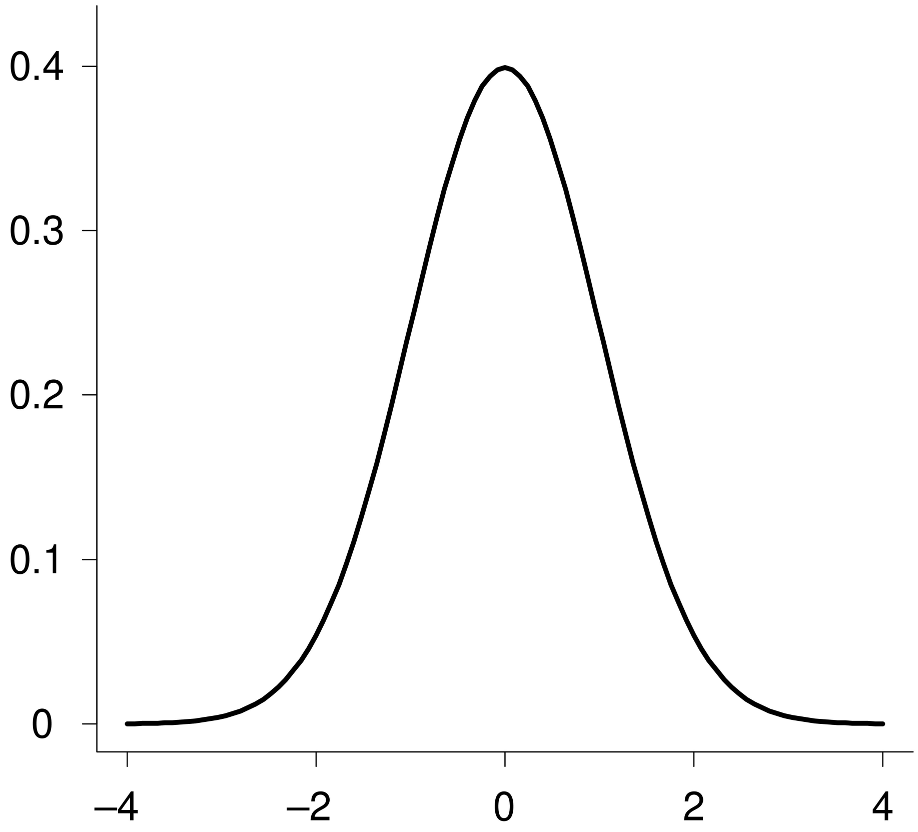
The foundation of most statistical procedures is a probability model, i.e., probability distributions are used as models for the data.

A probability distribution describes the likelihood or relative frequency of occurrence of particular values of the response (or dependent) variable.

Recall: The normal probability density for a single response variable, say Y_i, in the standard linear regression model is:

$$f(y_i|X_{i1},...,X_{ip}) = \frac{1}{\sqrt{2\pi\sigma^2}}\text{exp}\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

Equivalently, we assume that $e_i \sim N(0, \sigma^2)$



With repeated measures we have a vector of response variables and must consider joint probability models for the entire vector of responses.

A joint probability distribution describes the probability or relative frequency with which the vector of responses takes on a particular set of values.

The Multivariate Normal Distribution is an extension of the Normal distribution for a single response to a vector of responses.

Multivariate Normal Distribution

The multivariate normal probability density function for $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$ has the following representation:

$$f\left(Y_i | X_i\right) = f\left(Y_{i1}, Y_{i2}, \ldots, Y_{in} | X_i\right) =$$

$$(2\pi)^{-n/2} \left|\Sigma\right|^{-1/2} \exp\left[-\left(Y_i - X_i\beta\right)' \Sigma^{-1} \left(Y_i - X_i\beta\right)/2\right]$$

where $|\Sigma|$ is the *determinant* of Σ (also known as the *generalized variance*).

Note that $f\left(Y_i | X_i\right)$ describes the probability or relative frequency of occurrence of a particular set of values of $(Y_{i1}, Y_{i2}, \ldots, Y_{in})$.

Notable Features:

- $f\,(Y_i|X_i)$ is completely determined by the vector of means, $\mu_i = X_i\beta$, and by Σ
- $f\,(Y_i|X_i)$ depends to a very large extent on

$$(Y_i - X_i\beta)'\,\Sigma^{-1}\,(Y_i - X_i\beta)$$

- Although somewhat more complicated than in the univariate case, the latter has interpretation in terms of a measure of distance

In the *bivariate* case, it can be shown that

$$(Y_i - \mu_i)' \, \Sigma^{-1} \, (Y_i - \mu_i) =$$

$$(1-\rho^2)^{-1} \left\{ \frac{(Y_{i1}-\mu_1)^2}{\sigma_{11}} + \frac{(Y_{i2}-\mu_2)^2}{\sigma_{22}} - 2\rho \frac{(Y_{i1}-\mu_1)(Y_{i2}-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}} \right\}$$

where $\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$.

Note that this measure of *distance*

(i) down-weights deviations from the mean when the variance is large; this make intuitive sense because when the variance is large the "information" is somewhat poorer; and

(ii) modifies the distance depending on the magnitude of the correlation; when there is strong correlation, knowing that Y_{i1} is "close" to μ_1 also tells us something about how close Y_{i2} is to μ_2.

Maximum Likelihood and Generalized Least Squares

Next we consider a framework for estimation of the unknown parameters, β and Σ.

When full distributional assumptions have been made for vector of responses a standard approach is to use the method of *maximum likelihood* (ML).

Recall main idea behind ML: use as estimates of β and Σ the values that are most probable (or “likely”) for the data that we have observed.

That is, choose values of β and Σ that maximize the probability of the response variables evaluated at their observed values.

Regression with Independent Observations

For standard linear regression, with independent observations, the joint density is the product of the individual univariate normal densities.

We maximize

$$\begin{aligned}\prod_{i=1}^{N} f(y_i|X_{i1},...,X_{ip}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\sum_{i=1}^{N} \frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right\},\end{aligned}$$

with respect to the regression parameters, β,
or minimize

$$\sum_{i=1}^{N} (Y_i - X_i'\beta)^2 / 2\sigma^2$$

Generalized Least Squares

To find ML estimate of β in the repeated measures setting we first assume Σ is known (later, we will relax this unrealistic assumption).

Given that $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$ are assumed to have a multivariate normal distribution, we must maximize the following log-likelihood

$$\begin{aligned}
&\ln \{ \, (2\pi)^{-Nn/2} \, |\Sigma|^{-N/2} \\
&\exp \left[- \sum_{i=1}^{N} (Y_i - X_i\beta)' \, \Sigma^{-1} \, (Y_i - X_i\beta) \, /2 \right] \} \\
&= -\tfrac{Nn}{2} \ln (2\pi) - \tfrac{N}{2} \ln |\Sigma| \\
&- \left[\sum_{i=1}^{N} (Y_i - X_i\beta)' \, \Sigma^{-1} \, (Y_i - X_i\beta) \, /2 \right]
\end{aligned}$$

or minimize

$$\sum_{i=1}^{N} (Y_i - X_i\beta)' \, \Sigma^{-1} \, (Y_i - X_i\beta)$$

The estimate of β that minimizes this expression is known as the generalized least squares (GLS) estimate and can be written as

$$\widehat{\beta} = \left[\sum_{i=1}^{N} \left(X_i'\Sigma^{-1}X_i\right)\right]^{-1} \sum_{i=1}^{N} \left(X_i'\Sigma^{-1}Y_i\right)$$

This is the estimate that PROC MIXED in SAS provides.

Properties of GLS

(1) For any choice of Σ, GLS estimate of β is unbiased; that is, $E(\widehat{\beta}) = \beta$.

(2) $\text{Cov}(\widehat{\beta}) = \left[\sum_{i=1}^{N} \left(X_i'\Sigma^{-1}X_i\right)\right]^{-1}$

(3) Sampling Distribution of $\widehat{\beta}$:

$$\widehat{\beta} \sim N\left(\beta, \left[\sum_{i=1}^{N} \left(X_i'\Sigma^{-1}X_i\right)\right]^{-1}\right)$$

The most efficient generalized least squares estimate is the one that uses the true value of Σ.

Since we usually do not know Σ, we typically estimate it from the data.

In general, no simple expression for ML estimate of Σ.

It has to be found using numerical algorithms that maximize the likelihood.

Once ML estimate of Σ, say $\widehat{\Sigma}$, has been obtained, we substitute it in the GLS estimator to obtain ML estimate of β,

$$\widehat{\beta} = \left[\sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} X_i\right)\right]^{-1} \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} Y_i\right)$$

In large samples, resulting estimator of β has all the same properties as when Σ is known.

Statistical Inference

To test hypotheses about β we can make direct use of the ML estimate $\widehat{\beta}$ and its estimated covariance matrix,

$$\widehat{\text{Cov}}(\widehat{\beta}) = \left[\sum_{i=1}^{N} \left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}.$$

Let L denote a matrix or vector of known weights (often representing contrasts of interest) and suppose that it is of interest to test $H_0 : L\beta = 0$.

Note: Though we usually signify row vectors by transpose symbol, e.g, L', we assume here that L is either a matrix whose rows represent different linear combinations or a single linear combination (L is then a row vector).

Example: Suppose $\beta = (\beta_1, \beta_2, \beta_3)'$ and let $L = (0, 0, 1)$, then $H_0 : L\beta = 0$ is equivalent to $H_0 : \beta_3 = 0$.

Note: A natural estimate of $L\beta$ is $L\widehat{\beta}$ and the covariance matrix of $L\widehat{\beta}$ is given by $L\text{Cov}(\widehat{\beta})L'$.

Thus, the sampling distribution of $L\widehat{\beta}$ is:

$$L\widehat{\beta} \sim N\left(L\beta, L\text{Cov}(\widehat{\beta})L'\right).$$

Case 1: Suppose that L is a single row vector.

Then $L\text{Cov}(\widehat{\beta})L'$ is a single value (scalar) and its square root provides an estimate of the standard error for $L\widehat{\beta}$.

Thus an approximate 95% confidence interval is given by:

$$L\widehat{\beta} \pm 1.96\sqrt{L\widehat{\text{Cov}}(\widehat{\beta})L'}$$

Wald Test

In order to test $H_0 : L\beta = 0$ versus $H_A : L\beta \neq 0$, we can use the Wald statistic

$$Z = \frac{L\widehat{\beta}}{\sqrt{L\widehat{\text{Cov}}(\widehat{\beta})L'}}$$

and compare with a standard normal distribution.

Recall: If Z is a standard normal random variable, then Z^2 has a χ^2 distribution with 1 df. Thus, an identical test is to compare

$$W^2 = (L\widehat{\beta})(L\widehat{\text{Cov}}(\widehat{\beta})L')^{-1}(L\widehat{\beta})$$

to a χ^2 distribution with 1 df.

This approach readily generalizes to L having more than one row and this allows simultaneous testing of more than one hypothesis.

Case 2: Suppose that L has r rows.

Example: Suppose $\beta = (\beta_1, \beta_2, \beta_3)'$ and let

$$L = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix},$$

then $H_0 : L\beta = 0$ is equivalent to $H_0 : \beta_1 = \beta_2 = \beta_3$.

A simultaneous test of the r contrasts is given by

$$W^2 = (L\widehat{\beta})'(L\widehat{\text{Cov}}(\widehat{\beta})L')^{-1}(L\widehat{\beta})$$

which has a χ^2 distribution with r df.

This is how the "Tests of Fixed Effects" are constructed in PROC MIXED.

Likelihood Ratio Test

Suppose that we are interested in comparing two *nested* models, a “full” model and a “reduced” model.

<u>Aside:</u> Nested Models

When one model (the “reduced” model) is a special case of the other (the “full” model), the reduced model is said to be *nested* within the full model.

We can compare two nested models by comparing their maximized log-likelihoods, say $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$; the former is at least as large as the latter.

The larger the difference between $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$ the stronger the evidence that the reduced model is inadequate.

A formal test is obtained by taking

$$2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}})$$

and comparing the statistic to a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models.

Formally, this test is called the *likelihood ratio test* (LRT).

We can use LRTs for hypotheses about models for the mean and the covariance[1].

[1] Later in the course, we will discuss some potential problems with the use of the likelihood ratio test for comparing nested models for the covariance.

Residual Maximum Likelihood (REML) Estimation

Recall: ML estimate of β and Σ is obtained by maximizing the following log-likelihood

$$
\begin{aligned}
&-\tfrac{Nn}{2}\ln\left(2\pi\right)-\tfrac{N}{2}\ln\left|\Sigma\right| \\
&-\left[\sum_{i=1}^{N}\left(Y_i-X_i\beta\right)'\Sigma^{-1}\left(Y_i-X_i\beta\right)/2\right]
\end{aligned}
$$

Although the MLEs have the usual large sample (or asymptotic) properties, the MLE of Σ has well-known bias in small samples (e.g., the diagonal elements of Σ are underestimated).

To see problem, consider linear regression with independent errors.

If the N observations are independent we maximize

$$\prod_{i=1}^{N} f(y_i|X_{i1}, ..., X_{ip}) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\sum_{i=1}^{N}\frac{\left(Y_i - X_i'\beta\right)^2}{2\sigma^2}\right\}.$$

This gives the usual least squares estimator of β, but ML estimator of σ^2 is

$$\widehat{\sigma}^2 = \sum_{i=1}^{N}\left(Y_i - X_i'\widehat{\beta}\right)^2 / N$$

Note: The denominator is N. Furthermore, it can be shown that

$$E(\widehat{\sigma}^2) = \left(\frac{N-p}{N}\right)\sigma^2.$$

As a result, the ML estimate of σ^2 will be biased in small samples and will underestimate σ^2.

In effect, the bias arises because the ML estimate has not taken into account that β, also, is estimated. That is, in the estimator of σ^2 we have replaced β by $\widehat{\beta}$.

It should not be too surprising that similar problems arise in the estimation of Σ.

Recall: An unbiased estimator is given by using $N - p$ as the denominator instead of N.

The theory of residual or restricted maximum likelihood estimation was developed to address this problem.

The main idea behind REML is to eliminate the parameters β from the likelihood so that it is defined only in terms of Σ.

One possible way to obtain the restricted likelihood is to consider transformations of the data to a set of linear combinations of observations that have a distribution that does not depend on β.

For example, the residuals after estimating β by ordinary least squares can be used.

The likelihood for these residuals will depend only on Σ, and not on β.

Thus, rather than maximizing

$$-\frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - X_i\widehat{\beta}\right)'\Sigma^{-1}\left(Y_i - X_i\widehat{\beta}\right)$$

REML maximizes the following slightly modified log-likelihood

$$\begin{aligned} -\frac{N}{2}\ln|\Sigma| \quad &- \quad \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - X_i\widehat{\beta}\right)'\Sigma^{-1}\left(Y_i - X_i\widehat{\beta}\right) \\ &- \quad \frac{1}{2}\ln\left|\sum_{i=1}^{N} X_i'\Sigma^{-1}X_i\right| \end{aligned}$$

When the residual likelihood is maximized, we obtain less biased estimate of Σ.

That is, the extra determinant term effectively makes a correction or adjustments that is analogous to the correction to the denominator in $\hat{\sigma}^2$.

When REML estimation is used, we obtain the GLS estimates of β,

$$\widehat{\beta} = \left[\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}Y_i\right)$$

where $\widehat{\Sigma}$ is the REML estimate of Σ.

Note: The residual maximum likelihood (REML) can be used to compare different models for the covariance structure.

However, it should <u>not</u> be used to compare different regression models since the penalty term in REML depends upon the regression model specification.

Instead, the standard ML log-likelihood should be used for comparing different regression models for the mean.

In PROC MIXED, REML is the default maximization criterion.

ML estimates are obtained by specifying:

PROC MIXED METHOD = ML;

Some Remarks on Missing Data

Missing data arise in longitudinal studies whenever one or more of the sequences of measurements is incomplete, in the sense that some intended measurements are not obtained.

Let $Y^{(o)}$ denote the measurements observed and $Y^{(m)}$ denote the measurements that are missing.

For incomplete data to provide valid inference about a general linear model, the mechanism (probability model) producing the missing observations must satisfy certain assumptions.

Here. we distinguish two different types of missing data mechanisms:

1) Data are missing completely at random (MCAR) when the probability that an individual value will be missing is independent of $Y^{(o)}$ and $Y^{(m)}$. Many methods of analysis are valid when the data are MCAR. Valid methods include maximum likelihood and various ad hoc methods (e.g. 'complete case' analyses).
   Example: 'rotating panel' designs.

2) Data are missing at random (MAR) when the probability that an individual value will be missing is independent of $Y^{(m)}$ (but may depend on $Y^{(o)}$). If this assumption holds, likelihood-based inference is valid, but most ad hoc methods are not.
   Example: subject 'attrition' related to previous performance.

Note: Under assumptions 1) and 2), the missing data mechanism is often referred to as being 'ignorable'.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 6

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Modelling Longitudinal Data

Overview:

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

(1) mean response over time

(2) covariance among repeated measures

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

Modelling the Mean

Two main approaches can be distinguished:

(1) analysis of response profiles

(2) parametric or semi-parametric curves

Modelling the Covariance

Three broad approaches can be distinguished:

(1) "unstructured" or arbitrary pattern of covariance

(2) covariance pattern models

(3) random effects covariance structure

Modelling the Mean: Analysis of Response Profiles

Basic idea: Compare groups of subjects in terms of mean response profiles over time.

Useful for *balanced* longitudinal designs and when there is a single categorical covariate (perhaps denoting different treatment or exposure groups).

Analysis of response profiles can be extended to handle more than a single group factor.

Analysis of response profiles can also handle missing data.

Example

Treatment of Lead-Exposed Children (TLC) Trial

Recall data from TLC trial:

Children randomized to placebo or Succimer.

Measures of blood lead level at baseline, 1, 4 and 6 weeks.

The sequence of means over time in each group is referred to as the “mean response profile”.

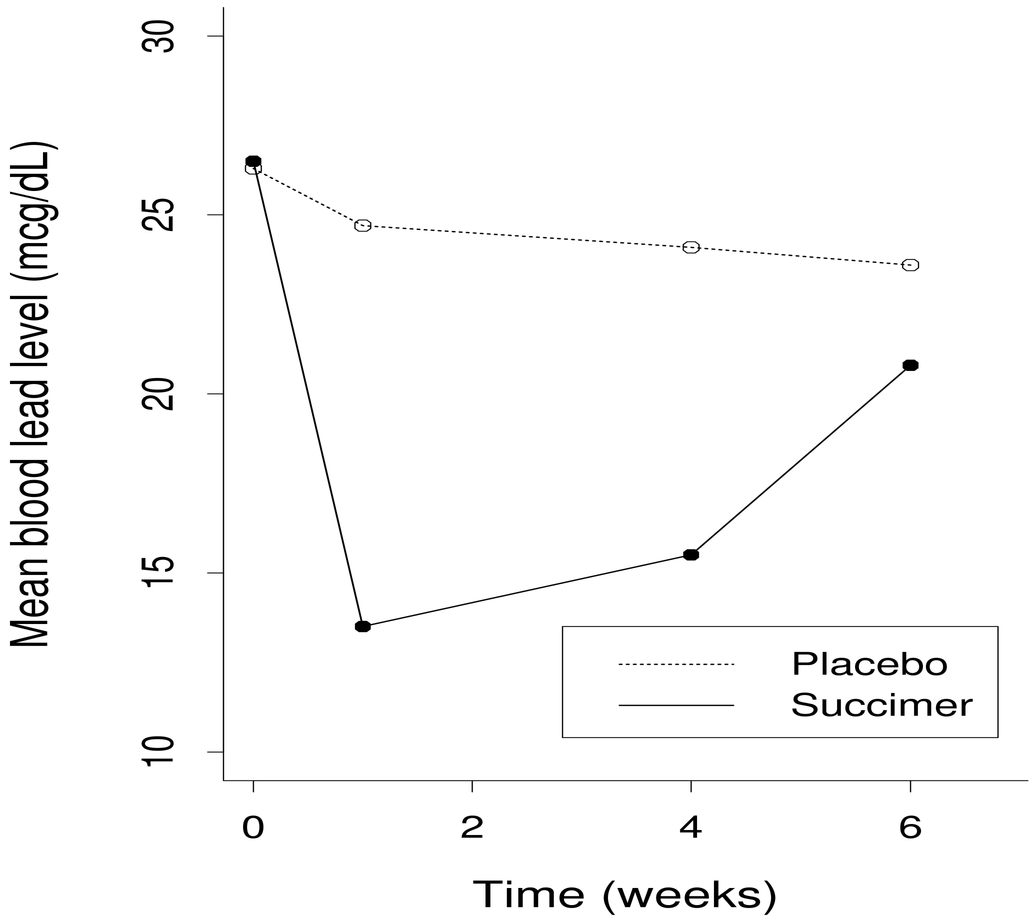


Figure 7: Mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

Hypotheses concerning response profiles

Given a sequence of n repeated measures on a number of distinct groups of individuals, three main questions:

(1) Are the mean response profiles similar in the groups, in the sense that the mean response profiles are parallel?
This is a question that concerns the *group* × *time interaction effect*.

(2) Assuming mean response profiles are parallel, are the means constant over time, in the sense that the mean response profiles are flat?
This is a question that concerns the *time effect*.

(3) Assuming that the population mean response profiles are parallel, are they also at the same level in the sense that the mean response profiles for the groups coincide?
This is a questions that concerns the *group effect*;

Note: For many longitudinal studies, especially longitudinal clinical trials, main interest is in Question 1: *group* × *time interaction effect*.

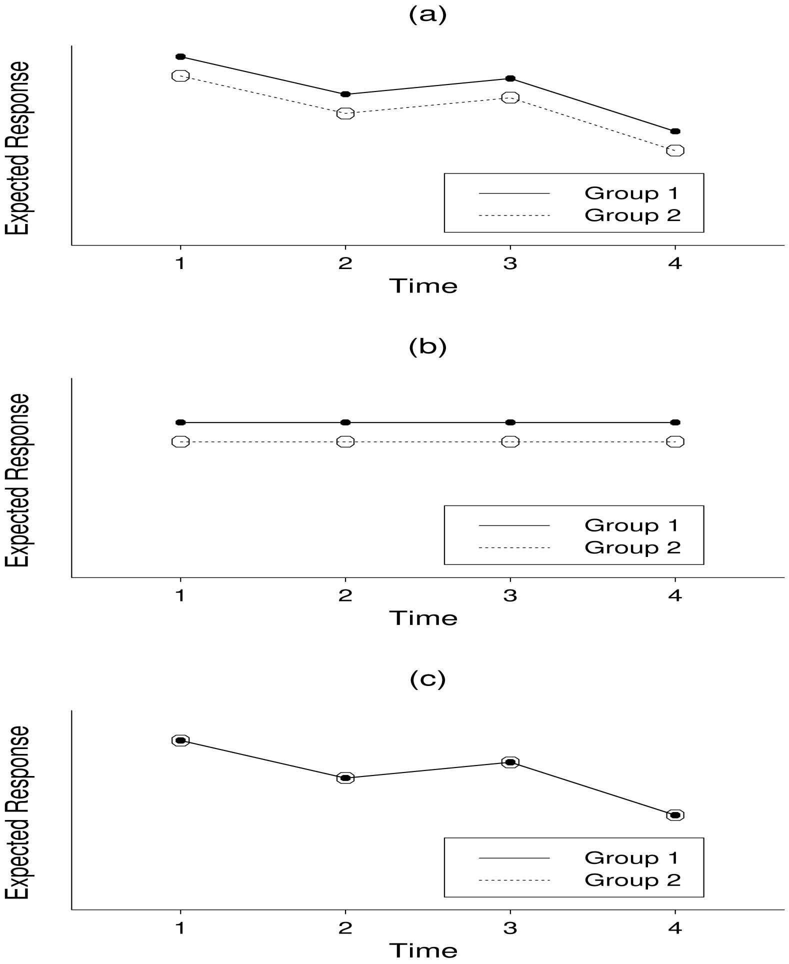


Figure 8: Graphical representation of the null hypotheses of (a) no group × time interaction effect, (b) no time effect, and (c) no group effect.

Table 8: Mean response profile over time in 2 groups.

	Measurement Occasion			
Group	1	2	...	n
1	$\mu_1(1)$	$\mu_2(1)$	...	$\mu_n(1)$
2	$\mu_1(2)$	$\mu_2(2)$	...	$\mu_n(2)$

Next, consider the differences between the group means at each occasion.

Define $\Delta_j = \mu_j(1) - \mu_j(2), \;\; j = 1, ..., n.$

The first hypothesis in an analysis of response profiles can be expressed as:

No *group* × *time interaction effect*:

$$H_0 : \Delta_1 = \Delta_2 = \cdots = \Delta_n.$$

With only 2 groups, the test of the null hypothesis of no group × time interaction effect has $(n - 1)$ degrees of freedom.

Note: Rejection of H_0, no *group* × *time interaction*, indicates groups differ in their patterns of change over time, but does not indicate how they differ.

Table 9: Mean response profile over time in G groups.

	Measurement Occasion			
Group	1	2	...	n
1	$\mu_1(1)$	$\mu_2(1)$	...	$\mu_n(1)$
2	$\mu_1(2)$	$\mu_2(2)$	...	$\mu_n(2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
g	$\mu_1(g)$	$\mu_2(g)$	...	$\mu_n(g)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
G	$\mu_1(G)$	$\mu_2(G)$	...	$\mu_n(G)$

Let G denote the number of groups, with $G \geq 2$.

Define $\Delta_j(g) = \mu_j(g) - \mu_j(G), j = 1, ..., n; g = 1, ..., G - 1$.

With $G \geq 2$, the test of the null hypothesis of no group × time interaction effect can be expressed as:

No *group* × *time interaction effect*:

$$H_{01} : \Delta_1(g) = \Delta_2(g) = \cdots = \Delta_n(g); \quad \text{for} \;\; g = 1, ..., G - 1.$$

With $G \geq 2$, the test of the null hypothesis of no group × time interaction effect has $(G - 1) \times (n - 1)$ degrees of freedom.

Remark on Baseline Measurement

Baseline measurement given same status as post-randomization outcomes.

Alternative methods:

(a) Subtract baseline from each subsequent observation and analyze differences

(b) Use baseline as a covariate

Covariate analysis based on (b) is generally more efficient than analyses of differences for pre-test post-test designs[2].

Note: Covariate analysis requires discarding subjects if there are missing baseline values.

[2] In the next lecture, we discuss alternative methods for handling baseline response.

Model for Variance-Covariance Matrix: Unstructured

Table 10: Assumed covariance matrix in analysis of response profiles.

Covariance Matrix				
σ_1^2	σ_{12}	σ_{13}	$\cdots$	σ_{1n}
σ_{21}	σ_2^2	σ_{23}	$\cdots$	σ_{2n}
σ_{31}	σ_{32}	σ_3^2	$\cdots$	σ_{3n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
σ_{n1}	σ_{n2}	σ_{n3}	$\cdots$	σ_n^2

The main focus of analysis is on a global test of the null hypothesis that the mean response profiles are similar in the groups.

Are the mean response profiles parallel?

This is a question that concerns the *group* × *time interaction effect*.

In testing this hypothesis, both *group* and *time* are regarded as categorical covariates (analogous to two-way ANOVA).

The analysis of response profiles can be specified as a regression model with "indicator variables" for *group* and *time*.

However, unlike standard regression, the correlation and variability among repeated measures on the same individuals must be properly accounted for.

In summary, analysis of response profiles can be specified as a regression model with "indicator variables" for *group* and *time*.

The global test of the null hypothesis of parallel profiles translates into a hypothesis concerning regression coefficients for the *group* × *time* interaction being equal to zero.

Beyond testing the null hypothesis of parallel profiles, the estimated regression coefficients have meaningful interpretations.

Analysis of Response Profiles using PROC MIXED

Note that PROC MIXED in SAS requires each repeated measurement in a longitudinal data set to be a separate "record". For example, in the TLC trial, the data are recorded as follows:

(ID	Group	Baseline	Week 1	Week 4	Week 6)
046	P	30.8	26.9	25.8	23.8
149	A	26.5	14.8	19.5	21.0
096	A	25.8	23.0	19.1	23.2
064	P	24.7	24.5	22.0	22.5
050	A	20.4	2.8	3.2	9.4
210	A	20.4	5.4	4.5	11.9
⋮	⋮	⋮	⋮	⋮	⋮
416	P	31.1	31.2	29.2	30.1

with a single "record" of 4 repeated measurements for each child in study.

The data set is in a *multivariate* mode (or “wide form”).

Prior to analysis, these data must be converted to a data set with 4 records for each child, one for each measurement occasion.

In the latter form, data set is in a *univariate* mode (or “long form”).

This can be accomplished using the illustrative SAS commands in Table 11 which produced the following:

(ID	Group	Time	Y)
046	P	0	30.8
046	P	1	26.9
046	P	4	25.8
046	P	6	23.8
149	A	0	26.5
149	A	1	14.8
149	A	4	19.5
149	A	6	21.0
⋮	⋮	⋮	⋮
416	P	0	31.1
416	P	1	31.2
416	P	4	29.2
416	P	6	30.1

Table 11: Illustrative commands in SAS for transforming data set with single record for each individual to data set with multiple records for each measurement occasion.

```
DATA lead;
   INFILE ’g:\shared\bio226\tlc.txt’;
   INPUT id group $ y1 y2 y3 y4;
   y=y1; time=0; OUTPUT;
   y=y2; time=1; OUTPUT;
   y=y3; time=4; OUTPUT;
   y=y4; time=6; OUTPUT;
   DROP y1-y4;
```

Table 12: Illustrative commands for an analysis of response profiles using PROC MIXED in SAS.

```
PROC MIXED ORDER=DATA;
   CLASS id group time;
   MODEL y=group time group*time /S CHISQ;
   REPEATED time / TYPE=UN SUBJECT=id R RCORR;
```

Case Study

Analysis of Response Profiles

Treatment of Lead-Exposed Children Trial

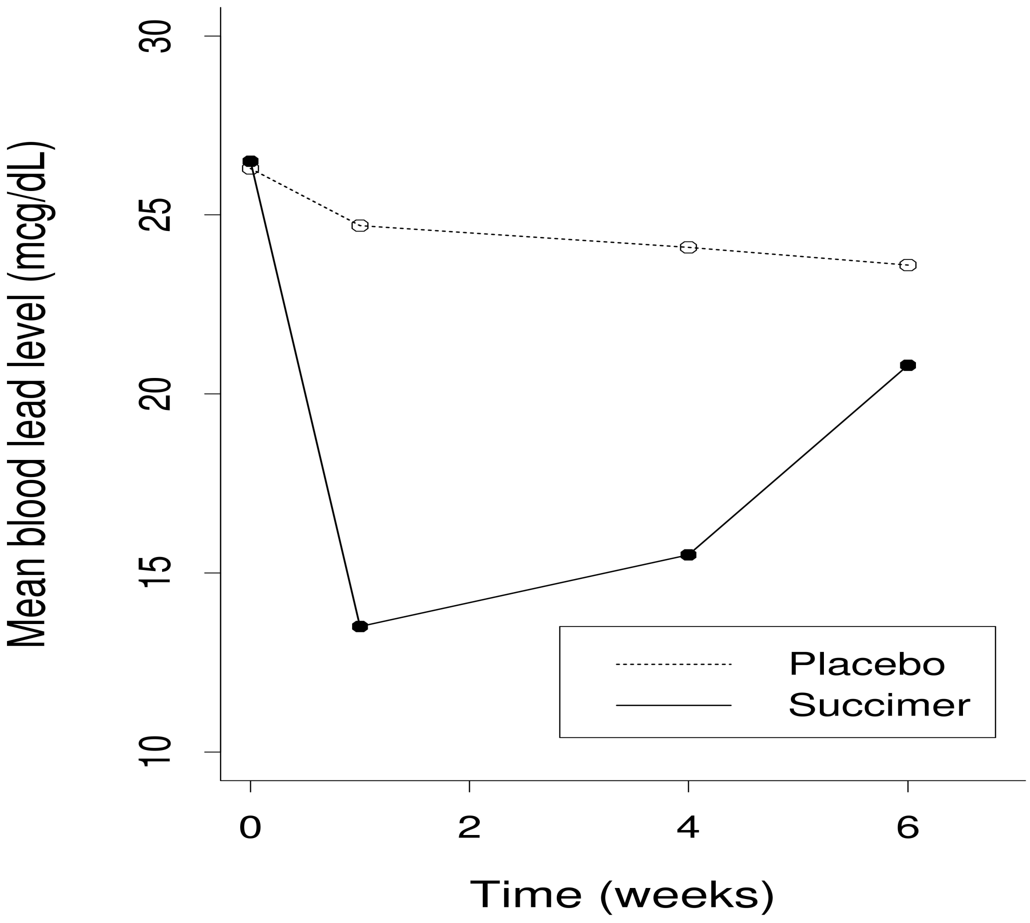


Figure 9: Plot of mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

Recall, the main focus of analysis is on a global test of the null hypothesis that the mean response profiles are similar in the groups.

Are the mean response profiles parallel?

This is a question that concerns the *group* × *time interaction effect.*

In testing this hypothesis, both *group* and *time* are regarded as categorical covariates (analogous to two-way ANOVA).

The analysis of response profiles can be specified as a regression model with "indicator variables" for *group* and *time.*

Choice of Reference Level

The usual choice of reference group:

(i) A natural baseline or comparison group, and/or

(ii) group with largest sample size

In longitudinal data setting, the "baseline" or first measurement occasion is a natural reference group for "time".

Treatment of Lead-Exposed Children Trial

In the TLC Trial there are two groups (placebo and succimer) and four measurement occasions (week 0, 1, 4, 6).

Let $X_1 = 1$ for all children at all occasions.

Creating indicator variables for group and time:

Group:

Let $X_2 = 1$ if child randomized to succimer, $X_2 = 0$ otherwise.

Time:

Let $X_3 = 1$ if measurement at week 1, $X_3 = 0$ otherwise
Let $X_4 = 1$ if measurement at week 4, $X_4 = 0$ otherwise
Let $X_5 = 1$ if measurement at week 6, $X_5 = 0$ otherwise

Recall: Hypothesis of main interest concerns *group* × *time interaction effect*.

Analysis of response profiles model can be expressed as:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_2 * X_3 + \beta_7 X_2 * X_4 + \beta_8 X_2 * X_5 + e$$

Test of *group* × *time interaction*: $H_0 : \beta_6 = \beta_7 = \beta_8 = 0$.

The analysis must also account for the correlation among repeated measures on the same child.

The analysis of response profiles estimates separate variances for each occasion (4 variances) and six pairwise correlations.

Treatment of Lead-Exposed Children Trial

Table 13 displays estimates of the covariance matrix.

Note the discernible increase in the variability in blood lead levels from pre- to post-randomization.

This increase in variability from baseline is probably due to:

(1) given the treatment group assignment, there may be natural heterogeneity in the individual response trajectories over time,

(2) the trial had an inclusion criterion that blood lead levels at baseline were in the range of 20-44 micrograms/dL.

Table 13: Estimated covariance matrix for the blood lead levels at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Covariance Matrix			
25.2	19.1	19.7	22.2
19.1	44.3	35.5	29.7
19.7	35.5	47.4	30.6
22.2	29.7	30.6	58.7

Table 14: Estimated correlation matrix for the blood lead levels at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

Correlation Matrix			
1.00	0.57	0.57	0.58
0.57	1.00	0.78	0.58
0.57	0.78	1.00	0.58
0.58	0.58	0.58	1.00

Table 15: Tests of fixed effects based on analysis of response profiles of the blood lead level data at baseline, weeks 1, 4, and 6.

Variable	DF	Chi-Squared	*P*-Value
Group	1	25.43	<0.0001
Week	3	184.48	<0.0001
Group × Week	3	107.79	<0.0001

Test of the group × time interaction is based on (multivariate) Wald test (comparison of estimates to SEs).

In the TLC trial, question of main interest concerns comparison of two treatment groups in terms of their patterns of change from baseline.

This question translates into test of group × time interaction.

The test of the group × time interaction yields a Wald statistic of 107.79 with 3 degrees of freedom ($p < 0.0001$).

Because this is a global test, it indicates that groups differ but does not tell us how they differ.

Recall, analysis of response profiles model can be expressed as:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_2 * X_3 + \beta_7 X_2 * X_4 + \beta_8 X_2 * X_5 + e$$

Test of *group* × *time interaction*: $H_0 : \beta_6 = \beta_7 = \beta_8 = 0$.

The 3 single df contrasts for group × time interaction have direct interpretations in terms of group comparisons of changes from baseline.

They indicate that children treated with succimer have greater decrease in mean blood lead levels from baseline at all occasions when compared to children treated with placebo (see Table 16).

Table 16: Estimated regression coefficients and standard errors based on analysis of response profiles of the blood lead level data.

Variable	Group	Week	Estimate	SE	Z
Intercept			26.272	0.710	36.99
Group	A		0.268	1.005	0.27
Week		1	−1.612	0.792	−2.04
Week		4	−2.202	0.815	−2.70
Week		6	−2.626	0.889	−2.96
Group × Week	A	1	−11.406	1.120	−10.18
Group × Week	A	4	−8.824	1.153	−7.66
Group × Week	A	6	−3.152	1.257	−2.51

Summary

“Analysis of response profiles” can be framed as a linear regression with correlated observations.

Extensions beyond the usual profile analysis:

missing observations

baseline covariates

time contrasts

area-under-the-curve-analyses

Strengths and Weaknesses of Analysis of Response Profiles

Strengths:

Allows arbitrary patterns in the mean response over time (no time trend assumed) and arbitrary patterns in the covariance.

Analysis has a certain robustness since potential risks of bias due to misspecification of models for mean and covariance are minimal.

Can accommodate an arbitrary pattern of missingness.

Drawbacks:

Requirement that the longitudinal design be balanced.

Analysis cannot incorporate mistimed measurements.

Analysis ignores the time-ordering (time trends) of the repeated measures in a longitudinal study.

Produces omnibus tests of effects that may have low power to detect group differences in specific trends in the mean response over time (e.g., linear trends in the mean response).

The number of estimated parameters, $G \times n$ mean parameters and $\frac{n(n+1)}{2}$ covariance parameters (variances and correlations), grows rapidly with the number of measurement occasions.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 7

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Single Degree of Freedom Contrasts

Recall: A drawback of analysis of response profiles is that it ignores the time-ordering (time trends) of the repeated measures in a longitudinal study.

Moreover, it produces omnibus tests of effects that may have low power to detect group differences in specific trends in the mean response over time.

In a certain sense, an omnibus test disperses statistical power among too many alternatives.

This lack of specificity is potentially a problem in studies with a large number of measurement occasions.

Recall that the test for group × time interaction has $(G-1) \times (n-1)$ degrees of freedom.

This general test becomes less sensitive to an interaction with a specific pattern as n increases.

One-Degree-of-Freedom Tests for Group by Time Interaction

In typical randomized trial, subjects are randomized to the intervention groups at baseline.

Investigator seeks to determine whether the pattern of response after randomization differs between groups.

A more powerful test of group by time interaction is obtained by specifying a single contrast that best represents direction in which patterns of response differ most markedly.

Example 1: Test for equality of the difference between the average response at occasions 2 through n and the baseline value in the two groups, using the contrast

$$L = (-L_1, L_1),$$

where

$$L_1 = \left(-1, \frac{1}{n-1}, \frac{1}{n-1}, \ldots, \frac{1}{n-1}\right).$$

Here, L_1 computes the mean response from occasions 2 through n and subtracts the mean response at baseline for a single group.

L is a group contrast of this average change in the two groups.

Example 2: A variant of this approach, known as *Area Under the Curve Minus Baseline*, or sometimes simply AUC.

AUC is the area under the trapezoidal curve created by connecting the mean responses at each occasion.

The AUC of the profile of blood lead levels for a single subject in the TLC trial is shown in Figure 10.

Can subtract baseline mean: $\mu_1 \times (t_n - t_1)$, the area of the rectangle of height μ_1 and width $t_n - t_1$.

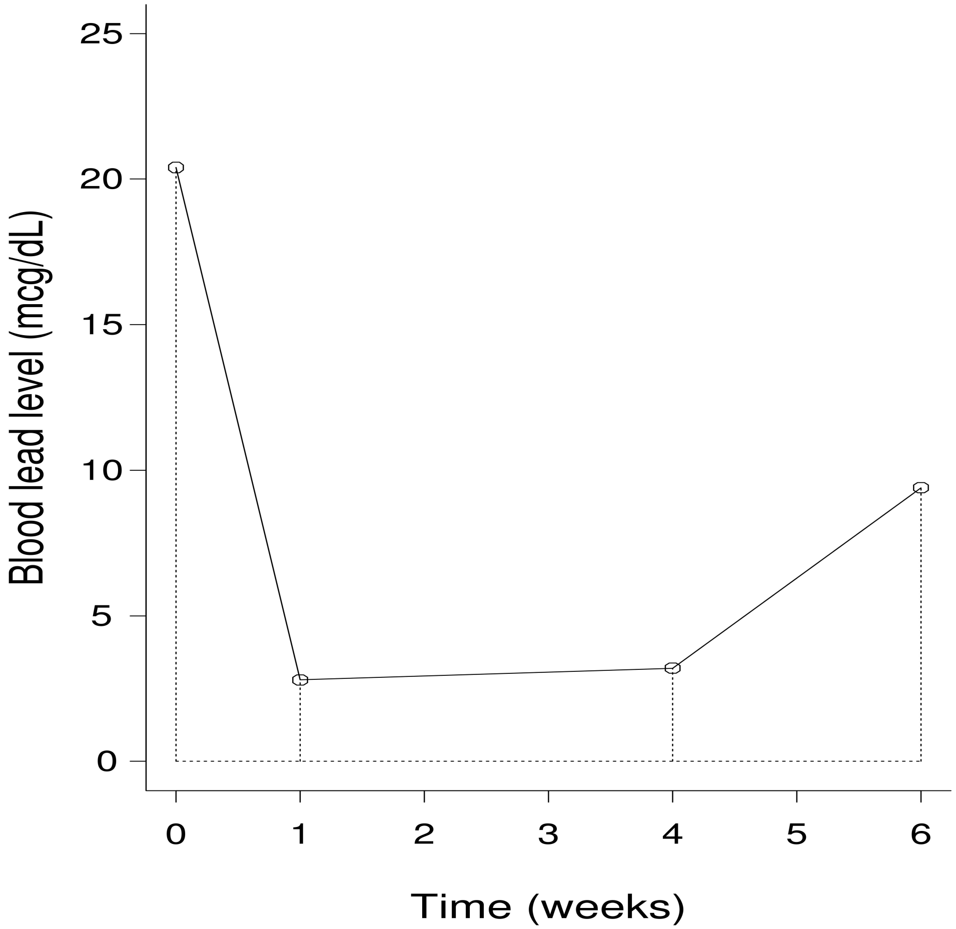


Figure 10: Area under the curve, calculated using the trapezoidal rule, for the profile of blood lead levels for a single subject in the TLC trial.

The test for equality of the AUC in two groups employs the contrast

$$L = (-L_2, L_2),$$

where

$$L_2 = \frac{1}{2} \times (t_1 + t_2 - 2\,t_n,\ t_3 - t_1, \ldots, t_{j+1} - t_{j-1}, \ldots, t_n - t_{n-1})$$

and $\frac{1}{2} \times (t_{j+1} - t_{j-1})$ is the value of the contrast vector for time points other than 1 (baseline) or n (the last occasion).

Note: These contrast weights are not intuitively obvious, but can be derived from formula for area of a trapezoid.

Example 3: A third popular method for constructing a single-degree-of-freedom test corresponds to a test of the hypothesis that the trend over time (e.g., linear) is the same in the groups.

A test of linear trend corresponds to

$$L = (-L_3, L_3),$$

where, for example,

$$L_3 = (-2, -1, 0, 1, 2)\,; \quad \text{when } n = 5.$$

Because this method is a special case of fitting parametric curves, we defer a discussion of this approach until next lecture.

Application to the Treatment of Lead-Exposed Children Trial

Recall: TLC trial measured blood lead levels at four occasions.

The vector representing the contrast based on the mean response at times 2 through n minus baseline is given by

$$L = (-L_1, L_1) = \left(1, -\frac{1}{3},\ -\frac{1}{3},\ -\frac{1}{3},\ -1,\ \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

We can apply the contrast weights to the means in Table 17.

Table 17: Mean blood lead levels (and standard deviation) at baseline, week 1, week 4, and week 6 for the children from the TLC trial.

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	23.6
	(5.0)	(5.5)	(5.8)	(5.6)

For the succimer group, the average value of the mean response minus baseline is -9.90.

For the placebo group, the average value of the mean response minus baseline is -2.17.

L, the group contrast, is 7.73 and the value of the Wald test statistic is $Z = 8.21$ (or $W^2 = 67.4$, with one degree of freedom).

This indicates a highly significant difference in the response pattern between treatment groups.

Similarly, because the time points in the TLC trial were 0, 1, 4, and 6 weeks, the contrast for comparing the AUC (minus baseline) in the two treatment groups is given by

$$L = (-L_2, L_2) = (5.5,\ -2,\ -2.5,\ -1,\ -5.5,\ 2,\ 2.5,\ 1).$$

The estimated mean AUC is -59.20 in the succimer group.

The estimated mean AUC is -11.40 in the placebo group.

L, the group contrast, is 47.8 and the value of the Wald statistic is $Z = 8.97$ (or $W^2 = 80.5$, with one degree of freedom).

Thus both methods of analysis provide a clear signal that the response profile differs in the two treatment groups.

Summary

In many applications, one-degree-of-freedom tests have increased sensitivity to group differences.

For valid application, the form of the contrast must be specified prior to data analysis.

Otherwise, one would be at risk of seeking the best contrast and testing its significance as if it had been chosen in advance.

Adjustment for Baseline Response

When data are complete (no missing data), the one-degree-of-freedom tests described earlier can be constructed as follows:

(1) calculate a univariate summary statistic for each subject

(2) perform a test for equality of means of these summary statistics in the G groups (e.g., t-test, ANOVA).

Furthermore, note that for the two tests described earlier, the summary statistic corresponds to subtracting the baseline value from a summary of the responses on occasions 2 through n.

For example, for the test for equality of mean response minus baseline, the summary statistic for i^{th} subject is

$$\frac{(Y_{i2} + Y_{i3} + \cdots + Y_{in})}{n-1} - Y_{i1}.$$

This suggests an alternative approach analogous to analysis of covariance (ANCOVA).

In the ANCOVA, the summary of the response at times 2 through n becomes the dependent variable and the baseline value becomes a covariate in the analysis.

For example, with two groups, the ANCOVA model is

$$Y_i^* = \beta_1 + \beta_2 Y_{i1} + \beta_3 \, \texttt{trt}_i + e_i^*,$$

where

$$Y_i^* = \frac{(Y_{i2} + Y_{i3} + \cdots + Y_{in})}{n-1},$$

$\texttt{trt}_\texttt{i}$ is an indicator for group, and e_i^* is the error term in the univariate model.

The ANOVA or ANCOVA analysis will be appealing where initial changes from baseline are expected to persist throughout the duration of follow up (see Figure 11).

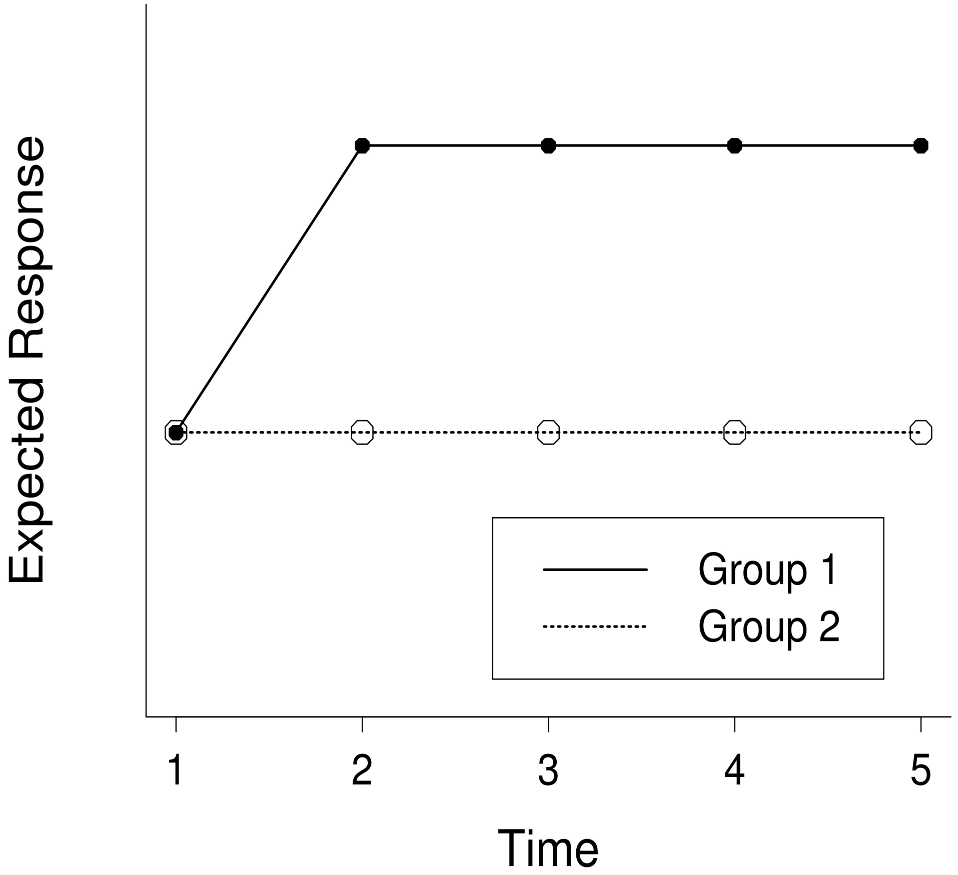


Figure 11: Graphical representation of changes in the mean response from baseline (in Group 1) that persist throughout the duration of follow up.

How best to adjust for baseline in the analysis?

Through a contrast (ANOVA) or via ANCOVA?

The answer depends on the study design: observational versus randomized study.

For observational study, usually not advisable to employ ANCOVA approach because baseline value may be associated with other variables whose effects are to be studied.

Example: Observational study comparing rates of decline of pulmonary function in asthmatics and non-asthmatics.

Suppose asthmatics have lower pulmonary function at all ages, but rates of decline are equal for asthmatics and non-asthmatics.

Suppose the model that best describes the data is:

$$Y_{ij} = \beta_1 + \beta_2 \text{Asthma}_i + \beta_3 \text{Age}_{ij} + e_{ij}$$

Thus the model for the non-asthmatics is,

$$E(Y_{ij}) = \beta_1 + \beta_3 \text{Age}_{ij}$$

and the model for the asthmatics is,

$$E(Y_{ij}) = (\beta_1 + \beta_2) + \beta_3 \text{Age}_{ij}$$

Clearly, the rate of change or decline, expressed by β_3, is the same in the two groups.

As a result, an analysis that compares the decline in the two groups would conclude that there are no differences.

However, if we introduce the baseline value as a covariate, the model is:

$$Y_{ij} = \beta_1 + \beta_2 \text{Asthma}_i + \beta_3 \text{Age}_{ij} + \beta_4 Y_{i1} + e_{ij}$$

This model gives the predicted values for asthmatics and non-asthmatics relative to a common baseline value.

As a result, the decline in pulmonary function for the asthmatics will appear to be greater than the decline for the non-asthmatics.

Why?

Note that the analysis with baseline value as a covariate addresses a somewhat different question.

It considers the conditional question:

“Is an asthmatic expected to show the same decline in pulmonary function as a non-asthmatic, given they both have the same initial level of pulmonary function?”

The answer to this questions is a resounding “No”.

The asthmatic will be expected to decline more.

Why?

If she is initially at the same level of pulmonary function as the non-asthmatic,

(1) either her level of function is very high and can be expected to decline or regress to the mean level for asthmatics, or

(2) the non-asthmatic's level of function is very low and can be expected to increase or regress to the mean level for non-asthmatics

As a result, the rates of decline, conditional on the same initial value, will not be the same in the two groups.

When subjects have been randomized to groups and the baseline value has been obtained before any interventions, adjustment for baseline through ANCOVA is of interest.

In a randomized study, the mean response at baseline is independent of treatment assignment.

In that setting, it can be shown that the 1 df test based on a contrast and the test based on ANCOVA represent alternative tests of the same null hypothesis.

Moreover, the ANCOVA approach will always be more efficient, yielding estimates of treatment group effects with smaller standard errors than those obtained by calculating contrasts.

The greater efficiency of ANCOVA can be highlighted by examining the relative efficiency (or ratio of variances) in simple settings.

Suppose the variance is homogeneous, with common variance σ^2, and the correlation between any pair of repeated measures is ρ, the relative efficiency is:

$$\frac{1}{n}\left\{1 + (n-1)\,\rho\right\}.$$

The greater efficiency of ANCOVA depends on both the number of repeated measures and magnitude of ρ.

For example, when $n = 5$ and $\rho = 0.4$ the analysis of covariance is approximately twice as efficient as subtracting the baseline response.

Alternative Adjustments for Baseline Response

The notion of adjustment for baseline can be applied more generally in the analysis of response profiles.

We consider four ways of handling the baseline value:

(1) Retain it as part of the outcome vector and make no assumptions about group differences in the mean response at baseline.
(2) Retain it as part of the outcome vector and assume the group means are equal at baseline, as might be appropriate in a randomized trial.
(3) Subtract the baseline response from all of the remaining post-baseline responses, and analyze the differences from baseline.
(4) Use baseline value as a covariate in the analysis of the post-baseline responses.

The first method retains the baseline response as part of the outcome vector.

This method produces the standard analysis of response profile results (see Tables 18 and 19).

The test of the group × time interaction from this model yields a Wald statistic of 107.79, with 3 degrees of freedom.

Table 18: Tests of fixed effects based on a profile analysis of the blood lead level data at baseline, weeks 1, 4, and 6.

Variable	DF	Chi-Squared	*P*-Value
Group	1	25.43	<0.0001
Week	3	184.48	<0.0001
Group × Week	3	107.79	<0.0001

Table 19: Estimated regression coefficients and standard errors based on analysis of response profiles of the blood lead level data.

Variable	Group	Week	Estimate	SE	Z
Intercept			26.272	0.710	36.99
Group	A		0.268	1.005	0.27
Week		1	−1.612	0.792	−2.04
Week		4	−2.202	0.815	−2.70
Week		6	−2.626	0.889	−2.96
Group × Week	A	1	−11.406	1.120	−10.18
Group × Week	A	4	−8.824	1.153	−7.66
Group × Week	A	6	−3.152	1.257	−2.51

The second method also retains the baseline response as part of the outcome vector.

This method corresponds to an analysis of response profiles where the group means at baseline are constrained to be equal.

Implemented by excluding the treatment group main effect from the model for the response profiles (see Table 20).

Note: Baseline (week 0) must be chosen as the reference level for time.

The test of the group × time interaction yields a Wald statistic of 111.96, with 3 degrees of freedom.

Table 20: Estimated regression coefficients and standard errors based on an analysis of response profiles of the blood lead level data assuming equal mean blood lead levels at baseline.

Variable	Group	Week	Estimate	SE	Z
Intercept			26.406	0.500	52.83
Week		1	−1.645	0.782	−2.10
Week		4	−2.231	0.807	−2.76
Week		6	−2.642	0.887	−2.98
Group × Week	A	1	−11.341	1.093	−10.38
Group × Week	A	4	−8.765	1.131	−7.75
Group × Week	A	6	−3.120	1.251	−2.49

The third method does not retain the baseline response as part of the outcome vector.

Baseline response is subtracted from post-baseline responses and analysis is based on these differences from baseline,

$$D_i = (Y_{i2} - Y_{i1}, Y_{i3} - Y_{i1}, \ldots, Y_{in} - Y_{i1})'.$$

Because outcome is a change score, this alters interpretation of the tests for all three effects.

Test for group × time interaction becomes a test for parallel profiles for the *changes* from baseline.

Test for group effect becomes a test that *changes* from baseline at occasion 2 are the same across groups (assuming occasion 2 is reference level).

Table 21: Estimated regression coefficients and standard errors based on an analysis of response profiles of the changes from baseline in blood lead levels at week 1, week 4, and week 6.

Variable	Group	Week	Estimate	SE	Z
Intercept			−1.612	0.792	−2.04
Group	A		−11.406	1.120	−10.18
Week		4	−0.590	0.643	−0.92
Week		6	−1.014	0.934	−1.09
Group × Week	A	4	2.582	0.909	2.84
Group × Week	A	6	8.254	1.321	6.25

Thus, the original test of "parallelism of profiles" now becomes a joint test of main effect of group and the group × time interaction.

Formally equivalent to the test of parallelism in standard analysis of response profiles.

Thus, first and third methods are completely equivalent.

The fourth method does not retain the baseline response as part of the outcome vector.

Instead, it focuses on *adjusted* changes from baseline and restricts the outcome vector to measurements obtained post-baseline.

Similar to the third method, test of interest is a joint test of main effect of group and the group × time interaction.

This yields a Wald statistic of 111.13, with 3 degrees of freedom.

Table 22: Estimated regression coefficients and standard errors based on an analysis of response profiles of the adjusted changes from baseline in blood lead levels at week 1, week 4, and week 6.

Variable	Group	Week	Estimate	SE	Z
Intercept			−1.638	0.777	−2.11
Baseline[†] ($Y_{i1} - 26.406$)			−0.196	0.094	−2.08
Group	A		−11.354	1.099	−10.34
Week		4	−0.590	0.643	−0.92
Week		6	−1.014	0.934	−1.09
Group × Week	A	4	2.582	0.909	2.84
Group × Week	A	6	8.254	1.321	6.25

[†] Centering baseline response on its overall mean (26.406) gives the intercept a meaningful interpretation.

Summary

In general, randomized studies are the only setting where we recommend adjustment for baseline through analysis of covariance.

In randomized studies, such an adjustment leads to meaningful tests of hypothesis of scientific interest.

Moreover, the tests based on the analysis of covariance approach will be more powerful.

Alternatively, and almost equivalently, can retain baseline as part of outcome vector and assume group means are equal at baseline.

Not advisable to make the above adjustments in observational studies.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 8

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Modelling the Mean: Parametric Curves

Fitting parametric or semi-parametric curves to longitudinal data can be justified on substantive and statistical grounds.

Substantively, in many studies true underlying mean response process changes over time in a relatively smooth, monotonically increasing/decreasing pattern.

Fitting parsimonious models for mean response results in statistical tests of covariate effects (e.g., treatment × time interactions) with greater power than in analysis of response profiles.

Polynomial Trends in Time

Describe the patterns of change in the mean response over time in terms of simple polynomial trends.

The means are modelled as an explicit function of time.

This approach can handle highly unbalanced designs in a relatively seamless way.

For example, mistimed measurements are easily incorporated in the model for the mean response.

Linear Trends over Time

Simplest possible curve for describing changes in the mean response over time is a straight line.

Slope has direct interpretation in terms of a constant rate of change in mean response for a single unit change in time.

Consider two-group study comparing *treatment* and *control*, where changes in mean response are approximately linear:

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Group}_i + \beta_4 \text{Time}_{ij} \times \text{Group}_i,$$

where $\text{Group}_i = 1$ if i^{th} individual assigned to treatment, and $\text{Group}_i = 0$ otherwise; and Time_{ij} denotes measurement time for the j^{th} measurement on i^{th} individual.

Model for the mean for subjects in control group:

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \mathrm{Time}_{ij},$$

while for subjects in treatment group,

$$E\left(Y_{ij}\right) = \left(\beta_1 + \beta_3\right) + \left(\beta_2 + \beta_4\right) \mathrm{Time}_{ij}.$$

Thus, each group's mean response is assumed to change linearly over time (see Figure 12).

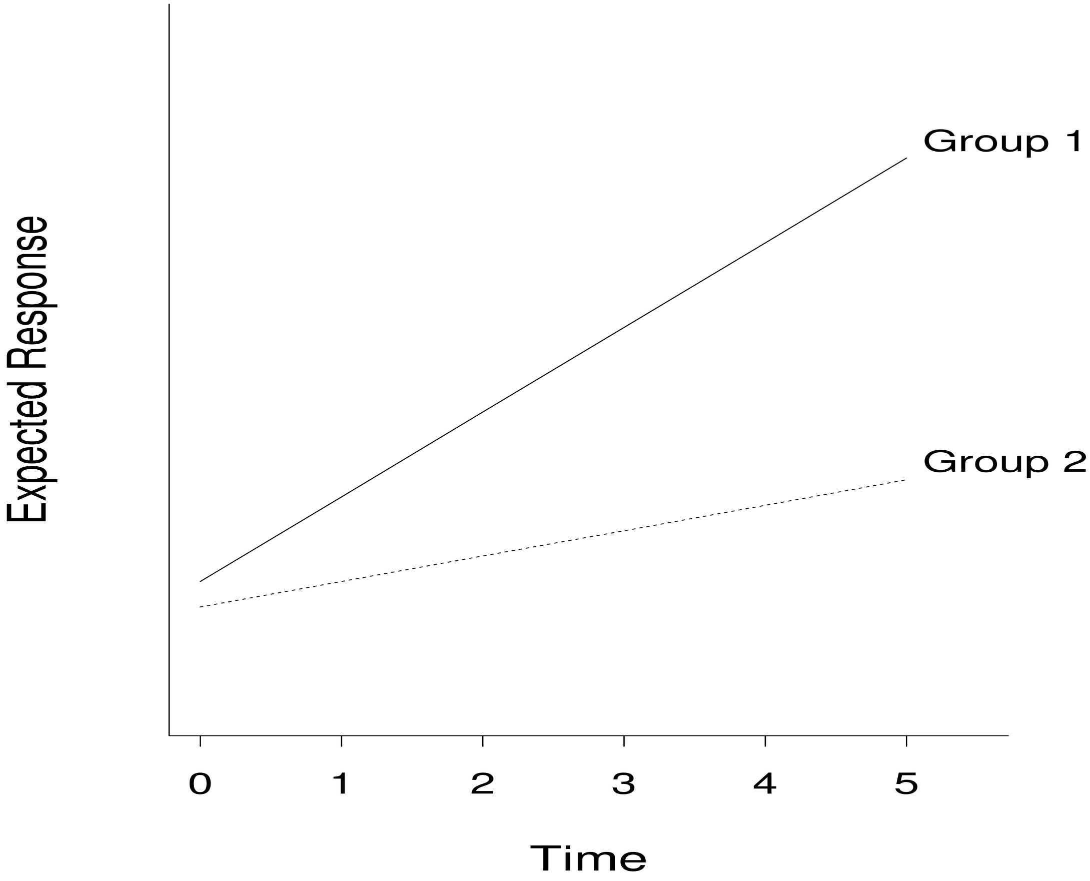


Figure 12: Graphical representation of model with linear trends for two groups.

Quadratic Trends over Time

When changes in the mean response over time are not linear, higher-order polynomial trends can be considered.

For example, if the means are monotonically increasing or decreasing over the course of the study, but in a curvilinear way, a model with quadratic trends can be considered.

In a quadratic trend model the rate of change in the mean response is not constant but depends on time.

Rate of change must be expressed in terms of two parameters.

Consider two-group study example:

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Time}_{ij}^2 + \beta_4 \text{Group}_i \\ &\quad + \beta_5 \text{Time}_{ij} \times \text{Group}_i + \beta_6 \text{Time}_{ij}^2 \times \text{Group}_i . \end{aligned}$$

Model for subjects in control group:

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Time}_{ij}^2 ;$$

while model for subjects in treatment group:

$$E\left(Y_{ij}\right) = \left(\beta_1 + \beta_4\right) + \left(\beta_2 + \beta_5\right) \text{Time}_{ij} + \left(\beta_3 + \beta_6\right) \text{Time}_{ij}^2 .$$

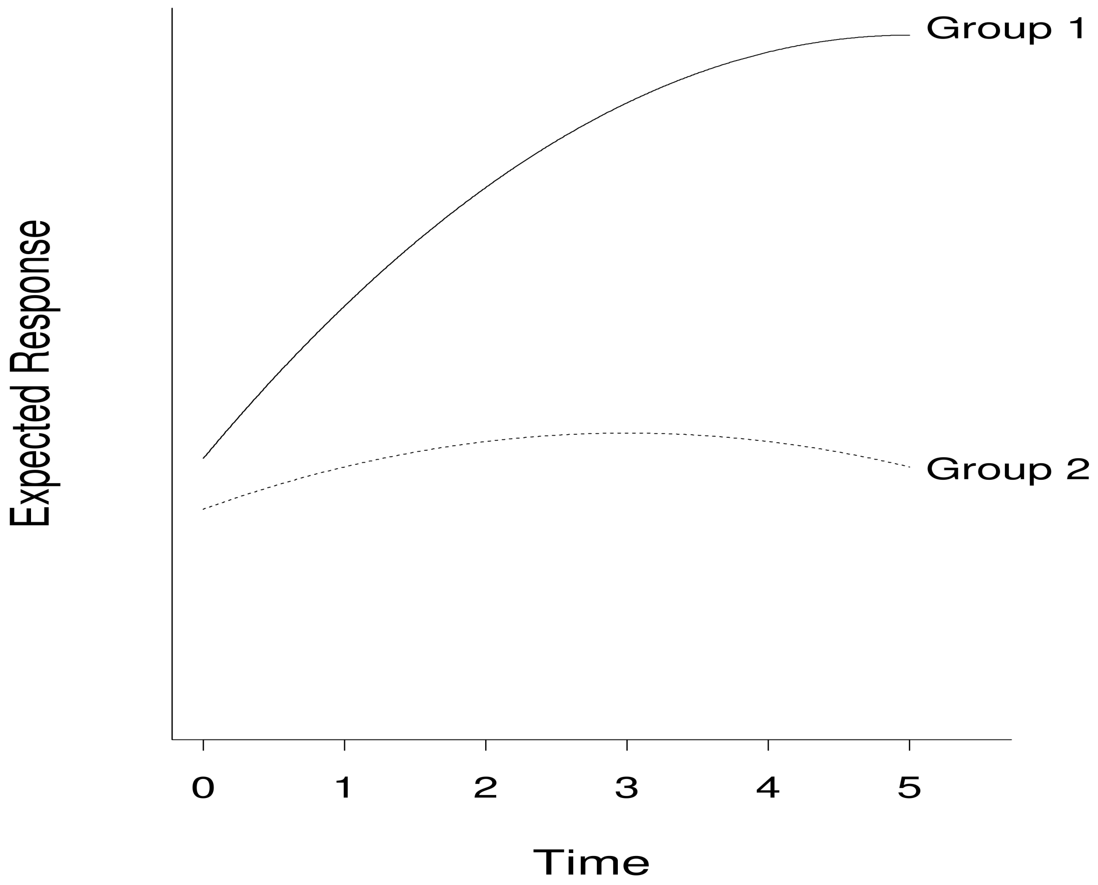


Figure 13: Graphical representation of model with quadratic trends for two groups.

Note: mean response changes at different rate, depending upon Time_{ij}.

Rate of change in control group is $\beta_2 + 2\beta_3\text{Time}_{ij}$
(derivation of this instantaneous rate of change straightforward with calculus).

Thus, early in the study when $\text{Time}_{ij} = 1$, rate of change is $\beta_2 + 2\beta_3$; while later in the study, say $\text{Time}_{ij} = 4$, rate of change is $\beta_2 + 8\beta_3$.

Regression coefficients, $(\beta_2 + \beta_5)$ and $(\beta_3 + \beta_6)$, have similar interpretations for treatment group.

“Centering”

To avoid problems of collinearity it is advisable to “center” Time_j on its mean value prior to the analysis.

Replace Time_j by its deviation from the mean of $(\text{Time}_1,\text{Time}_2,...,\text{Time}_n)$.

Note: centering of Time_{ij} at individual-specific values (e.g., the mean of the n_i measurement times for i^{th} individual) should be avoided, as the interpretation of the intercept becomes meaningless.

Linear Splines

If simplest curve is a straight line, then one way to extend the curve is to have sequence of joined line segments that produces a piecewise linear pattern.

Linear spline models provide flexible way to accommodate many non-linear trends that cannot be approximated by simple polynomials in time.

Basic idea: Divide time axis into series of segments and consider piecewise-linear trends, having different slopes but joined at fixed times.

Locations where lines are tied together are known as “knots”.

Resulting piecewise-linear curve is called a spline.

Piecewise-linear model often called “broken-stick” model.

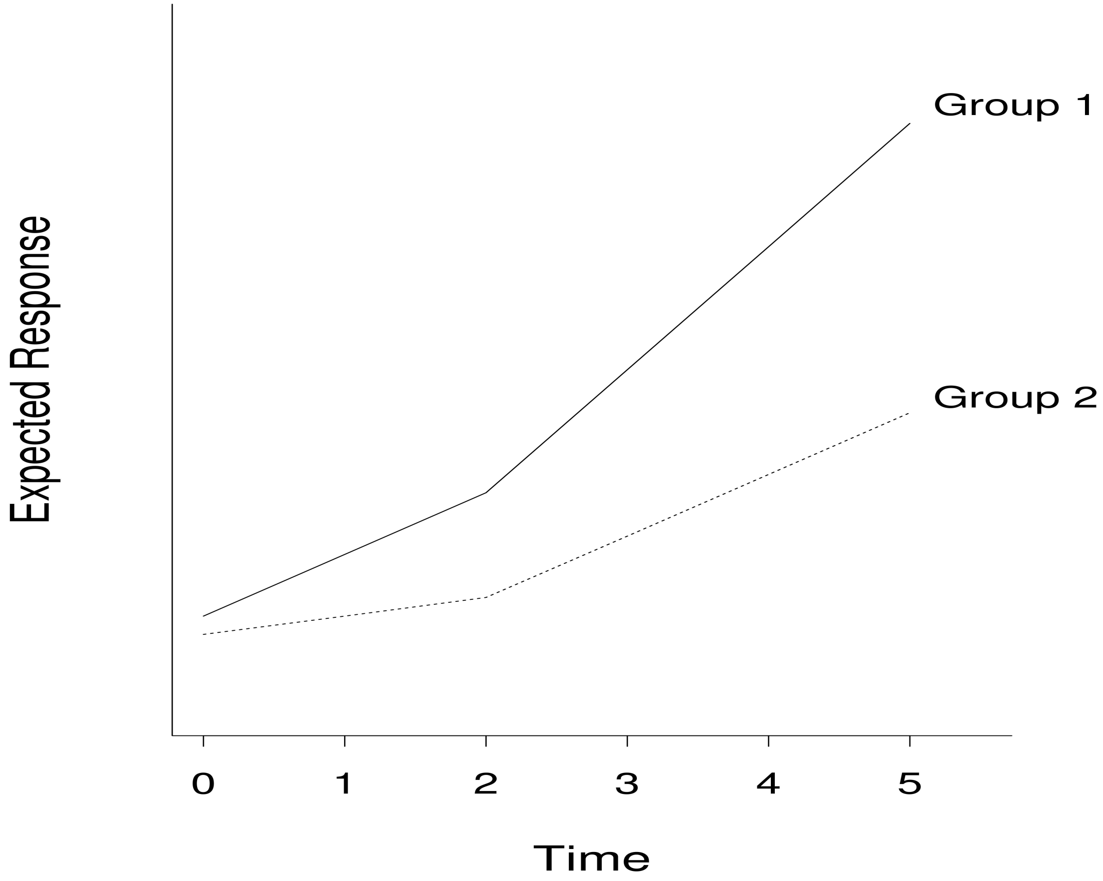


Figure 14: Graphical representation of model with linear splines for two groups, with common knot.

The simplest possible spline model has only one knot.

For two-group example, linear spline model with knot at t^*:

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_1 + \beta_2 \mathrm{Time}_{ij} + \beta_3 (\mathrm{Time}_{ij} - t^*)_+ + \beta_4 \mathrm{Group}_i \\ &\quad + \beta_5 \mathrm{Time}_{ij} \times \mathrm{Group}_i + \beta_6 (\mathrm{Time}_{ij} - t^*)_+ \times \mathrm{Group}_i, \end{aligned}$$

where $(x)_+$ is defined as a function that equals x when x is positive and is equal to zero otherwise.

Thus, $(\mathrm{Time}_{ij} - t^*)_+$ is equal to $(\mathrm{Time}_{ij} - t^*)$ when $\mathrm{Time}_{ij} > t^*$ and is equal to zero when $\mathrm{Time}_{ij} \leq t^*$.

Model for subjects in control group:

$$E\,(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 (\text{Time}_{ij} - t^*)_+.$$

When expressed in terms of mean response prior/after t^*:

$$\begin{aligned} E\,(Y_{ij}) &= \beta_1 + \beta_2 \text{Time}_{ij}, && \text{Time}_{ij} \leq t^*; \\ E\,(Y_{ij}) &= (\beta_1 - \beta_3 t^*) + (\beta_2 + \beta_3) \text{Time}_{ij}, && \text{Time}_{ij} > t^*. \end{aligned}$$

Slope prior to t^* is β_2 and following t^* is $(\beta_2 + \beta_3)$.

Model for subjects in treatment group:

$$E\left(Y_{ij}\right) \quad = \quad (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\text{Time}_{ij} + (\beta_3 + \beta_6)(\text{Time}_{ij} - t^*)_+.$$

When expressed in terms of mean response prior/after t^*:

$$\begin{aligned} E\left(Y_{ij}\right) &= (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\text{Time}_{ij}, & \text{Time}_{ij} \leq t^*; \\ E\left(Y_{ij}\right) &= [(\beta_1 + \beta_4) - (\beta_3 + \beta_6)t^*)] & \\ &\quad + (\beta_2 + \beta_3 + \beta_5 + \beta_6)\text{Time}_{ij}, & \text{Time}_{ij} > t^*. \end{aligned}$$

"Constant Effect" Model

In previous lecture, we discussed a simple model where an exposure or treatment might cause a shift in the mean response that remains constant across measurement occasions (e.g., Figure 11 on slide 212).

To fit such a model, we can create a new variable for time:

$\text{Posttime}_{ij} = 0$ if baseline ($\text{Time}_{ij} = 0$),
$\text{Posttime}_{ij} = 1$ if post-baseline ($\text{Time}_{ij} > 0$).

Then, in two group setting, the model is:

$$E(Y_{ij}) = \beta_1 + \beta_2 \text{Group}_i + \beta_3 \text{Posttime}_{ij} + \beta_4 \text{Posttime}_{ij} \times \text{Group}_i.$$

This model tests whether the differences between the group means, averaged over the $(n - 1)$ post-baseline measurement occasions, are significantly different from the corresponding differences at baseline.

That is, the hypothesis of no group effect on longitudinal change corresponds to the test of no group by post-baseline interaction.

In general, this test has $(G - 1)$ d.f., where G is the number of groups.

Case Study 1: Vlagtwedde-Vlaardingen Study

Epidemiologic study on prevalence of and risk factors for chronic obstructive lung disease.

Sample participated in follow-up surveys approximately every 3 years for up to 21 years.

Pulmonary function was determined by spirometry: FEV_1.

We focus on a subset of 133 residents aged 36 or older at their entry into the study and whose smoking status did not change over the 19 years of follow-up.

Each study participant was either a current or former smoker.

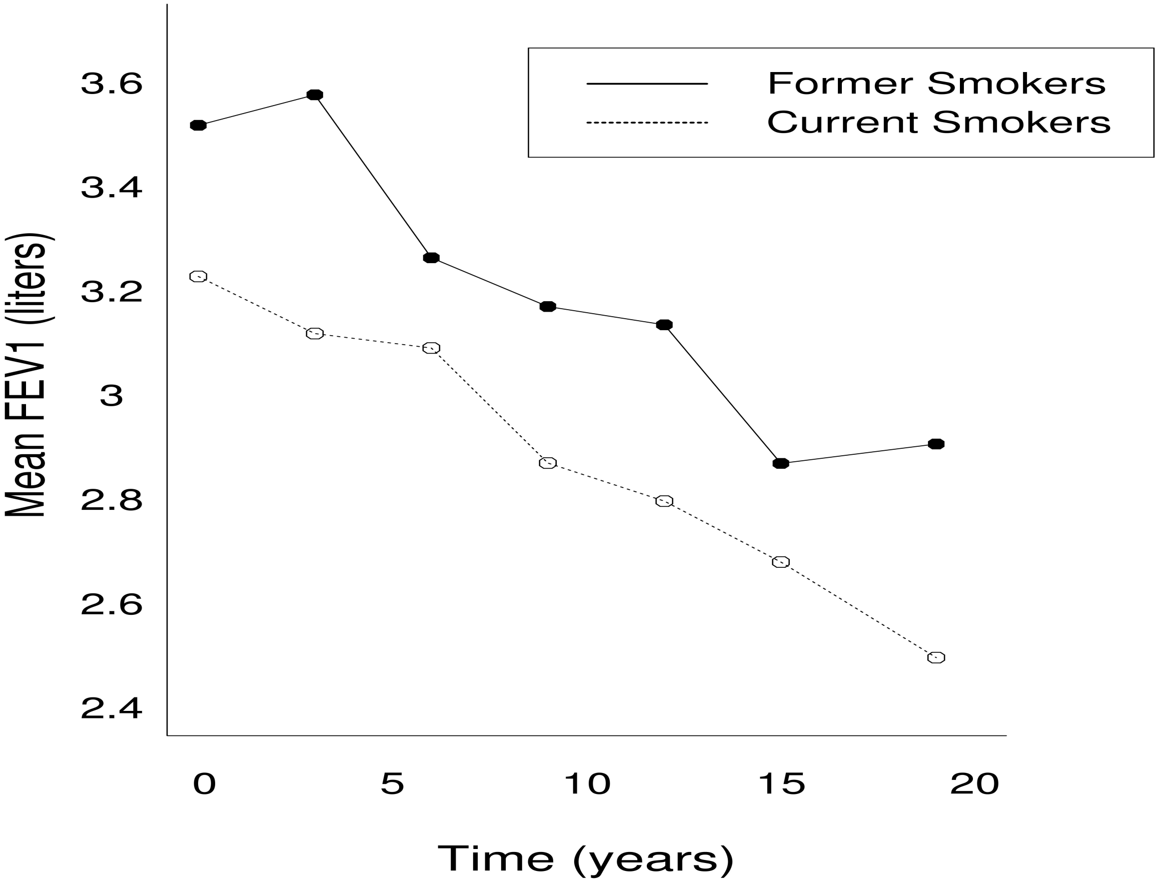


Figure 15: Mean FEV_1 at baseline (year 0), year 3, year 6, year 9, year 12, year 15, and year 19 in the current and former smoking exposure groups.

First we consider a linear trend in the mean response over time, with intercepts and slopes that differ for the two smoking exposure groups.

We assume an unstructured covariance matrix.

Based on the REML estimates of the regression coefficients in Table 23, the mean response for former smokers is

$$E\left(Y_{ij}\right) = 3.507 - 0.033\,\texttt{Time}_{\texttt{ij}},$$

while for current smokers,

$$\begin{aligned} E\left(Y_{ij}\right) &= (3.507 - 0.262) - (0.033 + 0.005)\ \texttt{Time}_{\texttt{ij}} \\ &= 3.245 - 0.038\,\texttt{Time}_{\texttt{ij}}. \end{aligned}$$

Table 23: Estimated regression coefficients for linear trend model for FEV_1 data from the Vlagtwedde-Vlaardingen study.

Variable	Smoking Group	Estimate	SE	Z
Intercept		3.5073	0.1004	34.94
$\texttt{Smoke}_{\texttt{i}}$	Current	−0.2617	0.1151	−2.27
$\texttt{Time}_{\texttt{ij}}$		−0.0332	0.0031	−10.84
$\texttt{Smoke}_{\texttt{i}} \times \texttt{Time}_{\texttt{ij}}$	Current	−0.0050	0.0035	−1.42

Thus, both groups have a significant decline in mean FEV_1 over time.

But there is no discernible difference between the two smoking exposure groups in the constant rate of change.

That is, the $\texttt{Smoke}_\texttt{i} \times \texttt{Time}_\texttt{ij}$ interaction (i.e., the comparison of the two slopes) is not significant, with $Z = -1.42$, $p > 0.15$.

But is the rate of change constant over time?

Adequacy of linear trend model can be assessed by including higher-order polynomial trends.

For example, we can consider a model that allows quadratic trends for changes in FEV_1 over time.

Recall that linear trend model is nested within the quadratic trend model.

The maximized log-likelihoods for the models with linear and quadratic trends are presented in Table 24.

LRT test statistic can be compared to a chi-squared distribution with 2 degrees of freedom (or 6, the number of parameters in the quadratic trend model, minus 4, the number of parameters in the linear trend model).

Note: Likelihood ratio test is based on the ML, not REML, log-likelihood.

Table 24: Maximized (ML) log-likelihoods for models with linear and quadratic trends for FEV_1 data from the Vlagtwedde-Vlaardingen study.

Model	-2 (ML) Log-Likelihood
Quadratic Trend Model	237.2
Linear Trend Model	238.5
$-2 \times$ Log-Likelihood Ratio: $G^2 = 1.3$, 2 df $(p > 0.50)$	

LRT comparing quadratic and linear trend models, produces $G^2 = 1.3$, with 2 degrees of freedom ($p > 0.50$).

Thus, when compared to quadratic trend model, linear trend model appears to be adequate.

Finally, for illustrative purposes, we can make a comparison with a cubic trend model.

This produces LRT statistic, $G^2 = 4.4$, with 4 degrees of freedom ($p > 0.35$), indicating again that the linear trend model is adequate.

Case Study 2: Treatment of Lead-Exposed Children Trial

Recall data from TLC trial:

Children randomized to placebo or Succimer.

Measures of blood lead level at baseline, 1, 4 and 6 weeks.

The sequence of means over time in each group is displayed in Figure 16.

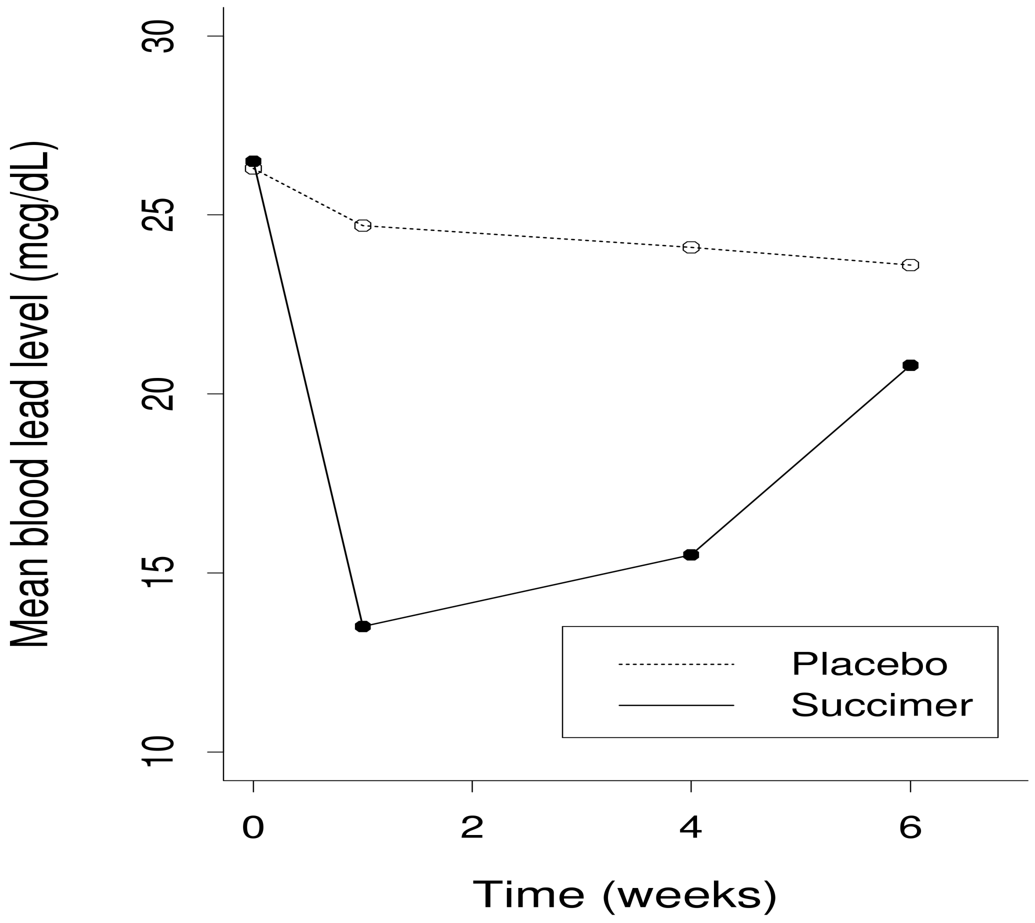


Figure 16: Mean blood lead levels at baseline, week 1, week 4, and week 6 in the succimer and placebo groups.

Given that there are non-linearities in the trends over time, higher-order polynomial models (e.g., a quadratic trend model) could be fit to the data.

Alternatively, we can accommodate the non-linearity with a piecewise linear model with common knot at week 1,

$$\begin{aligned} E\left(Y_{ij}\right) &= \beta_1 + \beta_2\,\mathtt{Week_{ij}} + \beta_3\left(\mathtt{Week_{ij}} - \mathtt{1}\right)_+ + \beta_4\,\mathtt{Group_i} \times \mathtt{Week_{ij}} \\ &\quad + \beta_5\,\mathtt{Group_i} \times \left(\mathtt{Week_{ij}} - \mathtt{1}\right)_+, \end{aligned}$$

where $\mathtt{Group_i} = 1$ if assigned to succimer, and $\mathtt{Group_i} = 0$ otherwise.

Because of randomization, model does not contain a main effect of `Group`.

That is, we assume a common mean blood lead level at baseline.

In this piecewise linear model, means for subjects in placebo group are

$$E\left(Y_{ij}\right) = \beta_1 + \beta_2\,\texttt{Week}_{\texttt{ij}} + \beta_3\left(\texttt{Week}_{\texttt{ij}} - \texttt{1}\right)_+,$$

while in the succimer group

$$E\left(Y_{ij}\right) \quad = \quad \beta_1 + \left(\beta_2 + \beta_4\right)\texttt{Week}_{\texttt{ij}} + \left(\beta_3 + \beta_5\right)\left(\texttt{Week}_{\texttt{ij}} - \texttt{1}\right)_+.$$

Table 25: Estimated regression coefficients and standard errors based on a piecewise linear model, with knot at week 1.

Variable	Group	Estimate	SE	Z
Intercept		26.3422	0.4991	52.78
$\texttt{Week}_{\texttt{ij}}$		−1.6296	0.7818	−2.08
$(\texttt{Week}_{\texttt{ij}} - \texttt{1})_+$		1.4305	0.8777	1.63
$\texttt{Group} \times \texttt{Week}_{\texttt{ij}}$	A	−11.2500	1.0924	−10.30
$\texttt{Group} \times (\texttt{Week}_{\texttt{ij}} - \texttt{1})_+$	A	12.5822	1.2278	10.25

When expressed in terms of mean response prior to/after week 1, estimated means in the placebo group are

$$\widehat{\mu}_{ij} = \widehat{\beta}_1 + \widehat{\beta}_2\,\texttt{Week}_{\texttt{ij}}, \qquad \texttt{Week}_{\texttt{ij}} \leq 1;$$

$$\widehat{\mu}_{ij} = (\widehat{\beta}_1 - \widehat{\beta}_3) + (\widehat{\beta}_2 + \widehat{\beta}_3)\,\texttt{Week}_{\texttt{ij}}, \qquad \texttt{Week}_{\texttt{ij}} > \texttt{1}.$$

Thus, in the placebo group, slope prior to week 1 is $\widehat{\beta}_2 = -1.63$ and following week 1 is $(\widehat{\beta}_2 + \widehat{\beta}_3) = -1.63 + 1.43 = -0.20$.

Similarly, when expressed in terms of the mean response prior to and after week 1, the estimated means for subjects in the succimer group are given by

$$\begin{aligned}\widehat{\mu}_{ij} &= \widehat{\beta}_1 + (\widehat{\beta}_2 + \widehat{\beta}_4)\,\texttt{Week}_{\texttt{ij}}, && \texttt{Week}_{\texttt{ij}} \leq \texttt{1};\\ \widehat{\mu}_{ij} &= \widehat{\beta}_1 - (\widehat{\beta}_3 + \widehat{\beta}_5)\\ &\quad + (\widehat{\beta}_2 + \widehat{\beta}_3 + \widehat{\beta}_4 + \widehat{\beta}_5)\,\texttt{Week}_{\texttt{ij}}, && \texttt{Week}_{\texttt{ij}} > \texttt{1}.\end{aligned}$$

The estimates of the mean blood lead levels for the placebo and succimer groups are presented in Table 26.

The estimated means from the piecewise linear model appear to adequately fit the observed mean response profiles for the two treatment groups.

Note that piecewise linear and quadratic trend models (with common intercept for two groups) are not nested.

They both have the same number of parameters and therefore their respective log-likelihoods can be directly compared.

The maximized log-likelihoods indicate that piecewise linear model fits these data better than quadratic trend model (-2 ML log-likelihood $= 2436.2$ for piecewise linear model versus -2 ML log-likelihood $= 2551.7$ for quadratic trend model).

Table 26: Estimated mean blood lead levels for placebo and succimer groups from linear spline model (knot at week 1). Observed means in parentheses.

Group	Week 0	Week 1	Week 4	Week 6
Succimer	26.3 (26.5)	13.5 (13.5)	16.7 (15.5)	19.1 (20.8)
Placebo	26.3 (26.3)	24.7 (24.7)	24.1 (24.1)	23.7 (23.2)

Parametric Curves using PROC MIXED in SAS

Table 27: Illustrative commands for a linear trend model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time group*time / SOLUTION CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

Note that the CLASS statement includes a variable t. This variable is an additional copy of the variable $time$.

The difference is that while t is declared as a categorical variable on the CLASS statement, $time$ is not and is treated as a quantitative covariate in the MODEL statement.

It is good practice to include, wherever possible, a REPEATED effect.

This ensures covariance is estimated correctly when the design is balanced but incomplete due to missingness or when repeated measures are not in same order for each subject in data set.

Table 28: Illustrative commands for a quadratic trend model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time timesqr group*time group*timesqr /S CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

Table 29: Illustrative commands for a spline model, with knot at $time = 4$, using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group t;
   MODEL y=group time time_4 group*time group*time_4 /S CHISQ;
   REPEATED t / TYPE=UN SUBJECT=id R RCORR;
```

The MODEL statement includes $time$ and $time\_4$, where $time\_4$ is a derived variable for $(time - 4)_+$.

The latter variable can easily be computed in SAS as

$$time\_4 = max(time - 4, 0);$$

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 9

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Modelling the Covariance

Longitudinal data present two aspects of the data that require modelling: mean response over time and covariance.

Although these two aspects of the data can be modelled separately, they are interrelated.

Choice of models for mean response and covariance are interdependent.

A model for the covariance must be chosen on the basis of some assumed model for the mean response.

Recall: Covariance between any pair of residuals, say $[Y_{ij} - \mu_{ij}(\beta)]$ and $[Y_{ik} - \mu_{ik}(\beta)]$, depends on the model for the mean, i.e., depends on β.

Modelling the Covariance

Three broad approaches can be distinguished:

(1) "unstructured" or arbitrary pattern of covariance

(2) covariance pattern models

(3) random effects covariance structure

Unstructured Covariance

Appropriate when design is “balanced” and number of measurement occasions is relatively small.

No explicit structure is assumed other than homogeneity of covariance across different individuals, $\mathrm{Cov}(Y_i) = \Sigma_i = \Sigma$.

Chief advantage: no assumptions made about the patterns of variances and covariances.

With n measurement occasions, "unstructured" covariance matrix has $\frac{n\times(n+1)}{2}$ parameters:

the n variances and $n \times (n-1)/2$ pairwise covariances (or correlations),

$$\mathrm{Cov}(Y_i) = \left(\begin{array}{cccc} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{array} \right).$$

Potential drawbacks:

Number of covariance parameters grows rapidly with the number of measurement occasions:

For $n = 3$ number of covariance parameters is 6

For $n = 5$ number of covariance parameters is 15

For $n = 10$ number of covariance parameters is 55

When number of covariance parameters is large, relative to sample size, estimation is likely to be very unstable.

Use of an unstructured covariance is appealing only when N is large relative to $\frac{n \times (n+1)}{2}$.

Unstructured covariance is problematic when there are mistimed measurements.

Covariance Pattern Models

When attempting to impose some structure on the covariance, a subtle balance needs to be struck.

With too little structure there may be too many parameters to be estimated with limited amount of data.

With too much structure, potential risk of model misspecification and misleading inferences concerning β.

Classic tradeoff between bias and variance.

Covariance pattern models have their basis in models for serial correlation originally developed for time series data.

Compound Symmetry

Assumes variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{i,k}) = \rho$ for all j and k.

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \ldots & 1 \end{pmatrix}.$$

Parsimonious: two parameters regardless of number of measurement occasions.

Strong assumptions about variance and correlation are usually not valid with longitudinal data.

Toeplitz

Assumes variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{i,j+k}) = \rho_k$ for all j and k.

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \dots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & 1 \end{pmatrix}.$$

Assumes correlation among responses at adjacent measurement occasions is constant, ρ_1.

Toeplitz only appropriate when measurements are made at equal (or approximately equal) intervals of time.

Toeplitz covariance has n parameters (1 variance parameter, and $n-1$ correlation parameters).

A special case of the Toeplitz covariance is the (first-order) autoregressive covariance.

Autoregressive

Assumes variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{i,j+k}) = \rho^k$ for all j and k, and $\rho \geq 0$.

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{pmatrix}.$$

Parsimonious: only 2 parameters, regardless of number of measurement occasions.

Only appropriate when the measurements are made at equal (or approximately equal) intervals of time.

Compound symmetry, Toeplitz and autoregressive covariances assume variances are constant across time.

This assumption can be relaxed by considering versions of these models with heterogeneous variances, $\text{Var}(Y_{ij}) = \sigma_j^2$.

A heterogeneous autoregressive covariance pattern:

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \ldots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \ldots & \rho^{n-2}\sigma_2\sigma_n \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \ldots & \rho^{n-3}\sigma_3\sigma_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \ldots & \sigma_n^2 \end{pmatrix},$$

and has $n + 1$ parameters (n variance parameters and 1 correlation parameter).

Banded

Assumes correlation is zero beyond some specified interval.

For example, a banded covariance pattern with a band size of 3 assumes that $\text{Corr}(Y_{ij}, Y_{i,j+k}) = 0$ for $k \geq 3$.

It is possible to apply a banded pattern to any of the covariance pattern models considered so far.

A banded Toeplitz covariance pattern with a band size of 2 is given by,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & 0 & \ldots & 0 \\ \rho_1 & 1 & \rho_1 & \ldots & 0 \\ 0 & \rho_1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix},$$

where $\rho_2 = \rho_3 = \cdots = \rho_{n-1} = 0$.

Banding makes very strong assumption about how quickly the correlation decays to zero with increasing time separation.

Exponential

When measurement occasions are not equally-spaced over time, autoregressive model can be generalized as follows.

Let $\{t_{i1}, ..., t_{in}\}$ denote the observation times for the i^{th} individual and assume that the variance is constant across all measurement occasions, say σ^2, and

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij}-t_{ik}|},$$

for $\rho \geq 0$.

Correlation between any pair of repeated measures decreases exponentially with the time separations between them.

Referred to as “exponential” covariance because it can be re-expressed as

$$\begin{aligned}\mathrm{Cov}(Y_{ij}, Y_{ik}) &= \sigma^2 \rho^{\left|t_{ij}-t_{ik}\right|} \\ &= \sigma^2 \exp\left(-\theta \left|t_{ij} - t_{ik}\right|\right),\end{aligned}$$

where $\theta = -\log(\rho)$ or $\rho = \exp(-\theta)$ for $\theta \geq 0$.

Exponential covariance model is invariant under linear transformation of the time scale.

If we replace t_{ij} by $(a + bt_{ij})$ (e.g., if we replace time measured in “weeks” by time measured in “days”), the same form for the covariance matrix holds.

Choice among Covariance Pattern Models

Choice of models for covariance and mean are interdependent.

Choice of model for covariance should be based on a “**maximal**” model for the mean that minimizes any potential misspecification.

With balanced designs and a very small number of discrete covariates, choose “saturated model” for the mean response.

Saturated model: includes main effects of time (regarded as a within-subject factor) and all other main effects, in addition to their two- and higher-way interactions.

Maximal model should be in a certain sense the most elaborate model for the mean response that we would consider from a subject-matter point of view.

Once maximal model has been chosen, residual variation and covariation can be used to select appropriate model for covariance.

For nested covariance pattern models, a likelihood ratio test statistic can be constructed that compares “full” and “reduced” models.

Recall: two models are said to be nested when the “reduced” model is a special case of the “full” model.

For example, compound symmetry model is nested within the Toeplitz model, since if the former holds the latter must necessarily hold, with $\rho_1 = \rho_2 = \cdots = \rho_{n-1}$.

Likelihood ratio test is obtained by taking twice the difference in the respective maximized REML log-likelihoods,

$$G^2 = 2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}}),$$

and comparing statistic to a chi-squared distribution with df equal to difference between the number of covariance parameters in full and reduced models.

To compare non-nested model, an alternative approach is the Akaike Information Criterion (AIC).

According to the AIC, given a set of competing models for the covariance, one should select the model that minimizes

$$\begin{aligned} \text{AIC} &= -2(\text{maximized log-likelihood}) + 2(\text{number of parameters}) \\ &= -2(\widehat{l} - c), \end{aligned}$$

where $\widehat{l}$ is the maximized REML log-likelihood and c is the number of covariance parameters.

Example: Exercise Therapy Trial

- subjects were assigned to one of two weightlifting programs to increase muscle strength.

- treatment 1: number of repetitions of the exercises was increased as subjects became stronger.

- treatment 2, number of repetitions was held constant but amount of weight was increased as subjects became stronger.

- Measurements of body strength were taken at baseline and on days 2, 4, 6, 8, 10, and 12.

- We focus only on measures of strength obtained at baseline (or day 0) and on days 4, 6, 8, and 12.

Before considering models for the covariance, it is necessary to choose a maximal model for the mean response.

We chose maximal model to be the saturated model for the mean.

First, we consider an unstructured covariance matrix.

Note that the variance appears to be larger by the end of the study when compared to the variance at baseline.

Correlations decrease as the time separation between the repeated measures increases.

Table 30: Estimated unstructured covariance matrix for the strength data at baseline, day 4, day 6, day 8, and day 12.

Day	0	4	6	8	12
0	9.668	10.175	8.974	9.812	9.407
4	10.175	12.550	11.091	12.580	11.928
6	8.974	11.091	10.642	11.686	11.101
8	9.812	12.580	11.686	13.990	13.121
12	9.407	11.928	11.101	13.121	13.944

Table 31: Estimated unstructured correlation matrix for the strength data at baseline, day 4, day 6, day 8, and day 12.

Day	0	4	6	8	12
0	1.0000	0.9237	0.8847	0.8437	0.8102
4	0.9237	1.0000	0.9597	0.9494	0.9017
6	0.8847	0.9597	1.0000	0.9577	0.9113
8	0.8437	0.9494	0.9577	1.0000	0.9394
12	0.8102	0.9017	0.9113	0.9394	1.0000

Despite apparent increase in variance over time, we consider an autoregressive model for the correlation.

Assume variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{i,j+k}) = \rho^k$ for all j and k, and $\rho \geq 0$.

This results in the following estimates of the variance and correlation parameters, $\widehat{\sigma}^2 = 11.87$ and $\widehat{\rho} = 0.94$.

This model was fit primarily for illustrative purposes; the model is not very appropriate as data are unequally spaced over time.

Table 32: Estimated autoregressive correlation matrix for the strength data at baseline, day 4, day 6, day 8, and day 12.

Day	0	4	6	8	12
0	1.0000	0.9402	0.8839	0.8311	0.7813
4	0.9402	1.0000	0.9402	0.8839	0.8311
6	0.8839	0.9402	1.0000	0.9402	0.8839
8	0.8311	0.8839	0.9402	1.0000	0.9402
12	0.7813	0.8311	0.8839	0.9402	1.0000

Instead, consider exponential model for the covariance, where

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{\left|t_{ij} - t_{ik}\right|},$$

for $t_i = (0, 4, 6, 8, 12)$ for all subjects.

Results: $\widehat{\sigma}^2 = 11.87$ and $\widehat{\rho} = 0.98$.

Table 33: Estimated exponential correlation matrix for the strength data at baseline, day 4, day 6, day 8, and day 12.

Day	0	4	6	8	12
0	1.0000	0.9169	0.8780	0.8408	0.7709
4	0.9169	1.0000	0.9576	0.9169	0.8408
6	0.8780	0.9576	1.0000	0.9576	0.8780
8	0.8408	0.9169	0.9576	1.0000	0.9169
12	0.7709	0.8408	0.8780	0.9169	1.0000

There is a hierarchy among the models: autoregressive and exponential are both nested within unstructured.

The autoregressive and exponential models are not nested but have the same number of parameters.

Any comparison between these two models can be made directly in terms of their maximized log-likelihoods.

LRT comparing autoregressive and unstructured covariance,

$$G^2 = 621.1 - 597.3 = 23.8,$$

with 13 (or 15 - 2) degrees of freedom ($p < 0.05$).

There is evidence that the autoregressive model does not provide an adequate fit to the covariance.

LRT comparing exponential and unstructured covariance, yields

$$G^2 = 618.5 - 597.3 = 21.2,$$

and when compared to a chi-squared distribution with 13 degrees of freedom, $p > 0.05$.

Exponential covariance provides an adequate fit to the data.

Also, in terms of AIC, the exponential model minimizes this criterion.

Table 34: Comparison of the maximized (REML) log-likelihoods and AIC for the covariance pattern models for the strength data from the exercise therapy trial.

Covariance Pattern Model	-2 (REML) Log-Likelihood	AIC
Unstructured	597.3	627.3
Autoregressive	621.1	625.1
Exponential	618.5	622.5

Strengths/Weaknesses of Covariance Pattern Models

Covariance pattern models attempt to characterize the covariance with a relatively small number of parameters.

However, many models (e.g., autoregressive, Toeplitz, and banded) appropriate only when repeated measurements are obtained at equal intervals and cannot handle irregularly timed measurements.

While there is a large selection of models for correlations, choice of models for variances is limited.

They are not well-suited for modelling data from inherently unbalanced longitudinal designs.

Table 35: Covariance pattern modelling options using PROC MIXED in SAS.

TYPE =	<pattern>	Specifies the covariance pattern
	UN	Unstructured
	CS	Compound symmetry
	AR(1)	First-order autoregressive
	TOEP	Toeplitz
	UN(n)	Banded unstructured, with n bands
	CSH	Heterogeneous compound symmetry
	ARH(1)	Heterogeneous first-order autoregressive

Table 36: Illustrative commands for an autoregressive model using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group time;
   MODEL y=group time group*time /S CHISQ;
   REPEATED time / TYPE=AR(1) SUBJECT=id R RCORR;
```

Table 37: Illustrative commands for an exponential model using PROC MIXED in SAS.

```
PROC MIXED;
  CLASS id group time;
  MODEL y=group time group*time /S CHISQ;
  REPEATED time / TYPE=SP(EXP)(ctime) SUBJECT=id R RCORR;
```

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 10

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Synthesis of Ideas for Analyzing Longitudinal Data

Primary goal of a longitudinal study is to characterize the change in response over time and the factors that influence change.

Longitudinal data require somewhat more sophisticated statistical techniques because: (i) repeated measures on the same individual are usually positively correlated, and (ii) variability is often heterogeneous across measurement occasions.

Correlation and heterogeneous variability must be accounted for in order to obtain valid inferences about change in response over time.

General Linear Model for Longitudinal Data

So far, we have considered linear regression models that

- permit individuals to be measured on different number of occasions and at different times
- can handle mixed discrete and continuous covariates
- allow a range of different covariance structures

Specifically, 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 Y_{ij} there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, ..., N; \;\; j = 1, ..., n_i.$$

Note: Information about the time of observation, treatment or exposure group, and other predictor and confounding variables can be expressed through this vector of covariates.

We consider linear regression models for changes in the mean response over time:

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

where $\beta_1, ..., \beta_p$ are unknown regression coefficients.

The e_{ij} are random errors, with mean zero, and represent deviations of the Y_{ij}'s from their means,

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

This model can also be represented in vector/matrix notation as:

$$E(Y_i|X_i) = X_i \beta.$$

Assumptions

(1) The individuals represent a random sample from the population of interest.

(2) Observations from different individuals are independent, while repeated measurements of the same individual are not assumed to be independent.

(3) The elements of the vector of repeated measures $Y_{i1}, \ldots, Y_{in_i}$, have a Multivariate Normal (MVN) distribution, with means

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

and covariance matrix Σ_i[3].

(4) If there are missing data they are assumed to be "ignorable", i.e., MAR or MCAR.

[3] Covariance matrix is indexed by i to permit individuals to have different numbers of repeated measures, n_i

Modelling Longitudinal Data

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

(1) mean response over time

(2) covariance among repeated measures

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

Modelling the Mean

Two main approaches can be distinguished:

(1) analysis of response profiles

(2) parametric or semi-parametric curves.

Modelling the Covariance

Three broad approaches can be distinguished:

(1) “unstructured” or arbitrary pattern of covariance

(2) covariance pattern models

(3) random effects covariance structure (discussed in next lecture)

Modelling Longitudinal Data

We have stressed that, in fitting linear models to longitudinal data, we have two modeling tasks:

(a) We must choose a covariance model that provides a good fit to the observed variances and covariances.

(b) We must fit a linear regression model that provides a good fit to the mean of the outcome variable.

Because the models for the mean and covariance are interdependent, we need to have a coherent strategy for model fitting.

Choosing a Covariance Structure

The choices of models for the mean and covariance are interdependent.

Since the residuals depend on the specification of the linear model for the mean, we choose a covariance structure for a particular linear model.

Substantial changes in the linear model could lead to a different choice of model for the covariance.

A balance needs to be struck:

With too little structure (e.g., unstructured), there may be too many parameters to be estimated with the limited amount of data available. This would leave too little information available for estimating β

$\Rightarrow$ weaker inferences concerning β.

With too much structure (e.g., compound symmetry), there is more information available for estimating β.
However, there is a potential risk of model misspecification

$\Rightarrow$ apparently stronger, but potentially biased, inferences concerning β.

General Strategy for Model Fitting

(1) To analyze longitudinal data we first need to choose a "working" covariance structure.

We must recognize that choices of model for the mean and covariance are interdependent.

Need to fit a "maximal model" for the mean response when choosing/comparing models for the covariance.

Can use REML log likelihood or AIC as criteria to guide the choice of model for the covariance.

When n is relatively small, and design is balanced, can simply use unstructured covariance matrix unless simpler model is clearly satisfactory.

When n is relatively large and/or there are mistimed measurements, alternative models for the covariance will need to be considered.

(2) Given choice of “working” covariance, select model for mean response. Need to decide how to model the pattern of change in the mean response:

(a) covariate by time interaction(s), where time is regarded as a categorical variable (analysis of response profiles)
(b) covariate by time interaction(s), where means are modeled as an explicit function of continuous time (parametric and semi-parametric curves)
(c) covariate effects in an analysis that includes the baseline measure as a covariate (e.g., randomized study)
(d) covariate by post-baseline time (posttime) interaction(s) in a “constant effect” model

Use the ML log likelihood to compare nested models for the mean differing by several degrees of freedom.

(3) Make an initial determination of the final form of the regression model.

(4) If necessary, re-fit the final regression model using REML to obtain standard errors.

Empirical Variance Estimation

We have focused on regression models for longitudinal data where the primary interest is in making inference about the regression parameters β.

For statistical inference about β we need

(i) an estimate, $\widehat{\beta}$

(ii) estimated standard error, $\text{SE}(\widehat{\beta})$

So far, we have made inferences about β using standard errors obtained under an assumed model for the covariance structure.

This approach is potentially problematic if the assumed covariance has been mis-specified.

How might the covariance be mis-specified?

For example, compound symmetry might be assumed but the correlations in fact decline over time.

Alternatively, an unstructured covariance might be assumed but the covariances also depend upon the treatment group.

If the assumed covariance has been mis-specified, we can correct the standard errors by using "empirical" or so-called "robust" variances.

Recall, the REML estimator of β is given by

$$\widehat{\beta} = \left[\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}Y_i\right)$$

where $\widehat{\Sigma}$ is the REML estimate of Σ.

It has covariance matrix,

$$\text{Cov}(\widehat{\beta}) =$$
$$\left[\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}\text{Cov}\left(Y_i\right)\widehat{\Sigma}^{-1}X_i\right)\left[\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}$$

If $\text{Cov}(Y_i)$ is replaced by $\widehat{\Sigma}$, the REML estimate of Σ, $\text{Cov}(\widehat{\beta})$ can be estimated by

$$\left[\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}$$

However, if the covariance has been mis-specified then an alternative estimator for $\text{Cov}\left(Y_i\right)$ is needed.

The empirical or so-called robust variance of $\widehat{\beta}$ is obtained by using

$$\widehat{V}_i = \left(Y_i - X_i\widehat{\beta}\right)\left(Y_i - X_i\widehat{\beta}\right)'$$

as an estimate of $\text{Cov}\left(Y_i\right)$.

Thus, the empirical variance of $\widehat{\beta}$ is estimated by

$$\left[\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}\widehat{V}_i\widehat{\Sigma}^{-1}X_i\right)\left[\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}$$

This empirical variance estimator is also known as the “sandwich estimator”.

The remarkable thing about the empirical estimator of $\text{Cov}(\widehat{\beta})$ is that it provides a consistent estimator of the variance even when the model for the covariance matrix has been misspecified.

That is, in large samples the empirical variance estimator yields correct standard errors.

In general, its use should be confined to cases where N (number of individuals) is relatively large and n (number of measurements) is relatively small.

The empirical variance estimator may not be appropriate when there is severe imbalance in the data.

In summary, (with large samples) the following procedure will produce valid estimates of the regression coefficients and their standard errors:

(1) Choose a "working" covariance matrix of some convenient form.
(2) Estimate the regression coefficients under the assumed working covariance matrix.
(3) Estimate the standard errors using the empirical variance estimator.

Why not be a clever ostrich?

Why not simply ignore potential correlation among repeated measures (i.e., put head in sand) and assume an independence "working" covariance. Then, obtain correct standard errors using empirical variance estimator.

Why should we bother to explicitly model the covariance?

Reasons:

(1) Efficiency: The optimal (most precise) estimator of β uses the true $\text{Cov}\,(Y_i)$. Given sufficient data, we can attempt to estimate $\text{Cov}\,(Y_i)$.

(2) When N (number of individuals) is not large relative to n (number of measurements) the empirical variance estimator is not recommended.

(3) Missing values: The empirical variance estimator uses the replications across individuals to estimate the covariance structure. This becomes problematic when there are missing data or when the times of measurement are not common.

In general, it is advantageous to model the covariance.

Table 38: Illustrative commands for an exponential model, with empirical standard errors, using PROC MIXED in SAS.

```
PROC MIXED EMPIRICAL;
  CLASS id group time;
  MODEL y=group time group*time /S CHISQ;
  REPEATED time / TYPE=SP(EXP)(ctime) SUBJECT=id R RCORR;
```

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 11

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Linear Mixed Effects Models for Longitudinal Data

Motivating Example: ***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.

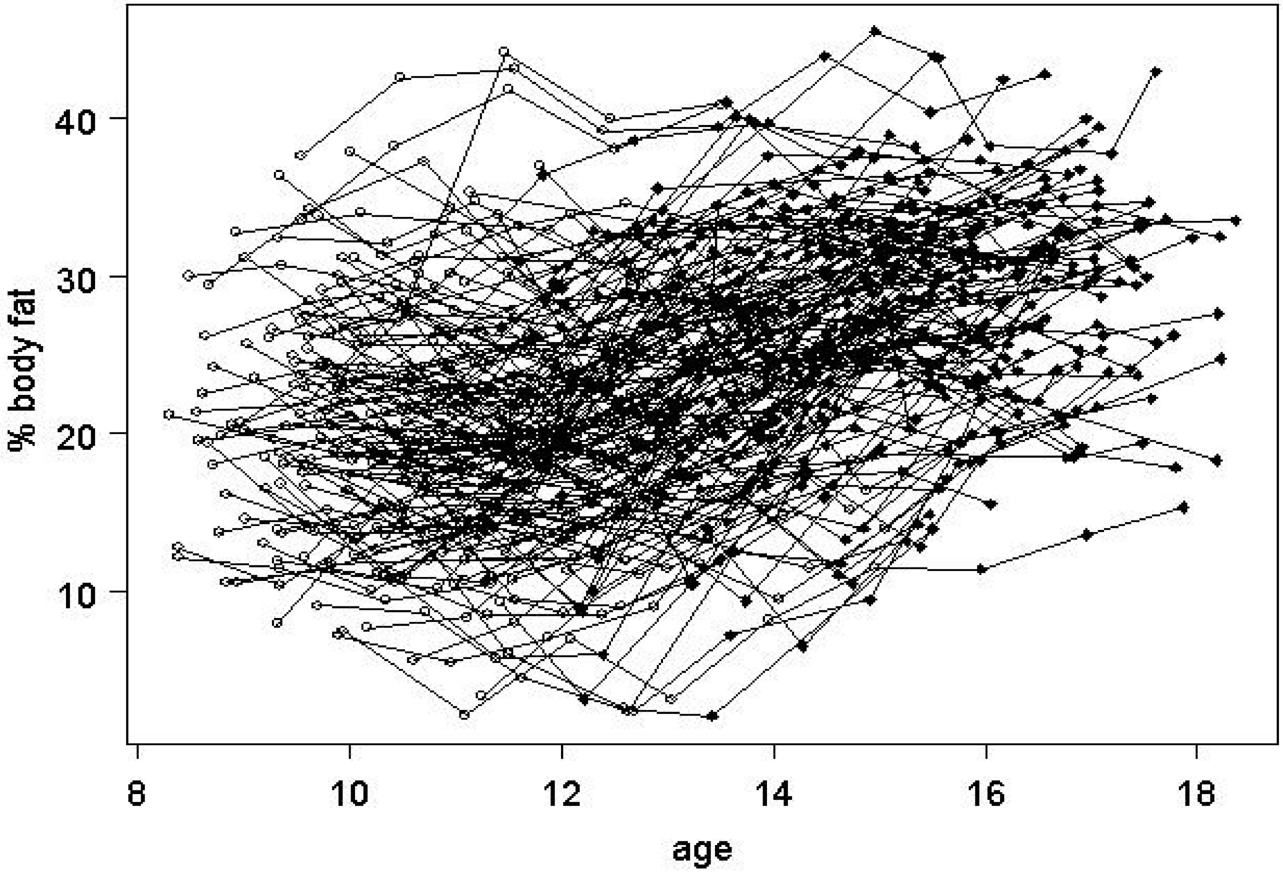


Figure 17: Timeplot of percent body fat against age (in years).

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.

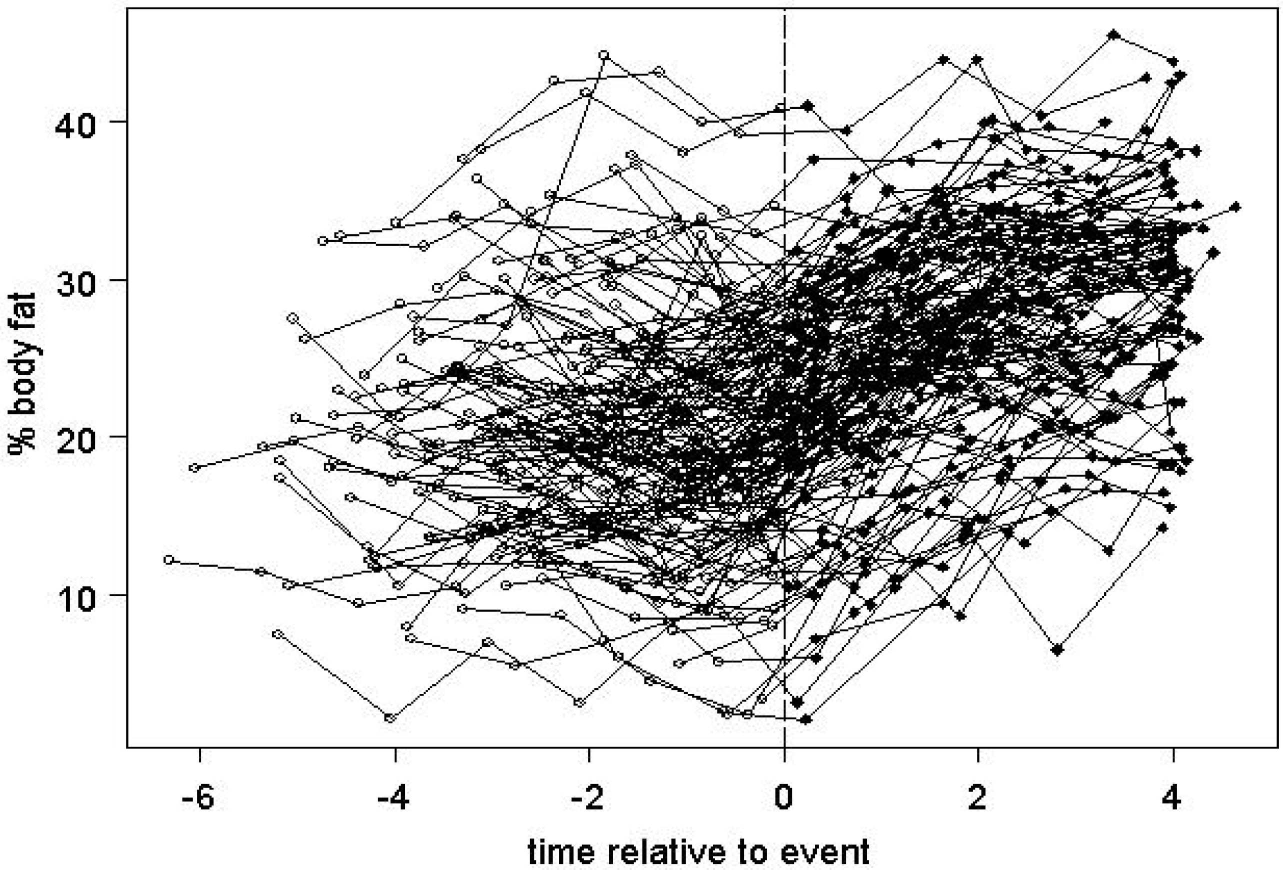


Figure 18: Timeplot of percent body fat against time, relative to age of menarche (in years).

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.

Linear Models for the Mean Response

The mean response can be modelled by a familiar regression model.

For example, with a linear trend over time, we may have

$$E(Y_{ij}) = \mu_{ij} = \beta_1 + \beta_2 t_{ij}.$$

With additional covariates, this can be written more generally

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

where t_{ij}, or possibly functions of t_{ij}, have been incorporated into the covariates, e.g., $X_{ij1} = 1$, $X_{ij2} = t_{ij}$, $X_{ij3} =$ treatment group indicator, and $X_{ij4} = t_{ij} \times$ treatment group indicator.

Model for Covariance: Random Intercept Model

One traditional approach for handling the covariance among repeated measures is to assume that it arises from a random subject effect.

That is, each subject is assumed to have an (unobserved) underlying level of response which persists across all of his/her repeated measurements.

This subject effect is treated as random and the model becomes

$$\begin{aligned} Y_{ij} &= \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij} \\ &= \beta_1 + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + b_i + \epsilon_{ij}, \end{aligned}$$

assuming $X_{ij1} = 1$ for all i and j, or

$$Y_{ij} = (\beta_1 + b_i) + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + \epsilon_{ij}$$

(also known as "random intercept model").

In the model

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

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

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

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

Furthermore, it is assumed that

$$b_i \sim N(0, \sigma_b^2); \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

and that b_i and ϵ_{ij} are mutually independent.
Note: Assumption of normality not always necessary.

Figure 19 provides graphical representation of linear trend model:

$$Y_{ij} = (\beta_1 + b_i) + \beta_2 t_{ij} + \epsilon_{ij}$$

Overall mean response over time in the (sub)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 (sub)population trend (denoted by the broken lines).

Individual A responds “higher” than the (sub)population average and thus has a positive b_i.

Individual B responds “lower” than the (sub)population average and has a negative b_i.

Inclusion of measurement errors, ϵ_{ij}, allows response at any occasion to vary randomly above/below subject-specific trajectories (see Figure 20).

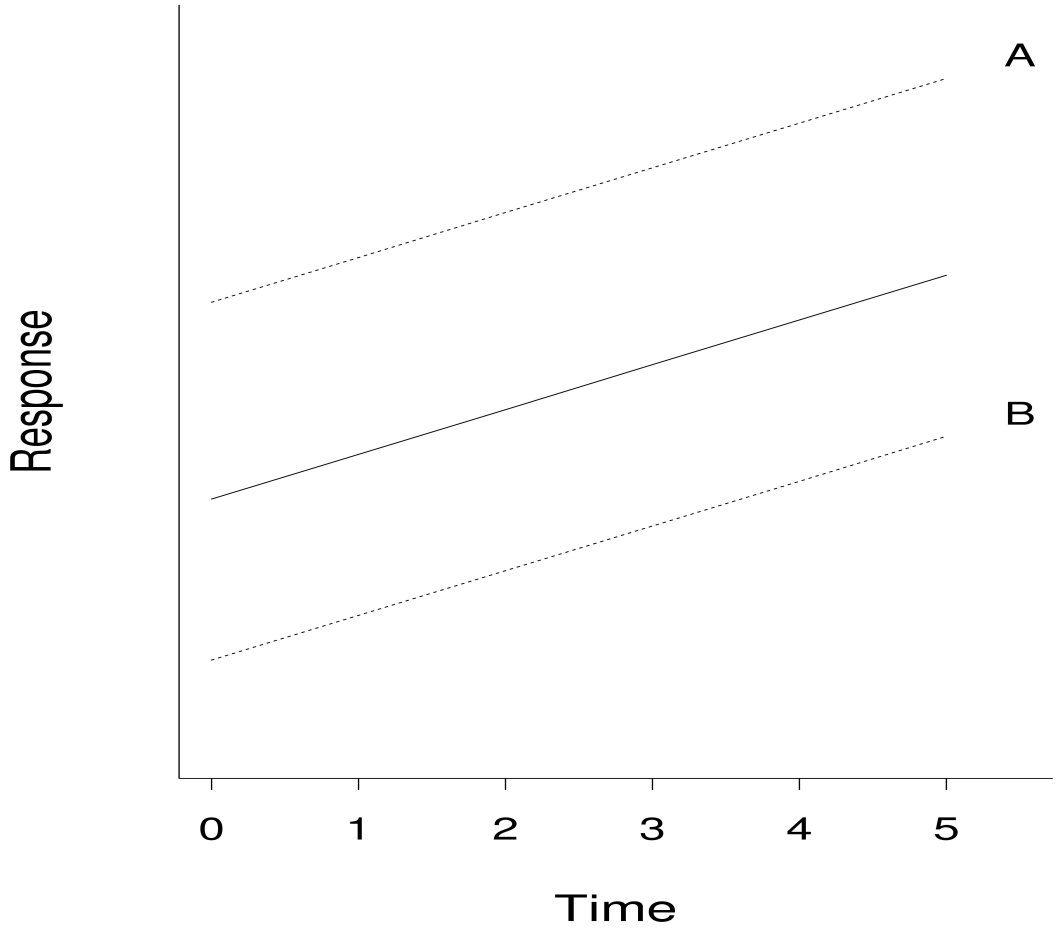


Figure 19: Graphical representation of the overall and subject-specific mean responses over time.

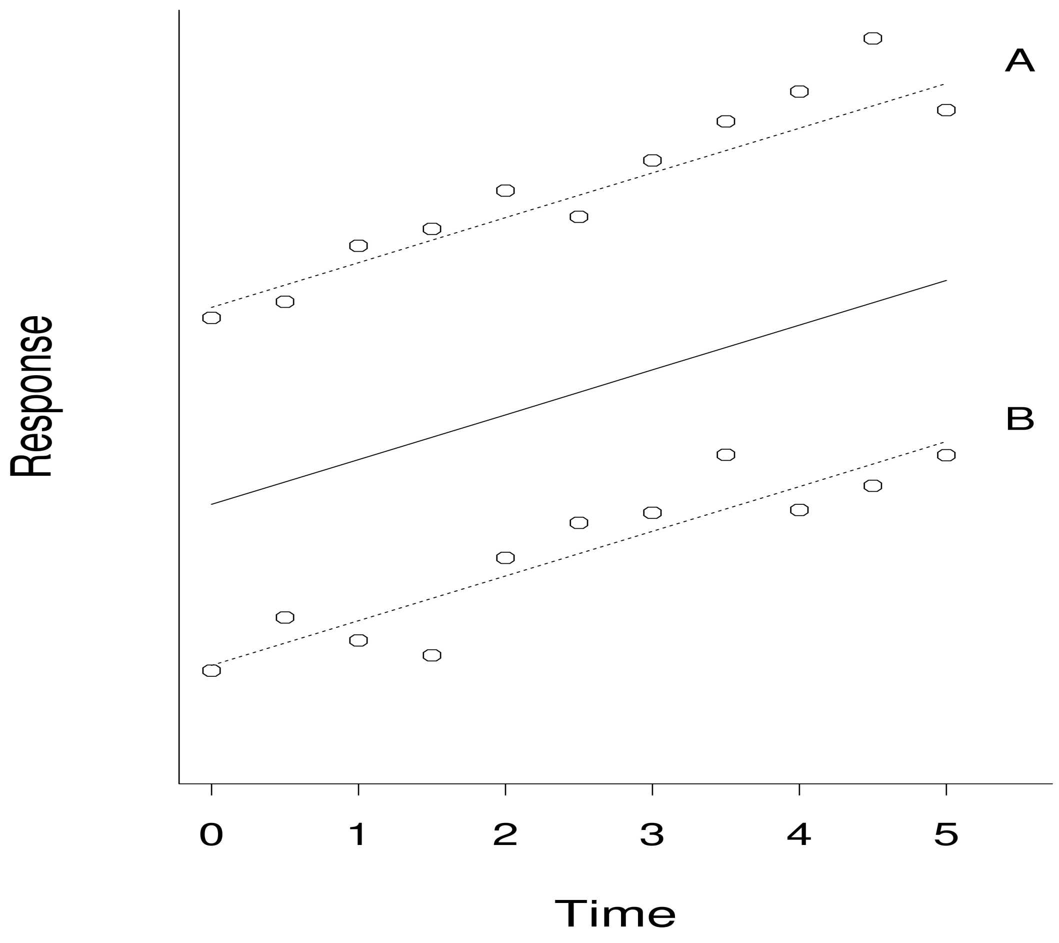


Figure 20: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

Covariance/Correlation Structure

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

If $\mathrm{Var}\,(b_i) = \sigma_b^2$ and $\mathrm{Var}\,(\epsilon_{ij}) = \sigma^2$, the covariance matrix of the repeated measurements has the compound symmetry form:

$$\begin{bmatrix} \sigma_b^2+\sigma^2 & \sigma_b^2 & \sigma_b^2 & \ldots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma^2 & \sigma_b^2 & \ldots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma^2 & \ldots & \sigma_b^2 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \ldots & \sigma_b^2+\sigma^2 \end{bmatrix}$$

$$\text{Var}(Y_{ij}) = \sigma_b^2 + \sigma^2$$

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

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

Note: The introduction of a random subject effect, b_i, induces correlation among the repeated measurements.

The compound symmetry model is the simplest possible example of a mixed effect model.

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

Solution: Allow for heterogeneity is 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,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}, \;\; j = 1, ..., n_i,$$

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 21).

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

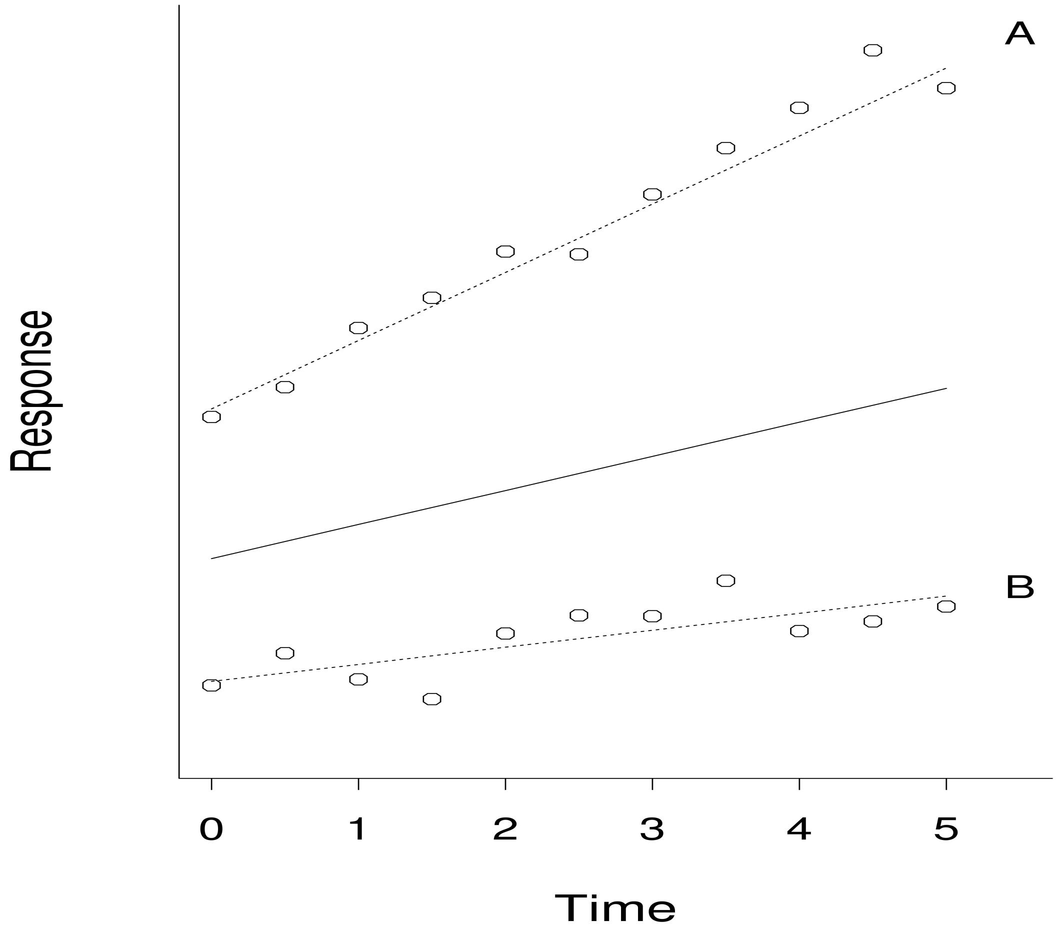


Figure 21: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

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

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 \mathrm{trt}_i + \beta_4 t_{ij} \times \mathrm{trt}_i + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij},$$

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

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

trt = 0: $Y_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i}) t_{ij} + \epsilon_{ij},$

trt = 1: $Y_{ij} = (\beta_1 + \beta_3 + b_{1i}) + (\beta_2 + \beta_4 + b_{2i}) t_{ij} + \epsilon_{ij},$

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

Assuming $b_{1i} \sim N(0, \sigma^2_{b_1})$, $b_{2i} \sim N(0, \sigma^2_{b_2})$ (with $\text{Cov}(\text{b}_{1i}, b_{2i}) = \sigma_{b_1,b_2}$) and $\epsilon_{ij} \sim N(0, \sigma^2)$, then

$$\begin{aligned}
\text{Var}\,(Y_{ij}) &= \text{Var}\,(b_{1i} + b_{2i}t_{ij} + \epsilon_{ij}) \\
&= \text{Var}(b_{1i}) + 2t_{ij}\text{Cov}(b_{1i}, b_{2i}) + t^2_{ij}\text{Var}(b_{2i}) + \text{Var}(\epsilon_{ij}) \\
&= \sigma^2_{b_1} + 2t_{ij}\sigma_{b_1,b_2} + t^2_{ij}\sigma^2_{b_2} + \sigma^2.
\end{aligned}$$

Similarly, it can be shown that

$$\text{Cov}\,(Y_{ij}, Y_{ik}) = \sigma^2_{b_1} + (t_{ij} + t_{ik})\,\sigma_{b_1,b_2} + t_{ij}t_{ik}\sigma^2_{b_2}.$$

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 covariate in X_{ij}.

For example, consider the random intercepts and slopes model introduced earlier,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 \text{trt}_i + \beta_4 t_{ij} \times \text{trt}_i + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}.$$

In this model, $X_{ij} = [1 \;\; t_{ij} \;\; \text{trt}_i \;\; t_{ij} * \text{trt}_{ij}]$ and $Z_{ij} = [1 \;\; t_{ij}]$.

In general, 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,

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

For example, in the random intercepts and slopes model,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 \text{trt}_i + \beta_4 t_{ij} \times \text{trt}_i + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij},$$

G is a 2×2 matrix with unique components $g_{11} = \text{Var}(b_{1i})$,
$g_{12} = \text{Cov}(b_{1i}, b_{2i})$, and $g_{22} = \text{Var}(b_{2i})$.

The within-subject errors, ϵ_{ij}, are assumed to have a multivariate normal distribution with mean zero and covariance matrix denoted by R_i,

$$\epsilon_{ij} \sim N(0, R_i).$$

Note: Usually, it is assumed that $R_i = \sigma^2 I$, where I is a $(n_i \times n_i)$ identity matrix.

That is, when $R_i = \sigma^2 I$, the errors ϵ_{ij} within a subject are uncorrelated, with homogeneous variance.

$\Rightarrow$ "conditional independence assumption".

In principle, a structured model for R_i could be assumed, e.g., AR(1).

Conditional and Marginal Means

In the linear mixed effects model,

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

there is an important distinction between the conditional mean,

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

and the marginal mean,

$$E(Y_{ij}|X_{ij}) = X'_{ij}\beta.$$

The former describes the mean response for an individual, the latter describes the mean response averaged over individuals.

The distinction between the conditional and marginal means is best understood with a simple example.

Consider the simple random intercepts and slopes model,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij},$$

In this model, we can distinguish the conditional mean for an individual,

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

(see broken lines for subjects A and B in Figure 22), and the marginal mean averaged over individuals,

$$E(Y_{ij}) = \beta_1 + \beta_2 t_{ij},$$

(see solid line in Figure 22).

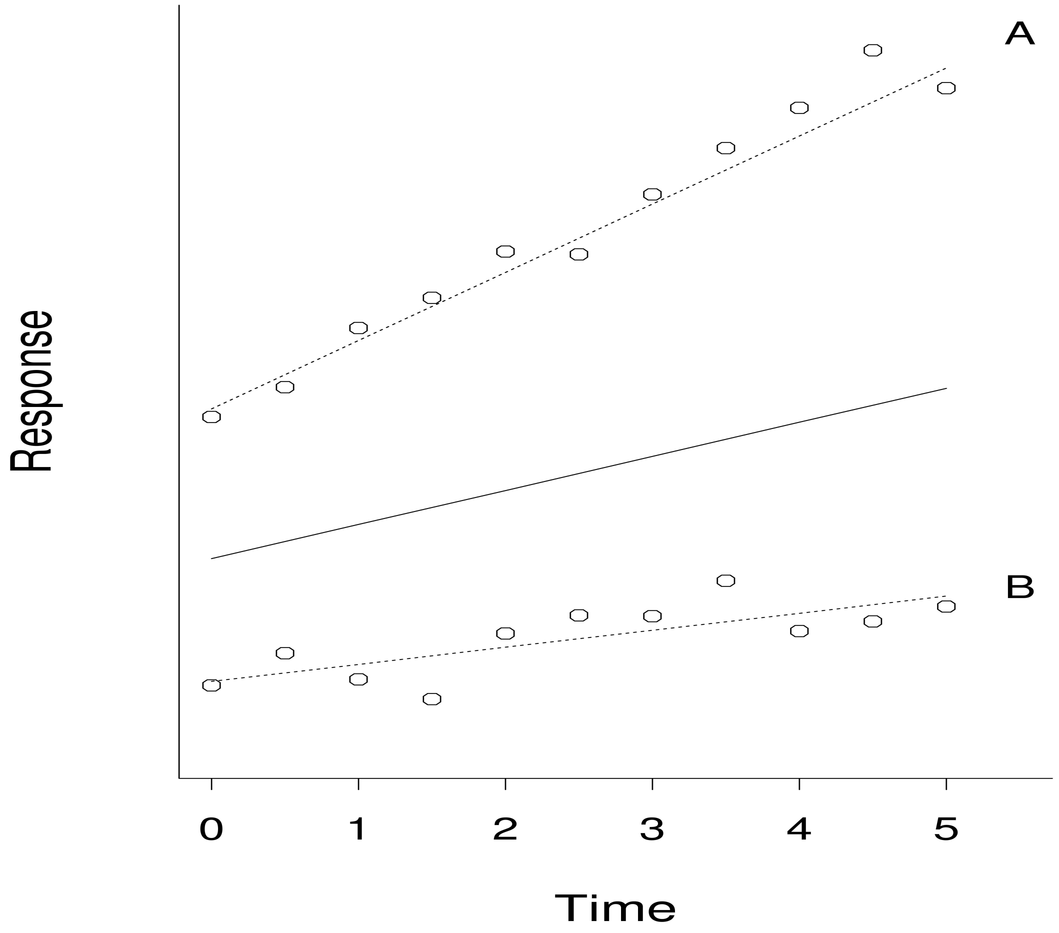


Figure 22: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

Conditional and Marginal Covariance

Variation and covariation can also be defined relative to the conditional and marginal means.

In the linear mixed effects model,

$$Y_{ij} = X_{ij}'\beta + Z_{ij}'b_i + \epsilon_{ij},$$

the conditional variance, $\text{Var}(Y_{ij}|X_{ij}, b_i) = \text{Var}(\epsilon_{ij}) = \sigma^2$ (when $R_i = \sigma^2 I$).

In contrast, the marginal covariance of the vector of responses Y_i is

$$\text{Cov}(Y_i|X_i) = Z_i G Z_i' + R_i = Z_i G Z_i' + \sigma^2 I.$$

Note: This matrix has non-zero off-diagonal elements (i.e., introduction of random effects, b_i, induces correlation marginally among the Y_i).

The distinction between conditional and marginal (co)variances is best understood by considering the simple random intercepts and slopes model,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}.$$

The conditional variance, $\text{Var}(Y_{ij}|b_{1i}, b_{2i}) = \text{Var}(\epsilon_{ij}) = \sigma^2$, describes variation in an individual's observations around her subject-specific mean (i.e., variation of observations around the broken line in Figure 23).

The marginal covariance describes (co)variation of the observations with respect to the marginal mean (i.e., variation and covariation of observations around the solid line in Figure 23):

$$\begin{aligned} \text{Var}\left(Y_{ij}\right) &= \sigma^2_{b_1} + 2t_{ij}\sigma_{b_1,b_2} + t^2_{ij}\sigma^2_{b_2} + \sigma^2. \\ \text{Cov}\left(Y_{ij}, Y_{ik}\right) &= \sigma^2_{b_1} + \left(t_{ij} + t_{ik}\right)\sigma_{b_1,b_2} + t_{ij}t_{ik}\sigma^2_{b_2}. \end{aligned}$$

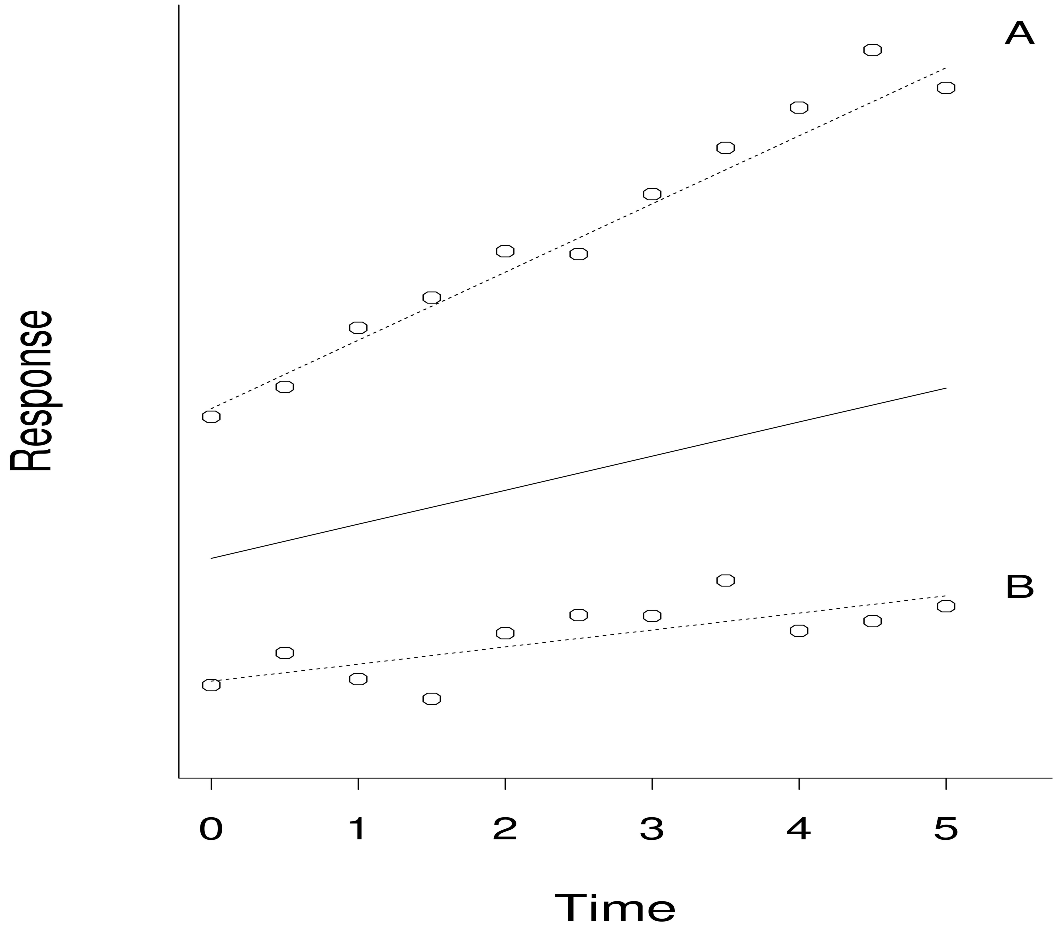


Figure 23: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

Estimation: Maximum Likelihood

ML estimator of $\beta_1, \beta_2, ..., \beta_p$ is the *generalized least squares* (GLS) estimator and depends on marginal covariance among the repeated measures (see Lecture 5).

In general, there is no simple expression for ML estimator of the covariance components - G and σ^2 (or R) - requires iterative techniques.

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

Example: Exercise Therapy Trial

- subjects were assigned to one of two weightlifting programs to increase muscle strength.

- treatment 1: number of repetitions of the exercises was increased as subjects became stronger.

- treatment 2, number of repetitions was held constant but amount of weight was increased as subjects became stronger.

- Measurements of body strength were taken at baseline and on days 2, 4, 6, 8, 10, and 12.

- We focus only on measures of strength obtained at baseline (or day 0) and on days 4, 6, 8, and 12.

Example: Exercise Therapy Trial

Consider a model with intercepts and slopes that vary randomly among subjects, and which allows the mean values of the intercept and slope to differ in the two treatment groups.

To fit this model, use the following code:

```
PROC MIXED DATA = stren;
     CLASS id trt;
     MODEL y=trt time time*trt / S CHISQ;
     RANDOM INTERCEPT time / TYPE=UN SUBJECT=ID G;
```

Selected Output from PROC MIXED

```
          Estimated G Matrix

Effect        id          col1         col2

Intercept      1        9.5469      0.05331
time           1        0.0533      0.02665


Residual: 0.6862


               Fit Statistics

-2 Res Log Likelihood          632.0
AIC (smaller is better)        640.0
AICC (smaller is better)       640.2
BIC (smaller is better)        646.4
```

```
                     Solution for Fixed Effects

                                    Standard
Effect        trt    Estimate          Error    DF    t Value    Pr > |t|

Intercept             81.2396         0.6910    35     117.57     <.0001
trt             1     -1.2349         1.0500    99      -1.18     0.2424
trt             2           0              .     .          .          .
time                   0.1729         0.0427    35       4.05     0.0003
time*trt        1     -0.0377         0.0637    99      -0.59     0.5548
time*trt        2           0              .     .          .          .
```

Recall:

$$\begin{aligned}\mathrm{Cov}\,(Y_i) &= Z_i \mathbf{G} Z_i' + R_i \\ &= Z_i \mathbf{G} Z_i' + \sigma^2 I\end{aligned}$$

Given estimates of G:

$$\begin{bmatrix} 9.54695 & 0.05331 \\ 0.05331 & 0.02665 \end{bmatrix}$$

and of $R_i = \sigma^2 I = (0.6862)I$,

and with

$$Z_i = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 12 \end{bmatrix}$$

We can obtain the following estimate of $\text{Cov}\,(Y_i)$:

$$\begin{bmatrix} 10.23 & 9.76 & 9.87 & 9.97 & 10.19 \\ 9.76 & 11.09 & 10.72 & 11.04 & 11.68 \\ 9.87 & 10.72 & 11.83 & 11.57 & 12.43 \\ 9.97 & 11.04 & 11.57 & 12.79 & 13.17 \\ 10.19 & 11.68 & 12.43 & 13.17 & 15.35 \end{bmatrix}$$

The corresponding correlation matrix is:

$$\begin{bmatrix} 1.000 & 0.916 & 0.897 & 0.872 & 0.813 \\ 0.916 & 1.000 & 0.936 & 0.927 & 0.895 \\ 0.897 & 0.936 & 1.000 & 0.941 & 0.922 \\ 0.872 & 0.927 & 0.941 & 1.000 & 0.940 \\ 0.813 & 0.895 & 0.922 & 0.940 & 1.000 \end{bmatrix}$$

These can be obtained using the following options in PROC MIXED:
RANDOM INTERCEPT time / TYPE=UN SUBJECT=id G V VCORR;

Based on the estimates of the fixed effects:

- the constant rate of increase in strength in group 1 is 0.173 per day
- the constant rate of increase in strength in group 2 is 0.135 (0.173 – 0.038) per day
- the difference between these two rates, -0.038 (SE = 0.064) is not statistically significant

There does not appear to be differences between the two groups in their pattern of increase in strength.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 12

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Two-Stage Random Effects Formulation

Recall main ideas underlying 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.

Linear Mixed Effects Model

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

$$Y_{ij} = X_{ij}'\beta + Z_{ij}'b_i + \epsilon_{ij}, \quad (j = 1, ..., n_i)$$

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}.

The linear mixed model can be motivated by a two-stage formulation.

To help fix ideas, consider a simple example from an animal study designed to compare clearance of iron particles from the lung and liver.

Example

Feldman (1988) describes a study in which iron oxide particles were administered to four rats by intravenous injection and to four other rats by tracheal installation.

The injected particles were taken up by liver endothelial cells and the installed particles by lung macrophages.

Each rat was followed for 30 days, during which time the quantity of iron oxide remaining in the lung was measured by magnetometry.

The iron oxide content declined linearly on the logarithmic scale.

The goal of the study was to compare the rate of particle clearance by liver endothelial cells and by lung macrophages.

Measurements during follow-up were expressed as a percentage of the baseline value, with the baseline value constrained to equal 100%.

Thus, in the analysis we will want to drop the baseline value.

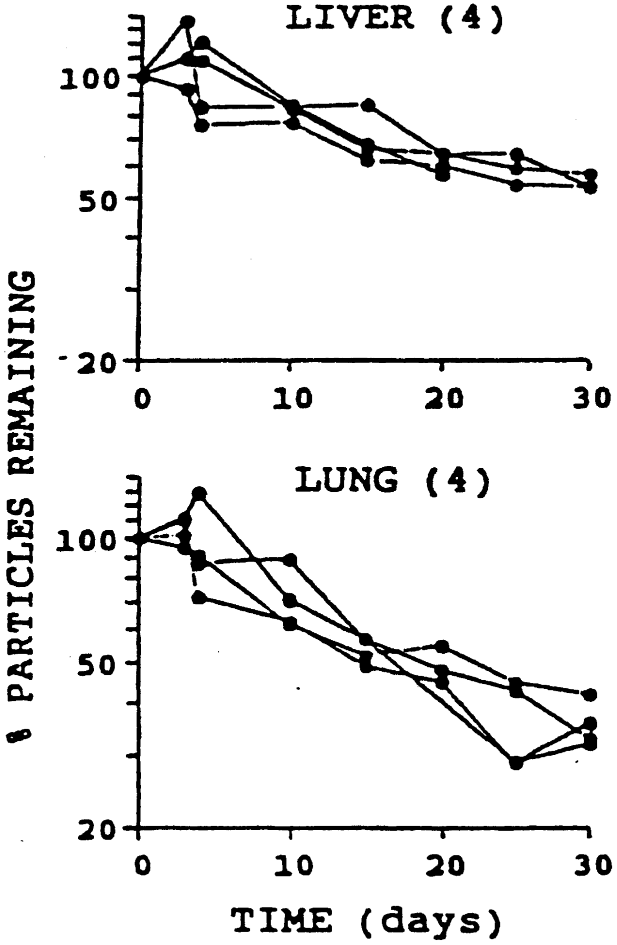


Figure 24: Timeplot of Feldman's clearance data.

Two-Stage Formulation

Linear mixed effects models can be motivated in terms of the following two-stage formulation of the model.

Basic idea: In the two-stage formulation of the model, we assume

1. A straight line (or curve) fits the observed responses for each subject (**first stage**)

2. A regression model relating the mean of the individual intercepts and slopes to subject-specific covariates (**second stage**)

Stage 1

In the first stage subjects are assumed to have their own unique individual-specific mean response trajectories,

$$Y_{ij} = Z'_{ij}\beta_i + \epsilon_{ij}, \quad (j = 1, ..., n_i)$$

where β_i is a vector of subject-specific regression parameters; the errors, ϵ_{ij}, are assumed to be independent within a subject.

For example, a simple model for subject-specific intercepts and slopes is given by,

$$Y_{ij} = \beta_{1i} + \beta_{2i}t_{ij} + \epsilon_{ij}.$$

Thus, in stage 1 we posit a regression model with separate or distinct coefficients for each individual.

This is equivalent to considering separate linear regression models for the data for each individual.

Note: Covariates in Z_{ij} are restricted to only within-individual or time-varying covariates (with the exception of the column of 1's for the intercept).

Time-invariant or between-individual covariates (e.g., gender, treatment group, exposure group) cannot be included in Z_{ij}; instead, they are introduced in the second stage of the model formulation.

Stage 2

In the second stage, we assume individual-specific effects, β_i, are random.

The mean and covariance of the β_i are the population parameters that are modelled in the second stage.

Specifically, the subject-specific coefficients are regressed on other between-subject covariates (e.g., gender, treatment group), say A_i,:

$$E(\beta_i) = A_i\beta.$$

The remaining covariation in the β_i that cannot be explained by A_i is expressed as

$$\mathrm{Cov}(\beta_i) = G.$$

Alternatively, model can be written as

$$\beta_i = A_i \beta + b_i,$$

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

For example, consider two-group setting and the simple model with subject-specific intercepts and slopes.

Allowing both the mean intercept and slope to depend on group

$$\begin{aligned} E(\beta_{1i}) &= \beta_1 + \beta_2 \, \texttt{Group}_\texttt{i} \\ E(\beta_{2i}) &= \beta_3 + \beta_4 \, \texttt{Group}_\texttt{i} \end{aligned}$$

where $\texttt{Group}_\texttt{i} = \texttt{1}$ if the i^{th} individual was assigned to the treatment, and $\texttt{Group}_\texttt{i} = \texttt{0}$ otherwise.

In this model, β_1 is the mean intercept in the control group, while $\beta_1 + \beta_2$ is the mean intercept in the treatment group.

Similarly, β_3 is the mean slope in the control group, while $\beta_3 + \beta_4$ is the mean slope in the treatment group.

In this model, the design matrix A_i of between-individual covariates has the following form:

$$A_i = \begin{pmatrix} 1 & \mathtt{Group_i} & 0 & 0 \\ \\ 0 & 0 & 1 & \mathtt{Group_i} \end{pmatrix}.$$

Thus, for the control group, the model for the mean is

$$E\begin{pmatrix}\beta_{1i}\\ \beta_{2i}\end{pmatrix}=\begin{pmatrix}1&0&0&0\\ 0&0&1&0\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}=\begin{pmatrix}\beta_1\\ \beta_3\end{pmatrix};$$

similarly, for the treatment group, the model for the mean is

$$E\begin{pmatrix}\beta_{1i}\\ \beta_{2i}\end{pmatrix}=\begin{pmatrix}1&1&0&0\\ 0&0&1&1\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}=\begin{pmatrix}\beta_1+\beta_2\\ \beta_3+\beta_4\end{pmatrix}.$$

It is also assumed that there is residual variation in β_i, that cannot be explained by the effect of group,

$$\beta_i = A_i\beta + b_i;$$

this is given by

$$\text{Cov}(\beta_i|A_i) = \text{Cov}(b_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = \text{Var}(b_{1i})$, $g_{22} = \text{Var}(b_{2i})$, and $g_{12} = g_{21} = \text{Cov}(b_{1i}, b_{2i})$.

Thus, g_{11} is the variance of β_{1i}, after adjusting for the effect of treatment group, and so on.

Finally, by combining the two components of the two-stage model, we obtain

$$\begin{aligned} Y_{ij} &= Z'_{ij}\beta_i + \epsilon_{ij} \\ &= Z'_{ij}(A_i\beta + b_i) + \epsilon_{ij} \\ &= (Z'_{ij}A_i)\beta + Z'_{ij}b_i + \epsilon_{ij} \\ &= X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij}, \end{aligned}$$

where $X'_{ij} = Z'_{ij}A_i$.

$\Longrightarrow$ Linear Mixed Effects Model (albeit with constraint, $X'_{ij} = Z'_{ij}A_i$).

Two-Stage Analysis: "NIH Method"

One classic approach with a long history for the analysis of longitudinal data is known as two-stage or two-step analysis.

It is sometimes called the "*NIH Method*" because it was popularized by statisticians working at NIH.

In two-stage method, we fit a straight line (or curve) to the response data for each subject (stage 1), and then regress the estimates of the individual intercepts and slopes on subject-specific covariates (stage 2).

One of the attractions of this method is that it is very easy to perform using existing statistical software for linear regression.

We illustrate the method using Feldman's clearance data.

Structure of Dataset

ORGAN	ID	DAYS	CFP	LOGCFP
lung	1	3	102	2.00860
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

SAS Code for Two-Stage Analysis

```
FILENAME rats ’g:\shared\bio226\rat.txt’;

DATA clear;
     INFILE rats;
     INPUT organ $ id days cfp logcfp;
     IF (days=0) THEN DELETE;
RUN;
```

Two-Stage Analysis

Stage 1:

```
PROC REG DATA=clear OUTEST=coeffs NOPRINT;
      BY id organ;
      MODEL logcfp=days;
RUN;
```

Note: This creates the following two variables that are of interest, intercept and days (the estimated intercepts and slopes respectively).

```
PROC PRINT DATA=coeffs;
      VAR id organ intercept days;
RUN;
```

```
OBS   ID   ORGAN   INTERCEPT        DAYS

  1    1    lung     2.05235   -0.017569
  2    2    lung     1.97683   -0.012858
  3    3    lung     1.99249   -0.017565
  4    4    lung     2.12824   -0.023480
  5   26    liver    2.06173   -0.011100
  6   28    liver    2.05379   -0.011425
  7   30    liver    1.95025   -0.008306
  8   31    liver    2.12560   -0.018886
```

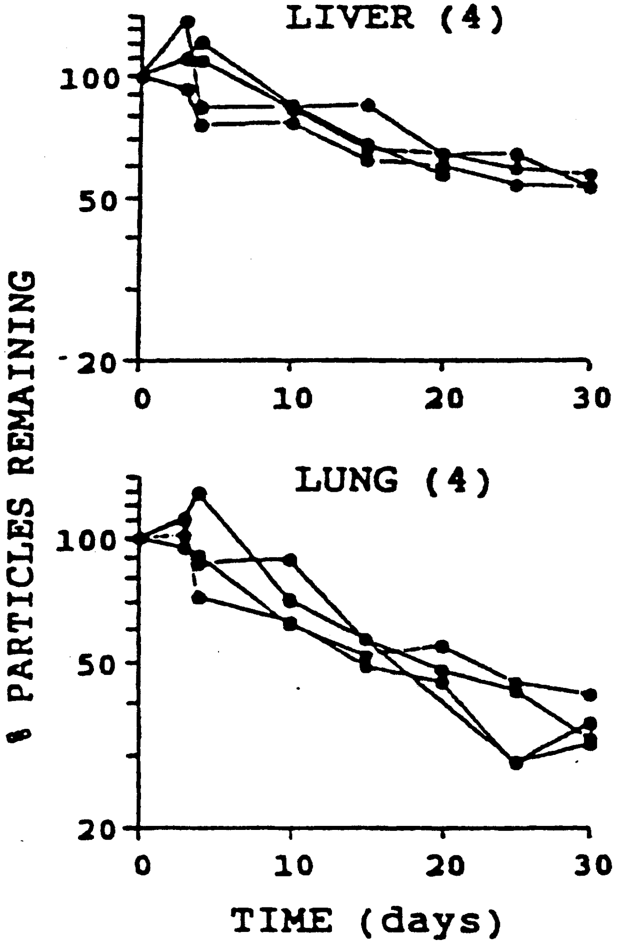


Figure 25: Timeplot of Feldman's clearance data.

Stage 2:

```
PROC GLM DATA=coeffs;
      CLASS organ;
      MODEL intercept=organ / SOLUTION;
      TITLE 'ANOVA for the Intercepts';
RUN;

PROC GLM DATA=coeffs;
      CLASS organ;
      MODEL days=organ / SOLUTION;
      TITLE 'ANOVA for the Slopes';
RUN;
```

ANOVA for the Intercepts

General Linear Models Procedure

```
Dependent Variable: Intercept

                       Sum of          Mean
Source    DF         Squares        Square      F Value    Pr > F

Model      1      0.00021482    0.00021482         0.04    0.8425
Error      6      0.02995984    0.00499331

Total      7      0.03017466


                                  Standard
Parameter          Estimate          Error    t Value    Pr > |t|

Intercept       2.037476922     0.03533167      57.67      <.0001
Organ liver     0.010363771     0.04996652       0.21      0.8425
```

ANOVA for the Slopes

```
              General Linear Models Procedure

Dependent Variable: days

                        Sum of          Mean
Source      DF        Squares         Square      F Value    Pr > F

Model        1     0.00005916     0.00005916         3.00    0.1339
Error        6     0.00011825     0.00001971

Total        7     0.00017741



                                     Standard
Parameter          Estimate             Error     t Value    Pr > |t|

Intercept      -.0178678950        0.00221968       -8.05      0.0002
Organ liver    0.0054387390        0.00313910        1.73      0.1339
```

Estimated slope in the lung group is -0.0178, representing a half time for clearance of 16.9 days (or $\frac{\log_{10}(0.5)}{-0.0178}$).

Estimated slope in the liver group is -0.0124 (-0.0178 + 0.0054), representing a half time for clearance of 24.2 days.

The mean slopes in the two groups are not discernibly different ($p = .13$).

The mean intercepts do not differ significantly in the two groups ($p = .84$), as would be expected given the normalization of each animal's data to baseline.

In summary, the two-stage analysis is easy to understand and nearly efficient when the dataset is balanced and complete.

It is somewhat less attractive when the number and timing of observations varies among subjects, because it does not take proper account of the weighting.

In contrast, we can consider the mixed effects model corresponding to the two-stage model, and obtain efficient (more precise) estimates of the regression coefficients.

Mixed Effects Model Representation

We can develop a mixed effects model in two stages corresponding to the two-stage model:

Stage 1:

$$Y_{ij} = \beta_{1i} + \beta_{2i} t_{ij} + \epsilon_{ij}$$

where β_{1i} is the intercept for the i^{th} subject,

β_{2i} is the slope for the i^{th} subject, and

errors, ϵ_{ij}, are assumed to be independent and normally distributed around the individual's regression line, that is, $\epsilon_{ij} \sim N\left(0, \sigma^2\right)$.

Stage 2:

Assume that the intercept and slope, β_{1i} and β_{2i}, are random and have a joint multivariate normal distribution, with mean dependent on covariates (e.g., the organ studied):

$$\beta_{1i} = \beta_1 + \beta_2 \text{ Organ } + b_{1i}$$

$$\beta_{2i} = \beta_3 + \beta_4 \text{ Organ } + b_{2i}$$

Also, let $\text{Var}\,(b_{1i}) = g_{11}$, $\text{Cov}\,(b_{1i}, b_{2i}) = g_{12}$, $\text{Var}\,(b_{2i}) = g_{22}$.

If we substitute the expressions for β_{1i} and β_{2i} into the equation in stage 1, we obtain

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 \text{ Organ } + \beta_3 t_{ij} + \beta_4 \text{ Organ } \times t_{ij} \\ &\quad + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij} \end{aligned}$$

The first four terms give the regression model for the mean response implied by the two-stage model.

The last three terms are the “error terms” (between and within-subject).

This model can be fit using the ***RANDOM*** statement in PROC MIXED.

PROC MIXED in SAS

```
FILENAME rats 'g:\shared\bio226\rat.txt';

DATA clear;
    INFILE rats;
    INPUT organ $ id days cfp logcfp;
    IF (days=0) THEN DELETE;
RUN;

PROC MIXED DATA=clear;
    CLASS id organ;
    MODEL logcfp=days organ days*organ / S CHISQ;
    RANDOM INTERCEPT days /
        TYPE=UN SUBJECT=ID G;
    TITLE 'Random Slopes and Intercepts';
RUN;
```

Random Slopes and Intercepts

Estimated G Matrix

Parameter	ID	Row	col1	col2
Intercept	1	1	0.002851	-0.00015
days	1	2	-0.00015	9.65E-6

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
UN(1,1)	ID	0.002851
UN(2,1)	ID	-0.00015
UN(2,2)	ID	9.65E-6
Residual		0.003155

Random Slopes and Intercepts
Solution for Fixed Effects

Effect	organ	Estimate	Standard Error	DF	t Value	Pr > $\|t\|$
Intercept		2.0375	0.03337	6	61.05	<.0001
days		-0.01785	0.001913	6	-9.33	<.0001
organ	liver	0.003814	0.04741	37	0.08	0.9363
organ	lung	0	.	.	.	.
days*organ	liver	0.006232	0.002760	37	2.26	.0299
days*organ	lung	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	Pr > ChiSq
days	1	6	114.05	<.0001
organ	1	37	0.01	0.9359
days*organ	1	37	5.10	0.0239

In contrast to results from two-stage analysis, results suggest that mean clearance of foreign particles is faster from the lung.

Estimated slope in the lung group is -0.0178, representing a half time for clearance of 16.9 days (or $\frac{\log_{10}(0.5)}{-0.0178}$).

Estimated slope in the liver group is -0.0116 (-0.0178 + 0.0062), representing a half time for clearance of 26.0 days.

The mean slopes in the two groups are different ($p < 0.05$).

Side-by-Side Comparison of Results

	Two-Stage		**GC**	
			(Mixed Effects)	
Intercept	2.0375	(.0353)	2.0375	(.0334)
Day	-0.0178	(.0022)	-0.0179	(.0019)
Organ	0.0104	(.0450)	0.0038	(.0474)
Organ*Time	0.0054	(.0031)	0.0062	(.0027)

Summary

The two-stage method is less attractive when the number and timing of observations varies among subjects, because it does not take proper account of the weighting.

Also, note that the two-stage formulation of the growth curve model imposes certain restrictions and structure on the covariates.

That is, in the two-stage approach covariates at the first stage (except for the intercept) must be *time-varying*, while covariates at the second stage must be *time-invariant*.

In contrast, in the mixed effects model the only restriction is that the components of Z_{ij} are a subset of the components of X_{ij}.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 13

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Linear Mixed Effects Model and Prediction

Recall that in the linear mixed effects model,

$$Y_{ij} = X_{ij}'\beta + Z_{ij}'b_i + \epsilon_{ij},$$

we can distinguish between the conditional mean,

$$E(Y_{ij}|X_{ij}, b_i) = X_{ij}'\beta + Z_{ij}'b_i,$$

and the marginal mean,

$$E(Y_{ij}|X_{ij}) = X_{ij}'\beta.$$

The former describes the mean response for an individual, the latter describes the mean response averaged over individuals.

The distinction between the conditional and marginal means is best understood with a simple example.

Consider the simple random intercepts and slopes model,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij},$$

In this model, we can distinguish the conditional mean for an individual,

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

(see broken lines for subjects A and B in Figure 26), and the marginal mean averaged over individuals,

$$E(Y_{ij}) = \beta_1 + \beta_2 t_{ij},$$

(see solid line in Figure 26).

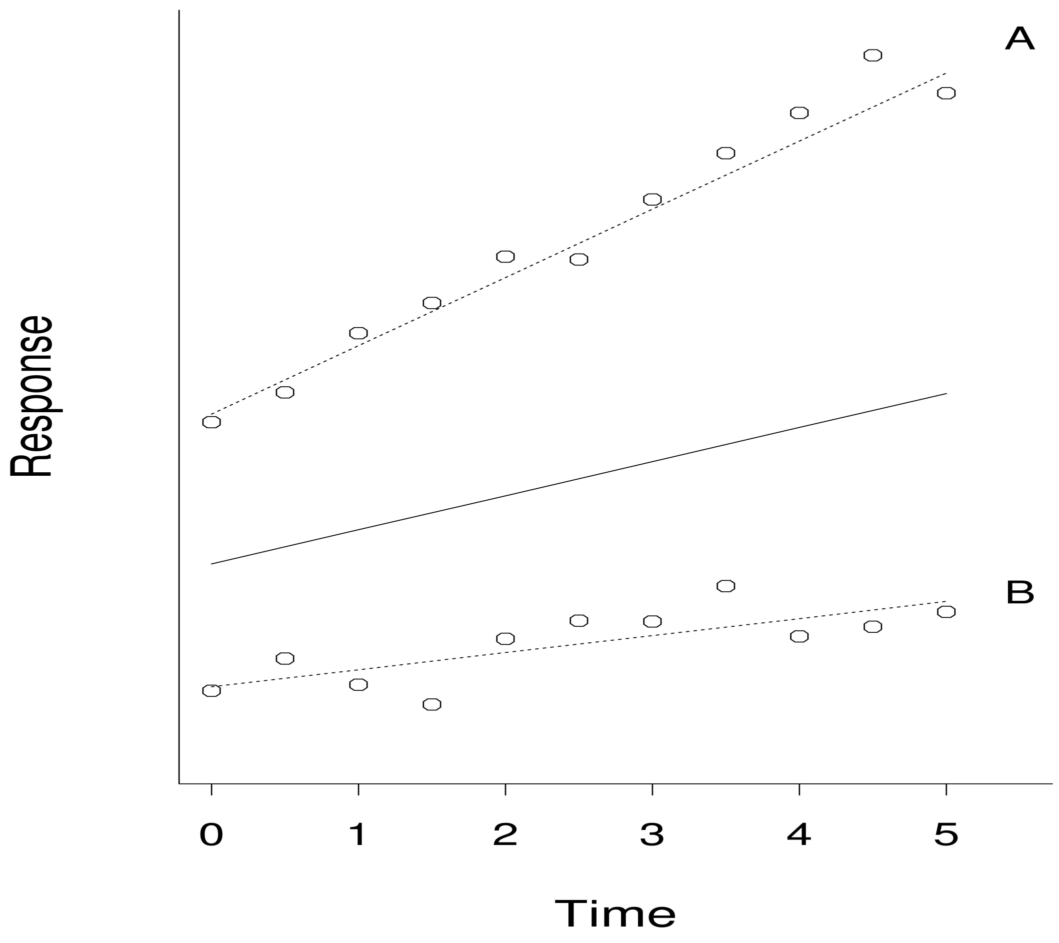


Figure 26: Graphical representation of the overall and subject-specific mean responses over time, plus measurement errors.

Prediction of Random Effects

In many applications, inference is focused on fixed effects, $\beta_1, \beta_2, ..., \beta_p$.

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

Technically, because the b_i are random, we customarily talk of "predicting" the random effects rather than "estimating" them.

Using maximum likelihood, the prediction of b_i, say $\widehat{b}_i$, is given by:

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

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

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

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

It can be shown that the BLUP for b_i is:

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

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_b^2 is large relative to σ^2.

For the general case, the prediction of b_i is given by:

$$\widehat{b}_i = E(b_i | Y_i; \widehat{\beta}, \widehat{G}, \widehat{\sigma}^2) = \widehat{G} Z_i' \widehat{\Sigma}_i^{-1} (Y_i - X_i \widehat{\beta}),$$

where $\Sigma_i = \text{Cov}\,(Y_i | X_i) = Z_i G Z_i' + R_i = Z_i G Z_i' + \sigma^2 I.$

When the unknown covariance parameters have been replaced by their ML or REML estimates, the resulting predictor is often referred to as the "Empirical BLUP" or the "Empirical Bayes" (EB) estimator.

Finally, the i^{th} subject's predicted response profile is,

$$\begin{aligned} \widehat{Y}_i &= X_i \widehat{\beta} + Z_i \widehat{b}_i \\ &= X_i \widehat{\beta} + Z_i \widehat{G} Z_i' \widehat{\Sigma}_i^{-1} (Y_i - X_i \widehat{\beta}) \\ &= (\widehat{R}_i \widehat{\Sigma}_i^{-1}) X_i \widehat{\beta} + (I - \widehat{R}_i \widehat{\Sigma}_i^{-1}) Y_i \end{aligned}$$

That is, the i^{th} subject's predicted response profile is a weighted combination of the population-averaged mean response profile, $X_i\widehat{\beta}$, and the i^{th} subject's observed response profile Y_i.

Subject's predicted response profile is "shrunk" towards population-averaged mean response profile.

Amount of "shrinkage" depends on relative magnitude of R_i and Σ_i.

Note that R_i characterizes the within-subject variability, while Σ_i incorporates both within-subject and between-subject sources of variability.

$$R_i\Sigma_i^{-1} = \frac{\text{within-subject variability}}{\text{within-subject + between-subject variability}}.$$

Thus, $R_i\Sigma_i^{-1}$ denotes the fraction of total variability that is due to within-subject (or measurement error) variation.

Similarly, $(I - R_i\Sigma_i^{-1})$ denotes the fraction of total variability that is due to between-subject variation.

When within-subject variability is large relative to between-subject variability, more weight is given to $X_i\widehat{\beta}$, the population-averaged mean response profile (more "shrinkage").

PROC MIXED in SAS

The Empirical Bayes (EB) estimates, $\widehat{b}_i$, can be obtained by using the following option on the RANDOM statement in PROC MIXED:

RANDOM INTERCEPT time / TYPE=UN SUBJECT=id S;

Alternatively, a subject's predicted response profile,

$$\widehat{Y}_i = X_i\widehat{\beta} + Z_i\widehat{b}_i,$$

can be obtained by using the following option on the MODEL statement:

MODEL y = trt time trt*time / OUTP=*SAS-data-set***;**

Example: *Exercise Therapy Study*

Consider a model with randomly varying intercepts and slopes, and which allows the mean values of the intercept and slope to differ in the two treatment groups.

To fit this model, use the following SAS code:

```
PROC MIXED DATA=stren;
      CLASS id trt;
      MODEL y=trt time time*trt / S CHISQ;
      RANDOM INTERCEPT time / TYPE=UN SUBJECT=id G S;
```

Empirical Bayes Estimates of b_i

Solution for Random Effects

			Std Err		
Effect	id	Estimate	Pred	t Value	Pr > $\lvert t \rvert$
Intercept	1	-1.0111	0.9621	-1.05	0.2959
time	1	-0.03812	0.08670	-0.37	0.7144
Intercept	2	3.3772	0.9621	1.07	0.0007
time	2	0.1604	0.08670	1.85	0.0672
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Example: *Exercise Therapy Study*

Next, we consider how to obtain a subject's predicted response profile.

```
PROC MIXED DATA=stren;
      CLASS id trt;
      MODEL y=trt time time*trt / S CHISQ OUTP=predict;
      RANDOM INTERCEPT time / TYPE=UN SUBJECT=id G S;

PROC PRINT DATA=predict;
      VAR id trt time y Pred StdErrPred Resid;
```

Predicted Response Profiles

					StdErr	
id	trt	time	y	Pred	Pred	Resid
1	1	0	79	78.9937	0.59729	0.00634
1	1	4	79	79.4071	0.39785	-0.40707
1	1	6	80	79.6138	0.36807	0.38623
1	1	8	80	79.8205	0.40451	0.17952
1	1	12	80	80.2339	0.61057	-0.23389
2	1	0	83	83.3820	0.59729	-0.38202
2	1	4	85	84.5644	0.39785	0.43562
2	1	6	85	85.1556	0.36807	-0.15557
2	1	8	86	85.7468	0.40451	0.25325
2	1	12	87	86.9291	0.61057	0.07088
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

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.

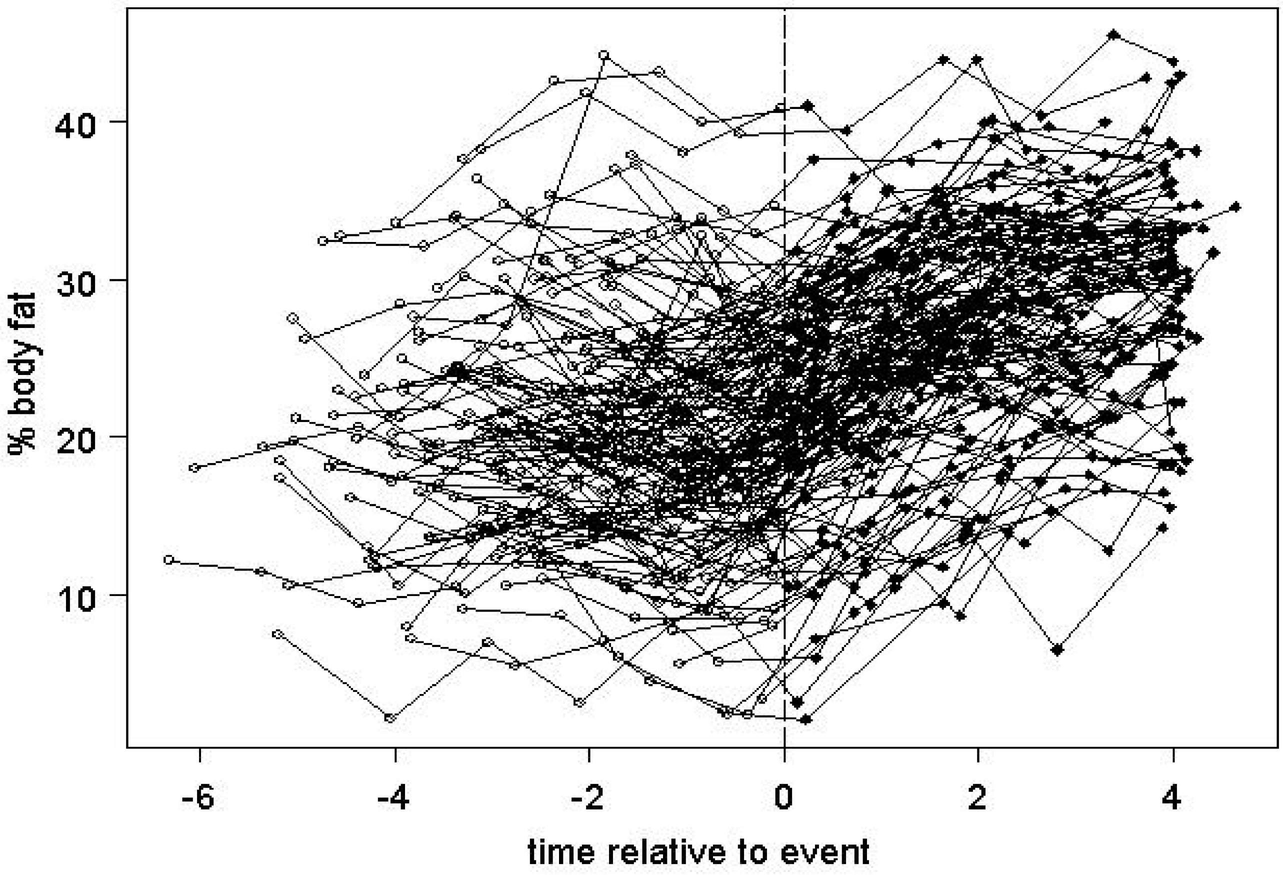


Figure 27: Timeplot of percent body fat against time, relative to age of menarche (in years).

Consider hypothesis that %body fat 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 28).

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).

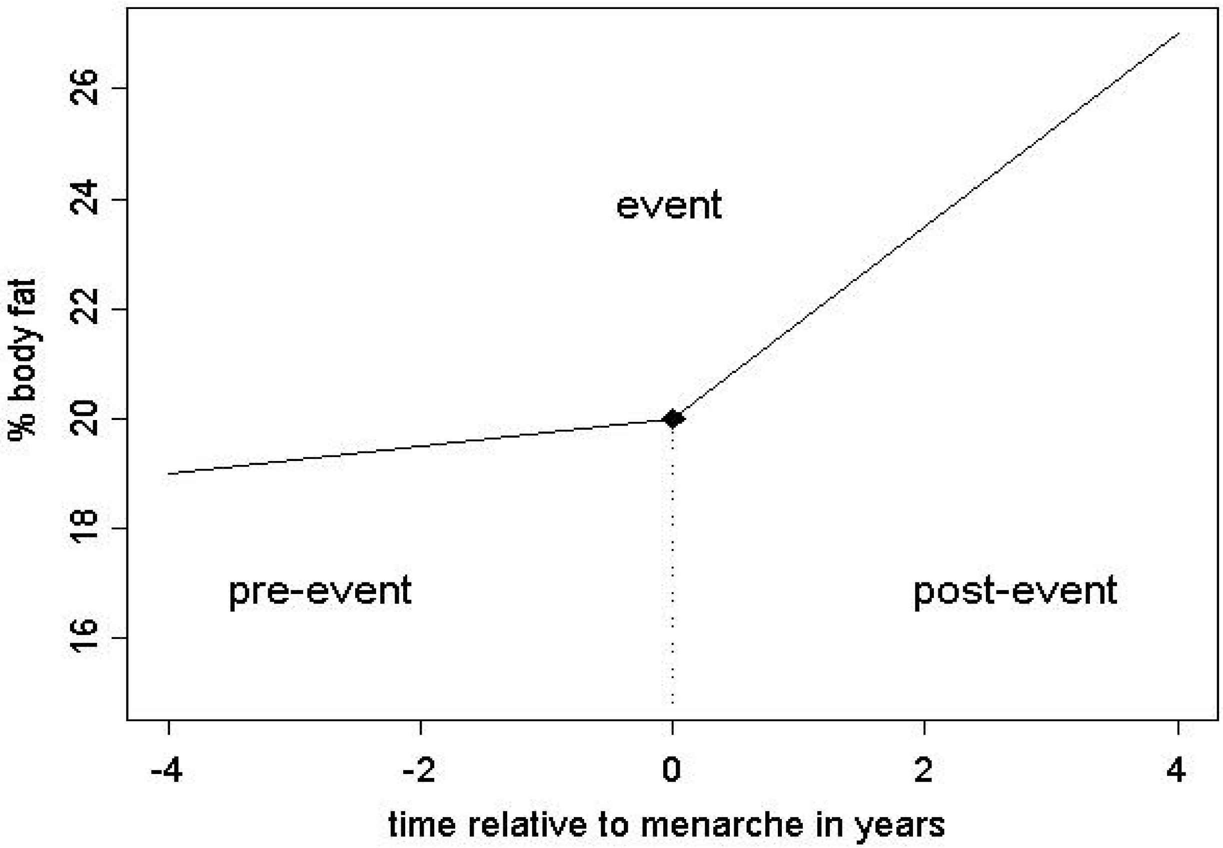


Figure 28: Graphical representation of piecewise linear trajectory.

We consider the following linear mixed effects model

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+,$$

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_1 is the average %body fat at menarche (when $t_{ij} = 0$).

The slope β_2 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 $(\beta_2 + \beta_3)$.

Goal: Assess whether population slopes differ before and after menarche, i.e., $H_0 : \beta_3 = 0$.

Similarly, $(\beta_1 + b_{1i})$ is intercept for i^{th} subject and is the true %body fat at menarche (when $t_{ij} = 0$).

$(\beta_2 + b_{2i})$ 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 $[(\beta_2 + \beta_3) + (b_{2i} + b_{3i})]$.

Interpretation of variance components:

Recall that the subject-specific intercepts, $(\beta_1 + b_{1i})$, have mean β_1 and variance $g_{11} = \sigma^2_{b_{1i}}$.

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

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

Variance components for b_{2i} and b_{3i} can be interpreted in similar fashion.

Table 39: 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 40: Estimated covariance of the random effects and standard errors for the percent body fat data.

PARAMETER	ESTIMATE	SE	Z
$\text{Var}(b_{1i})$	45.9413	5.7393	8.00
$\text{Var}(b_{2i})$	1.6311	0.4331	3.77
$\text{Var}(b_{3i})$	2.7497	0.9635	2.85
$\text{Cov}(b_{1i}, b_{2i})$	2.5263	1.2185	2.07
$\text{Cov}(b_{1i}, b_{3i})$	-6.1096	1.8730	-3.26
$\text{Cov}(b_{2i}, b_{3i})$	-1.7505	0.5980	-2.93
$\text{Var}(\epsilon_i) = \sigma^2$	9.4732	0.5443	17.40

Results

Estimated intercept, $\widehat{\beta}_1 = 21.36$, 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_2, 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, $\beta_2 + \beta_3$, 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 $\widehat{\beta}_3$, 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 45.94, indicating substantial variability from girl to girl in true percent body fat at menarche, $\beta_1 + b_{1i}$.

For example, approximately 95% of girls have true percent body fat between 8.08% and 34.65% $\left(\text{i.e., } 21.36 \pm 1.96 \times \sqrt{45.94}\right)$.

Estimated variance of b_{2i} 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 during the post-menarcheal period, $\text{Var}(b_{2i} + b_{3i})$, is 0.88 (or $[1.63 + 2.75 - 2 \times 1.75]$), indicating less variability in the slopes after menarche.

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

Results indicate that more than 95% of girls are expected to have increases in body fat during the post-menarcheal period.

Substantially fewer (approximately 63%) are expected to have increases in body fat during the pre-menarcheal period.

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 40.

Strength of correlation declines over time, but does not decay to zero even when measurements are taken 8 years apart (see Table 41).

Table 41: Marginal correlations (off-diagonals) 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	0.78	0.71	0.61	0.60	0.57	0.52	0.47
0.82	54.9	0.81	0.76	0.70	0.68	0.64	0.60	0.54
0.78	0.81	51.8	0.80	0.76	0.74	0.71	0.66	0.60
0.71	0.76	0.80	52.0	0.81	0.79	0.76	0.71	0.64
0.61	0.70	0.76	0.81	55.4	0.81	0.78	0.73	0.66
0.60	0.68	0.74	0.79	0.81	49.1	0.79	0.76	0.70
0.57	0.64	0.71	0.76	0.78	0.79	44.6	0.77	0.74
0.52	0.60	0.66	0.71	0.73	0.76	0.77	41.8	0.76
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, based on the $\widehat{\beta}$'s and $\widehat{b}_i$'s.

Figure 29 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 30).

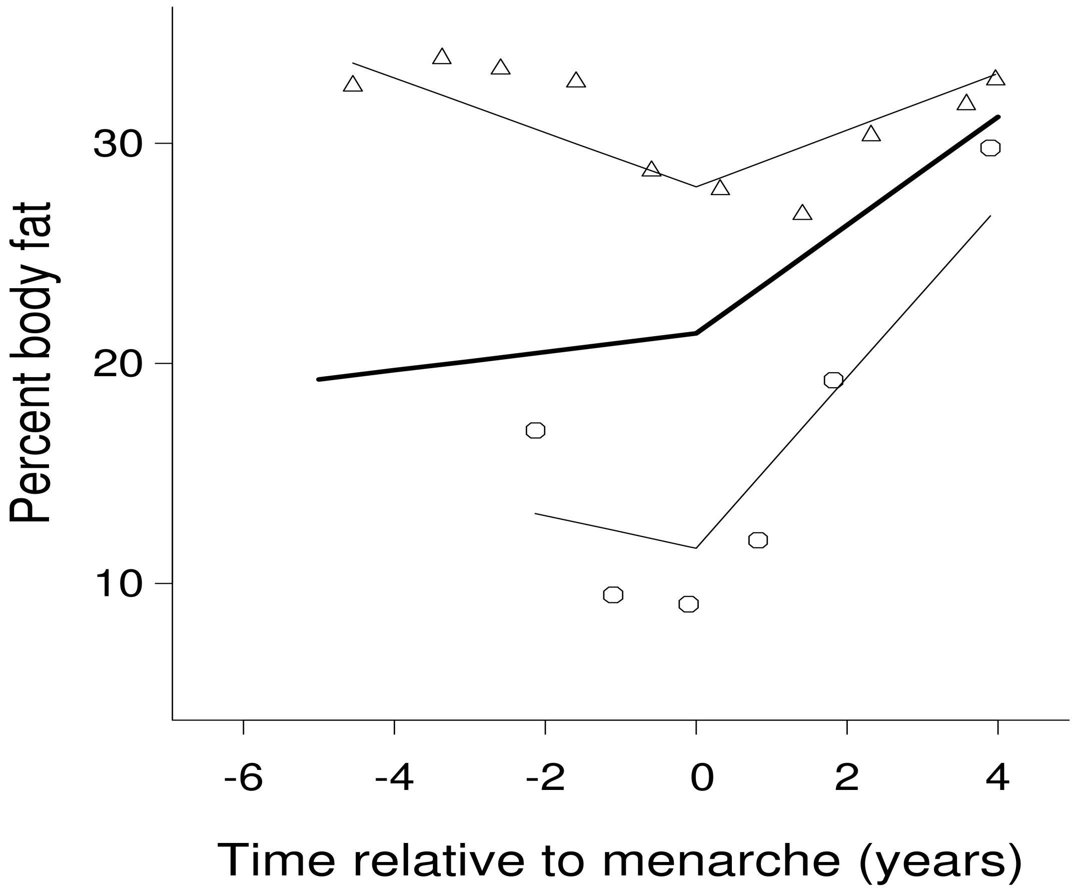


Figure 29: Population average curve and empirical BLUPs for two randomly selected girls.

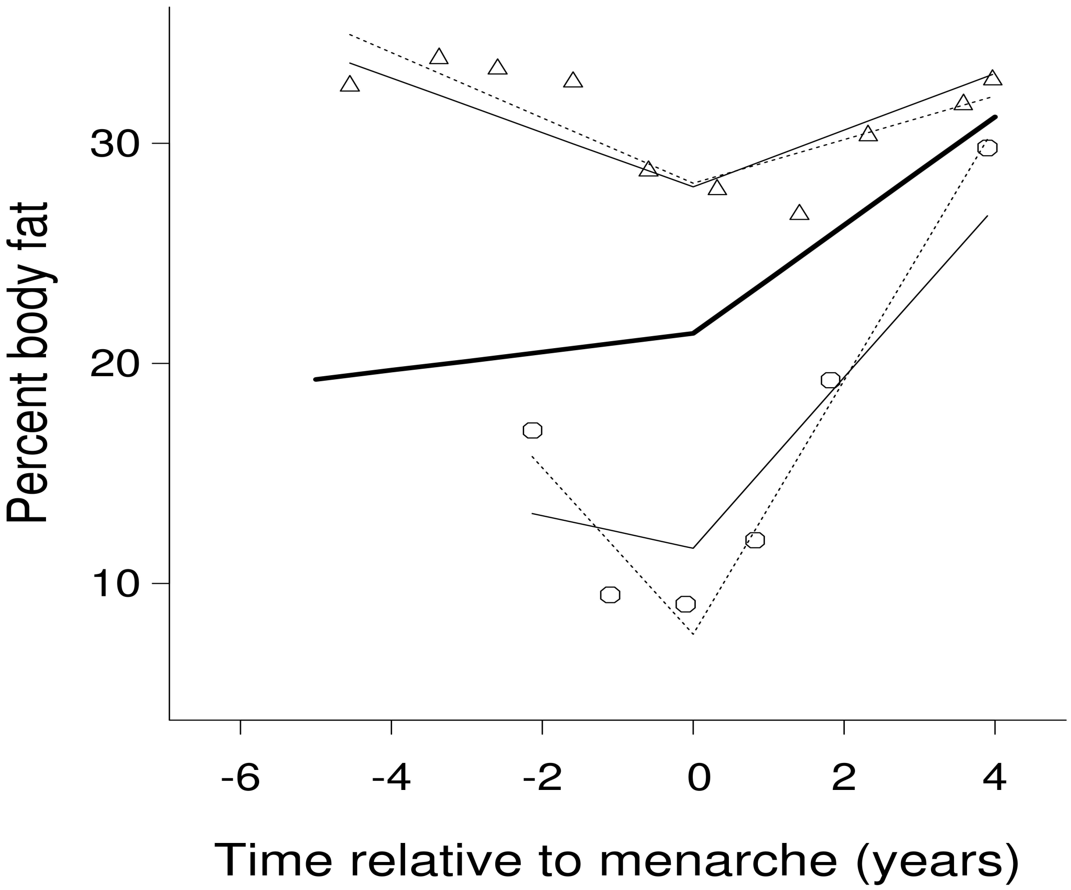


Figure 30: Population average curve, empirical BLUPs, and OLS predictions for two randomly selected girls.

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, but does not change the model for $E(Y_{ij}|X_{ij})$.

The inclusion of random slopes or random trajectories induces a random effects covariance structure for $Y_{i1}, ..., Y_{in_i}$ where the variances and correlations are a function of the times of measurement.

In general, the random effects covariance structure is relatively parsimonious (e.g., random intercepts and slopes model has only four parameters, $\sigma^2_{b_1}, \sigma^2_{b_2}, \sigma_{b_1,b_2},$ and σ^2).

Linear mixed effects models are appealing because of

- their flexibility in accommodating a variety of study designs, data models and hypotheses.

- their flexibility in accommodating any degree of imbalance in the data (e.g., due to mistimed measurements and/or missing data)

- their ability to parsimoniously model the variance and correlation

- their ability to predict *individual* trajectories over time

Note 1: Tests of fixed effects rely on asymptotic normality of the fixed effects (not Y_{ij}); need reasonable (say > 30) number of subjects.

Note 2: Missing observations can be accommodated easily, validity of results depends upon assumption about missingness (see Lecture 14).

Linear Mixed Models using PROC MIXED in SAS

Table 42: Illustrative commands for a linear mixed effects model, with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
PROC MIXED;
   CLASS id group;
   MODEL y=group time group*time / SOLUTION CHISQ;
   RANDOM INTERCEPT time / SUBJECT=id TYPE=UN G V;
```

Table 43: Illustrative commands for obtaining the estimated BLUPs and predicted responses from model with randomly varying intercepts and slopes, using PROC MIXED in SAS.

```
PROC MIXED;
  CLASS id group;
  MODEL y=group time group*time / SOLUTION CHISQ OUTPRED=yhat;
  RANDOM INTERCEPT time / SUBJECT=id TYPE=UN SOLUTION;

PROC PRINT;
  VAR id group time y PRED;
```

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 14

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Missing Data and Dropout

In longitudinal studies missing data are the rule not the exception.

The term “missing data” is used to indicate that an intended measurement could not be obtained.

With missing data there must necessarily be some loss of information.

Of greater concern, missing data can introduce bias and result in misleading inferences about change over time.

When data are missing we must carefully consider the reasons for missingness.

Unequal n_i per Subject or Unbalanced Designs

Basically, the methods that we have discussed so far can handle situations in which $n_i \neq n$ for all i (i.e., unequal number of observations per subject).

That is, modern regression methods can handle unbalanced longitudinal designs with relative ease.

However, we do need to be more careful when $n_i \neq n$ for all i due to missingness.

Missing Data

Why might we have $n_i \neq n$ for all i?

For most designed studies, we *plan* on measuring the same number of outcomes, so if $n_i \neq n$ for all i, then some outcomes are *missing*.

Let Y denote the complete response vector which can be partitioned into two sub-vectors:

(i) Y^O the measurements observed
(ii) Y^M the measurements that are missing

If there were no missing data, we would have observed the complete response vector Y.

Instead, we get to observe Y^O.

The main problem with missing data is that distribution of the observed data may not be the same as distribution of the complete data.

Consider the following simple illustration:

Suppose we intend to measure subjects at 6 months (Y_1) and 12 months (Y_2) post treatment.

All of the subjects return for measurement at 6 months, but many do not return at 12 months.

If subjects fail to return at 12 months because they are not well (say, values of Y_2 are low), then distribution of observed Y_2's will be positively skewed compared to distribution of Y_2's in the population of interest.

When data are missing we must carefully consider the reasons for missingness.

Estimation of β with missing data depends on the missing data mechanism.

The missing data mechanism is a probability model for missingness:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Not Missing at Random (NMAR)

Missing Completely at Random (MCAR)

MCAR: probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained (the *missing* responses) or the set of *observed* responses.

MCAR: probability responses are missing is independent of Y^O and Y^M.

Missingness is simply the result of a chance mechanism that is unrelated to either observed or unobserved components of the outcome vector.

Consequently, observed data can by thought of as a random sample of the complete data.

Covariate-Dependent Missingness

If missingness depends only on X, then technically it is MCAR. However, sometimes this is referred to as *covariate dependent* non-response.

In general, if non-response depends on covariates, X, it is harmless and same as MCAR provided you always condition on the covariates (i.e., incorporate the covariate in the analysis).
This type of missingness is only a problem if you do not condition on X.

Example 1: Consider the case where missingness depends on treatment group. Then the observed means in each treatment group are unbiased estimates of the population means.

However, the marginal response mean, averaged over the treatment groups, is not unbiased for the corresponding mean in the population (the latter, though, is usually not of subject-matter interest).

Sometimes it may be necessary to introduce additional covariates, or stratifying variables, into the analysis to control for potential bias due to missingness.

Example 2: Suppose the response Y is some measure of health, and X_1 is an indicator of treatment, and X_2 is an indicator of side-effects. Suppose missingness depends on side-effects.

If side-effects and outcome are uncorrelated, then there will be no bias.

If side-effects and outcome are correlated, then there will be bias unless you stratify the analysis on both treatment and side-effects (analogous to confounding).

Features of MCAR

The means, variances, and covariances are preserved.

So, if

$$\underset{n\times 1}{E\ (Y_i)} = \underset{n\times p}{X_i}\beta \quad \text{with complete data}$$

then

$$\underset{n_i\times 1}{E\ (Y_i)} = \underset{n_i\times n}{I_i} \ \underset{n\times p}{X_i}\beta = \underset{n_i\times p}{X_i}\,\beta$$

$$\underset{n_i\times n_i}{\text{Cov}\ (Y_i)} = \underset{n_i\times n}{I_i}\ \Sigma\ I_i' = \Sigma_i$$

where I_i is identity matrix with rows corresponding to missing values removed.

MCAR:

- Can use ML/REML estimators for β
- More generally, we can use GLS estimator with any "working" assumption for the covariance; normality assumption for Y_{ij} is not necessary
- If we use GLS estimator with incorrect "working" assumption for the covariance, then must use "empirical" or "sandwich" variance estimator for $\mathrm{Cov}(\widehat{\beta})$

Any method of analysis that yields valid inferences in absence of missing data is also valid when missing data are MCAR and analysis is based on all available data, or even when restricted to so-called "completers".

Given that valid estimates of the means, variances, and covariances can be obtained, GLS provides valid estimates of β without requiring any distributional assumptions for Y_i.

The GLS estimator of β is valid provided the model for the mean response has been correctly specified; it does not require any assumptions about the joint distribution of the longitudinal responses.

$\Longrightarrow$ With complete data or data MCAR, normality assumption is not required.

Missing at Random (MAR)

MAR: probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that, in principle, should have been obtained.

MAR: probability that responses are missing depends on Y^O, but is conditionally independent of Y^M.

Note 1: If subjects are stratified on the basis of similar values for the responses that have been observed, then within strata missingness is simply the result of a chance mechanism unrelated to unobserved responses.

Note 2: Because missingness depends on observed responses, the distribution of Y_i in each of the distinct strata defined by the patterns of missingness is not the same as the distribution of Y_i in the target population.

The “completers” are a biased sample from the target population.

Features of MAR

Means, variances, and covariances are <u>not</u> preserved:

So, if

$$\underset{n\times 1}{E\ (Y_i)} \;=\; \underset{n\times p}{X_i\beta} \quad \text{with complete data}$$

In general

$$\underset{n_i\times 1}{E\ (Y_i)} \;\neq\; \underset{n_i\times p}{X_i\,\beta}, \quad \text{Cov}(Y_i) \neq \Sigma_i$$

This implies that sample means, variances, and covariances based on either the "completers" or the available data are biased estimates of the corresponding parameters in the target population.

However, the likelihood is preserved.

For example, in the linear models for longitudinal data the appropriate likelihood assumes

$$Y_i \sim N(X_i\beta, \Sigma_i).$$

Because missingness only depends on observed data, likelihood factors into one piece depending on (β, Σ_i), another depending on Y_i and missingness indicators.

Valid inferences for (β, Σ_i) are obtained by maximizing the first piece (and "ignoring" the second piece) of the likelihood.

Note: Observed responses are not necessarily normally distributed.

ML estimation (e.g., PROC MIXED) of β is valid when data are MAR provided the multivariate normal distribution has been correctly specified.

This requires correct specification of not only the model for the mean response, but also the model for the covariance among the responses.

In a sense, ML estimation allows the missing values to be validly "predicted" or "imputed" using the observed data and a correct model for the joint distribution of the responses.

Not Missing at Random (NMAR)

NMAR: probability that responses are missing is related to the specific values that should have been obtained.

An NMAR mechanism is often referred to as “non-ignorable” missingness.

Challenging problem and requires modelling of missing data mechanism; moreover, specific model chosen can drive results of analysis.

Sensitivity analyses is recommended.

Dropout

Longitudinal studies often suffer from problem of attrition; i.e., some individuals “drop out” of the study prematurely.

This is where an individual is observed from baseline up until a certain point in time, thereafter no more measurements are made.

Term *dropout* refers to special case where if Y_{ik} is missing, then $Y_{ik+1}, ..., Y_{in}$ are also missing.

This gives rise to so-called “monotone” missing data pattern displayed in Figure 31.

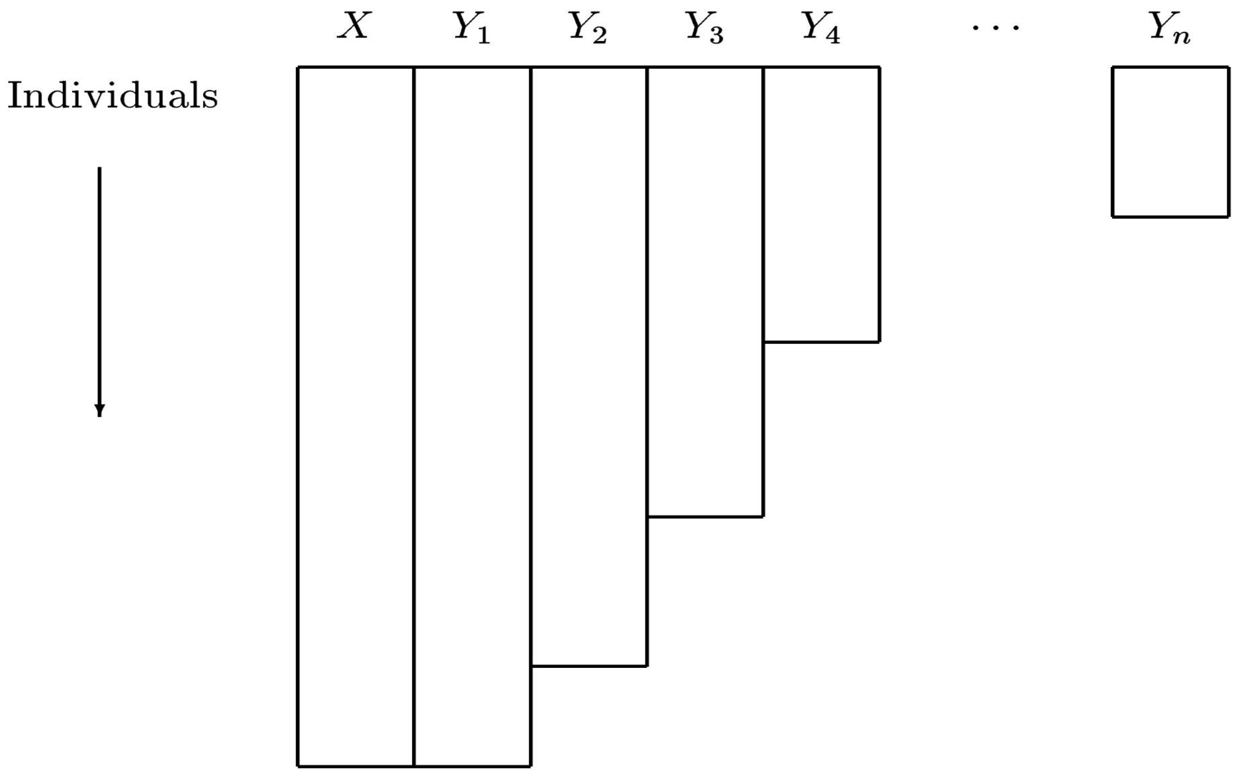


Figure 31: Schematic representation of a monotone missing data pattern for dropout, with Y_j more observed than Y_{j+1} for $j = 1, ..., n-1$.

Possible reasons for dropout:

1. Recovery
2. Lack of improvement or failure
3. Undesirable side effects
4. External reasons unrelated to specific treatment or outcome
5. Death

Examples

In clinical trials, monotone missing data can arise from a variety of circumstances:

a) ***Late entrants*:** If the study has staggered entry, at any interim analysis some individuals may have only partial response data.

   Usually, this sort of missing data does not introduce any bias.

b) ***Dropout*:** Individuals may drop out of a clinical trial because of side effects or lack of efficacy.

   Usually, this type of missing data is of concern, especially if dropout is due to lack of efficacy.
   Dropout due to lack of efficacy suggests that those who drop out come from the lower end of the spectrum.
   Dropout due to side effects may or may not be a problem, depending upon the relationship between side effects and the outcome of interest.

When there is dropout, key issue is whether those who “drop out” and those who remain in the study differ in any further relevant way.

If they do differ, then there is potential for bias.

The taxonomy of missing data mechanisms (MCAR, MAR, and NMAR) discussed earlier can be applied to dropout.

Common Approaches for Handling Dropout

Complete-Case Analysis:

Exclude all data from the analysis on any subject who drops out.

That is, a so-called "complete-case" analysis can be performed by excluding any subjects that do not have data at all intended measurement occasions.

This method is very problematic and is rarely an acceptable approach to the analysis.

It will yield unbiased estimates of mean response trends only when dropout is MCAR.

Even when MCAR assumption is tenable, complete-case analysis can be immensely inefficient.

Available-Data Analysis:

General term that refers to a wide collection of techniques that can readily incorporate vectors of repeated measures of unequal length in the analysis.

Standard applications of GLS are available-data methods.

In general, available-data methods are more efficient than complete-case methods.

Drawbacks of available-data methods:

(i) Sample base of cases changes over measurement occasions.
(ii) Pairwise available-data estimates of correlations can lie outside (-1, 1).
(iii) Many available-data methods yield biased estimates of mean response trends unless dropout is MCAR.

Imputation

Imputation: substitute or fill-in the values that were not recorded with imputed values.

Once a filled-in data set has been constructed, standard methods for complete data can be applied.

Validity of method depends on how imputation is done.

Methods that rely on just a single imputation fail to acknowledge the uncertainty inherent in the imputation of the unobserved responses.

“Multiple imputation” circumvents this difficulty.

Multiple Imputation (MI): Missing values are replaced by a set of m plausible values, thereby acknowledging uncertainty about what values to impute.

Typically, a small number of imputations, for instance, $5 \leq m \leq 10$, is sufficient.

The m filled-in data sets produce m different sets of parameter estimates and their standard errors.

These are then combined to provide a single estimate of the parameters of interest, together with standard errors that reflect the uncertainty inherent in the imputation.

"Last Value Carried Forward" (LVCF):

One widely used imputation method, especially in clinical trials, is LVCF.

Regulatory agencies such as FDA seem to encourage the continuing use of LVCF.

LVCF makes a strong, and often very unrealistic, assumption that the responses following dropout remain constant at the last observed value prior to dropout.

There appears to be some statistical folklore that LVCF yields a *conservative* estimate of the comparison of an active treatment versus the control.

This is a gross misconception!

Except in very rare cases, we do not recommend the use of LVCF as a method for handling dropout.

Variations on the LVCF theme include baseline value carried forward and worst value carried forward.

Imputation methods based on drawing values of missing responses from the conditional distribution of the missing responses given the observed responses have a much firmer theoretical foundation.

Then subsequent analyses of the observed and imputed data are valid when dropouts are MAR (or MCAR).

Furthermore, multiple imputation ensures that the uncertainty is properly accounted for.

Model-Based Imputation:

There is a related form of “imputation” where missing responses are *implicitly* imputed by modelling joint distribution of Y_i, $f(Y_i|X_i)$.

When dropout is MCAR or MAR, likelihood-based methods can be used based solely on the marginal distribution of the observed data.

In a certain sense, the missing values are validly predicted by the observed data via the model for the conditional mean of the missing responses given the observed responses (and covariates).

However, likelihood-based approaches require model for $f(Y_i|X_i)$ must be correctly specified (e.g., any misspecification of the covariance will, in general, yield biased estimates of the mean response trend).

Weighting Methods

In weighting methods, under-representation of certain response profiles in the observed data is taken into account and corrected.

These approaches are often called “propensity weighted” or “inverse probability weighted” methods.

Basic Idea: Base estimation on the observed responses but weight them to account for the probability of remaining in the study.

Intuition: Each subject’s contribution to the weighted analysis is replicated to count for herself and for those subjects with the same history of responses and covariates, but who dropout.

Propensities for dropout can be estimated as a function of observed responses prior to dropout and covariates.

Inverse probability weighted methods were first proposed in sample survey literature, where the weights are known.

In contrast, with dropout the weights are not known, but must be estimated from the observed data.

In general, weighting methods are valid provided model that produces the estimated weights is correctly specified.

Summary

In longitudinal studies missing data are the rule not the exception.

Missing data have two important implications:
(i) loss of information, and
(ii) validity of analysis.

The loss of information is directly related to the amount of missing data; it will lead to reduced precision (e.g., larger SEs, wider CIs) and reduced statistical power (e.g., larger p-values).

The validity of the analysis depends on assumptions about the missing data mechanism.

Likelihood-based methods (e.g., PROC MIXED) are valid under MAR or MCAR.

The distinction between MAR and MCAR determines the appropriateness of ML estimation under the assumption of normality versus GLS estimation without requiring distributional assumptions.

With complete data or data MCAR, normality assumption is not required.

With data MAR, normality assumption is required and correct models for both the mean response and the covariance.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 15

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Aspects of Design of Longitudinal Studies

In this lecture, we consider two issues concerning the design of longitudinal studies:

(1) Sample size and Power

(2) Longitudinal and Cross-Sectional Information

The first issue has important implications for planning a longitudinal study, the second has implications for analysis.

Sample Size and Power

Investigators typically need to know the answer to the following question: "How *large* should my study be?"

Answer is straightforward with only a single, univariate response: the *size* of a study = sample size.

For a longitudinal study the question of *size* is more complex, e.g., number of subjects, duration of study, frequency and spacing of repeated measurements on the subjects.

Before discussing sample size/power in context of longitudinal studies, we review sample size/power formulas for a univariate response.

We then describe a simple, albeit approximate, method that allows direct application of these formulas in longitudinal setting.

Sample Size for a Univariate Response

Interested in comparing two treatments (or exposures), denoted A and B.

Plan to randomize an equal number of subjects, say N, to two groups.

Two groups are to be compared in terms of the mean response.

Let $\mu^{(A)}$ and $\mu^{(B)}$ denote the mean response in the two populations.

Define effect of interest as $\delta = \mu^{(A)} - \mu^{(B)}$.

The null hypothesis of no group difference is H_0: $\delta = 0$.

Type I and Type II Errors

Recall: Two types of errors can arise when testing H_0: $\delta = 0$.

Type I error: If we reject the null hypothesis when in fact it is true.

Thus, for our example where H_0: $\delta = 0$,

$$\alpha = \Pr(\text{Reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ is true}).$$

The probability of type I error, also known as the significance level, is usually denoted by α.

Conventionally, α is chosen to be no greater than 0.05.

Type II error: If we fail to reject the null hypothesis when in fact it is false.

We denote probability of a type II error by γ, with

$$\gamma = \Pr(\text{Fail to reject H}_0 \mid \text{H}_0 \text{ is false}).$$

Note: γ necessarily depends upon the particular choice of value for $\delta \neq 0$ under the alternative hypothesis.

Finally, power of test is defined as $1 - \gamma$, that is,

$$\text{power} = 1 - \gamma = \Pr(\text{Reject H}_0 \mid \text{H}_0 \text{ is false}).$$

Two-Group Sample Size Formula

For two group comparison, a formula for approximate sample size in each group, N, is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\,\sigma^2}{\delta^2}, \text{ where}$$

σ^2 is variance of response (assumed to be common in two groups), and $Z_{(1-\alpha/2)}$ and $Z_{(1-\gamma)}$ denote the $(1 - \alpha/2) \times 100\%$ and $(1 - \gamma) \times 100\%$ percentiles of a standard normal distribution (e.g., the 97.5th percentile of a standard normal distribution is 1.96).

Note: N denotes sample size in each group; total sample size is $2N$.

Closer examination of formula reveals that the determination of sample size requires

(i) significance level, α;
(ii) power, $1 - \gamma$;
(iii) effect size, δ; and
(iv) common variance, σ^2.

Conventionally, α is fixed at mythical 0.05 level (with $Z_{(1-\alpha/2)} = 1.96$ for a 2-tailed test).

Similarly, lower bound on acceptable power is usually set at approx. 80% (with $Z_{(1-\gamma)} = 0.842$ for power $= 0.8$, or $Z_{(1-\gamma)} = 1.282$ for power $= 0.9$).

Investigators must provide information on: minimum effect size of scientific interest, δ, and an estimate of σ^2.

Sample Size for Longitudinal Response

Interested in comparing two treatments (or exposures), denoted A and B.

Plan to randomize an equal number of subjects, say N, to two groups.

Plan to take n repeated measurements of the response (not necessarily equally spaced measurements).

Two groups to be compared in terms of changes in the mean response over duration of study.

For simplicity, we assume linear trends and define effect of interest as difference in slopes or rates of change, say δ.

Under null hypothesis of no group difference, i.e., no group × linear trend interaction, H_0: $\delta = 0$.

Sample size calculation can be simplified so that earlier formula can be used.

This is achieved by considering two-stage model described in Lecture 12.

Stage 1: assume a simple parametric curve (e.g., linear) fits the observed responses for each subject.

Stage 2: individual-specific parameters are then related to covariates that describe the two groups.

Stage 1:

$$Y_{ij} = \beta_{1i} + \beta_{2i}\, t_j + \epsilon_{ij},$$

where the errors, ϵ_{ij}, are assumed to be independent and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

Stage 2:

Let $\beta_i = (\beta_{1i}, \beta_{2i})'$.

Allow the mean of β_i (i.e., the mean intercept and slope) to depend on group,

$$\begin{aligned} E(\beta_{1i}) &= \beta_1 + \beta_2\, \mathtt{Group_i} \\ E(\beta_{2i}) &= \beta_3 + \beta_4\, \mathtt{Group_i}. \end{aligned}$$

Note: β_4 is the group difference in the mean slope; $\delta = \beta_4$.

The between-individual variation in the β_i that cannot be explained by group is

$$\text{Cov}(\beta_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = \text{Var}(\beta_{1i})$, $g_{22} = \text{Var}(\beta_{2i})$, and $g_{12} = g_{21} = \text{Cov}(\beta_{1i}, \beta_{2i})$.

Recall: Each subject is measured at common set of occasions, $t_1, ..., t_n$.

Let $\widehat{\beta}_{2i}$ denote the ordinary least squares (OLS) estimate of the slope for the i^{th} subject.

Variability of $\widehat{\beta}_{2i}$, say σ^2, is given by

$$\sigma^2 = \mathrm{Var}(\widehat{\beta}_{2i}) = \sigma^2_\epsilon \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \; g_{22},$$

where

$$\bar{t} = \frac{1}{n} \sum_{j=1}^{n} t_j.$$

To test if mean slopes are equal in two groups, we can construct the following z-test based on the $\widehat{\beta}_{2i}$:

$$Z = \frac{\overline{\beta}_2^{(A)} - \overline{\beta}_2^{(B)}}{\sigma\sqrt{\frac{1}{N} + \frac{1}{N}}} = \frac{\overline{\beta}_2^{(A)} - \overline{\beta}_2^{(B)}}{\sigma\sqrt{\frac{2}{N}}},$$

where $\overline{\beta}_2^{(A)}$ and $\overline{\beta}_2^{(B)}$ are the sample averages of $\widehat{\beta}_{2i}$ in groups A and B, respectively, and $\sigma^2 = \text{Var}(\widehat{\beta}_{2i})$.

Given estimates of g_{22}, the between-subject variability in slopes, and σ_ϵ^2, the within-subject variability, the sample size can be determined from

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\,\sigma^2}{\delta^2},$$

where now

$$\sigma^2 = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} \; + \; g_{22},$$

and δ is group difference in slopes.

Note: This sample size formula is virtually identical to previous formula except σ^2 has two components:
a within-subject variance component, $\sigma_\epsilon^2 \{\sum_{j=1}^{n} (t_j - \bar{t})^2\}^{-1}$, and
a between-subject variance component, $g_{22} = \mathrm{Var}(\beta_{2i})$.

Finally, in a study of length τ, if the n repeated measurements are taken at equally-spaced times $t_1 = 0, t_2 = \tau/(n-1)$, $t_3 = 2\tau/(n-1), ..., t_n = \tau$,

$$\sum_{j=1}^{n} (t_j - \bar{t})^2 = \frac{\tau^2 \, n \, (n+1)}{12 \, (n-1)}.$$

Further examination of this simple formula reveals how sample size (and power) is impacted by:

(i) the length of the study;

(ii) the number of repeated measures; and

(iii) the spacing of the repeated measures.

Note: In general, investigators have little control over the natural heterogeneity of the study population, $g_{22} = \text{Var}(\beta_{2i})$.

Magnitude of σ^2 can be reduced by increasing magnitude of

$$\sum_{j=1}^{n} (t_j - \bar{t})^2.$$

For example, increase length of study or number of repeated measures.

Example

Interested in comparing two treatments (or exposures), denoted A and B.

Plan to randomize an equal number of subjects, say N, to two group.

Plan to take $n = 5$ repeated measurements of the response; 1 at month 0, remainder at 6-month intervals ($\tau = 2$ years).

For simplicity, we assume linear trends and define effect of interest as difference in slopes or rates of change, say δ.

Suppose investigators want to detect minimum $\delta = 1.2$ (e.g., difference in the annual rates of change in the two groups of no less than 1.2).

Based on historical data, investigators posit that between-subject variability in the rate of change, $\text{Var}(\beta_{2i}) \approx 2$ and the within-subject variability, $\sigma^2_\epsilon \approx 7$.

Finally, investigators desire to have power of 90% when conducting a 2-sided test at the 5% significance level (i.e., $\gamma = 0.1$ and $\alpha = 0.05$).

Given these specifications,

$$\sigma^2_\epsilon \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} = \frac{12\,(n-1)\,\sigma^2_\epsilon}{\tau^2\, n\,(n+1)} = \frac{12 \times 4 \times 7}{4 \times 5 \times 6} = 2.8,$$

and

$$\sigma^2 = \sigma^2_\epsilon \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \;\; g_{22} = 2.8 + 2.0 = 4.8.$$

The projected N in each group is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, 2\sigma^2}{\delta^2} = \frac{(1.96 + 1.282)^2 \times 2 \times 4.8}{1.44} = 70.1.$$

Thus, to ensure power of at least 90% investigators will need to enroll a total of 142 subjects, randomizing an equal number (71) to each group.

Note study of same duration ($\tau = 2$ years) with $n = 3$ repeated measurements, 12 months apart, would require a total of 162 subjects to achieve comparable power.

Alternatively, study over 3 instead of 2 years (and with same retention rate), with $n = 5$ repeated measurements taken 9 months apart, would require a total of 96 subjects to achieve power of at least 90%.

Table 44 displays power as a function of N and n.

Table 44: Power as a function of sample size and the number of equally spaced repeated measurements in a longitudinal study of fixed duration.

	Number of Repeated Measures (n)				
Sample Size (N)	2	4	6	8	10
20	0.37	0.39	0.43	0.47	0.50
40	0.63	0.66	0.72	0.76	0.79
60	0.80	0.83	0.87	0.90	0.93
80	0.90	0.92	0.95	0.97	0.98
100	0.95	0.96	0.98	0.99	0.99

Power when conducting a 2-sided test at the 5% significance level ($\alpha = 0.05$) when $\tau = 2$, $\delta = 1.2$, $\text{Var}(\beta_{2i}) = 2$, and $\sigma_e^2 = 7$.

Longitudinal and Cross-Sectional Information

In certain longitudinal designs, we have cohorts that differ in age measured repeatedly over time.

In such designs, it is possible to estimate the effect of growth or aging from two different sources of information: longitudinal and cross-sectional.

It is possible for these two sources of information to provide conflicting estimates of effects.

For example, when effect of aging is determined from cross-sectional information, it is potentially confounded by cohort effects.

Therefore, important to consider models that allow for separate parameters for the longitudinal and cross-sectional effects.

In slight departure from notation, let t_{ij} denote age of i^{th} subject at j^{th} occasion.

In another departure from notation, let X_{ij} denote vector of time-varying covariates and Z_i denote vector of time-stationary covariates.

The following linear model simultaneously models cross-sectional and longitudinal effects:

$$Y_{ij} = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)} + e_{ij}.$$

This representation allows both cross-sectional effects, $\beta^{(C)}$, and longitudinal effects, $\beta^{(L)}$, to be modelled simultaneously.

Interpretation of $\beta^{(C)}$ and $\beta^{(L)}$ becomes more transparent when implied models for initial response and subsequent within-subject changes are considered.

First, consider the model for the initial response, Y_{i1},

$$Y_{i1} = Z_i'\beta_0 + X_{i1}'\beta^{(C)} + e_{i1},$$

since $(X_{i1}' - X_{i1}') = 0$.

$\beta^{(C)}$ represents a vector of regression parameters for cross-sectional effects.

Next, consider the model for within-subject changes from the initial response, $Y_{ij} - Y_{i1}$,

$$\begin{aligned}(Y_{ij} - Y_{i1}) &= Z_i'\beta_0 + X_{i1}'\beta^{(C)} + (X_{ij}' - X_{i1}')\beta^{(L)} + e_{ij} \\ &- (Z_i'\beta_0 + X_{i1}'\beta^{(C)} + e_{i1}) \\ &= (X_{ij}' - X_{i1}')\beta^{(L)} + (e_{ij} - e_{i1}).\end{aligned}$$

$\beta^{(L)}$ represents a vector of regression parameters for longitudinal effects.

Simple Example

$$Y_{ij} = \beta_0 + \beta_1 \text{Gender}_i + \beta^{(C)} \text{Age}_{i1} + \beta^{(L)} (\text{Age}_{ij} - \text{Age}_{i1}) + e_{ij}.$$

First, consider the model for the initial response, Y_{i1},

$$Y_{i1} = \beta_0 + \beta_1 \text{Gender}_i + \beta^{(C)} \text{Age}_{i1} + e_{i1},$$

$\beta^{(C)}$ describes how mean response at baseline changes with age at baseline (cross-sectional change).

Next, consider the model for within-subject changes from the initial response, $Y_{ij} - Y_{i1}$,

$$(Y_{ij} - Y_{i1}) = \beta^{(L)} (\text{Age}_{ij} - \text{Age}_{i1}) + (e_{ij} - e_{i1}).$$

$\beta^{(L)}$ describes how within-subject changes in the response are related to within-subject changes in age (longitudinal change).

Formal comparisons can be made by testing H$_0$: $\beta^{(C)} = \beta^{(L)}$.

Note: When $\beta^{(C)} = \beta^{(L)} = \beta$, the model simplifies to

$$Y_{ij} = Z_i'\beta_0 + X_{ij}'\beta + e_{ij}.$$

In the simple example:

$$Y_{ij} = \beta_0 + \beta_1 \text{Gender}_i + \beta^{(C)} \text{Age}_{i1} + \beta^{(L)} (\text{Age}_{ij} - \text{Age}_{i1}) + e_{ij},$$

the model simplifies to

$$Y_{ij} = \beta_0 + \beta_1 \text{Gender}_i + \beta \text{Age}_{ij} + e_{ij}.$$

However, when $\beta^{(C)} \neq \beta^{(L)}$ but the model does not allow separate estimation of cross-sectional and longitudinal effects,

$$Y_{ij} = Z_i' \beta_0 + X_{ij}' \beta + e_{ij},$$

then β is some weighted combination of $\beta^{(C)}$ and $\beta^{(L)}$ and may not reflect effects of interest.

β confounds the longitudinal effects with the cross-sectional.

Illustration

Suppose three age-cohorts of children, initially aged 5, 6, and 7 years, are measured at baseline and followed annually for three years.

Suppose cross-sectional effect of age on the baseline response is linear, with

$$E(Y_{i1}) = \beta^{(C)}\text{Age}_{i1},$$

(for simplicity, model with intercept=0 is assumed).

Mean response increases linearly with changes in age in each cohort

$$E(Y_{ij} - Y_{i1}) = \beta^{(L)}(\text{Age}_{ij} - \text{Age}_{i1}),$$

but with slope $\beta^{(L)} \neq \beta^{(C)}$.

Figure 32 gives graphical representation of model for mean response versus age, when $\beta^{(C)} = 0.75$ and $\beta^{(L)} = 0.25$.

Note the discernible difference between longitudinal (solid line) and cross-sectional (dotted line) effects of aging.

When these differences are ignored, changes in the mean response (averaged over the three age-cohorts) with age of measurement (dashed line) reflect a combination of $\beta^{(C)}$ and $\beta^{(L)}$.

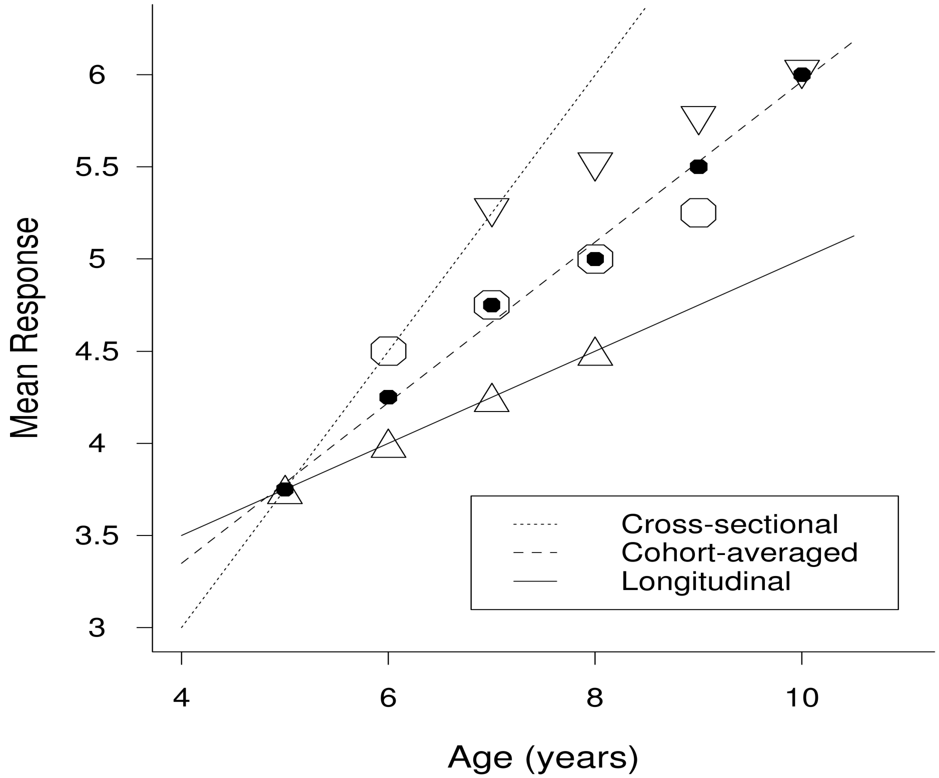


Figure 32: Longitudinal, cross-sectional, and cohort-averaged regression lines for the three age-cohorts: △ denotes mean response of children initially aged 5 years; ○ denotes mean response of children initially aged 6 years; and ∇ denotes mean response of children initially aged 7 years. (● denotes averages over cohorts).

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 16

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Generalized Linear Models for Longitudinal Data

When the response variable is categorical (e.g., binary 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.

In this course we will consider two main extensions of generalized linear models: Marginal Models and Mixed Effects Models.

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

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/Poisson regression models for count data

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

In the remainder of the course, we consider extensions of generalized linear models to longitudinal data.

First, we review logistic and Poisson regression models for a single response.

Review: Logistic Regression

So far, we have considered linear regression models for a continuous response, Y, of the following form

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e$$

The response variable, Y, is assumed to have a normal distribution with mean

$$E(Y) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

and with variance, σ^2.

Recall that the population intercept (for $X_1 = 1$), β_1, has interpretation as the mean value of the response when all of the covariates take on the value zero.

The population slope, say β_k, has interpretation in terms of the expected change in the mean response for a single-unit change in X_k given that all of the other covariates remain constant.

In many studies, however, we are interested in a response variable that is dichotomous/binary rather than continuous.

Next, we consider a regression model for a binary (or dichotomous) response.

Let Y be a binary response, where

$Y = 1$ represents a "success";

$Y = 0$ represent a "failure".

Then the mean of the binary response variable, denoted π, is the *proportion* of successes or the probability that the response takes on the value 1.

That is,

$$\pi = E(Y) = \Pr(Y = 1) = \Pr(\text{“success”})$$

With a binary response, we are usually interested in estimating the probability π, and relating it to a set of covariates.

To do this, we can use *logistic regression.*

A naive strategy for modeling a binary response is to consider a linear regression model

$$\pi = E(Y) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

However, in general, this model is not feasible since π is a probability and is restricted to values between 0 and 1.

Also, the usual assumption of homogeneity of variance would be violated since the variance of a binary response depends on the mean, i.e.

$$\mathrm{Var}(Y) = \pi\,(1 - \pi)$$

Instead, we can consider a logistic regression model where

$$\ln\left[\pi/\left(1-\pi\right)\right] = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

This model accommodates the constraint that π is restricted to values between 0 and 1.

Recall that $\pi/\left(1-\pi\right)$ is defined as the odds of success.

Therefore, modeling π with a logistic function can be considered equivalent to a linear regression model where the mean of the continuous response has been replaced by the logarithm of the odds of success.

Note that the relationship between π and the covariates is non-linear.

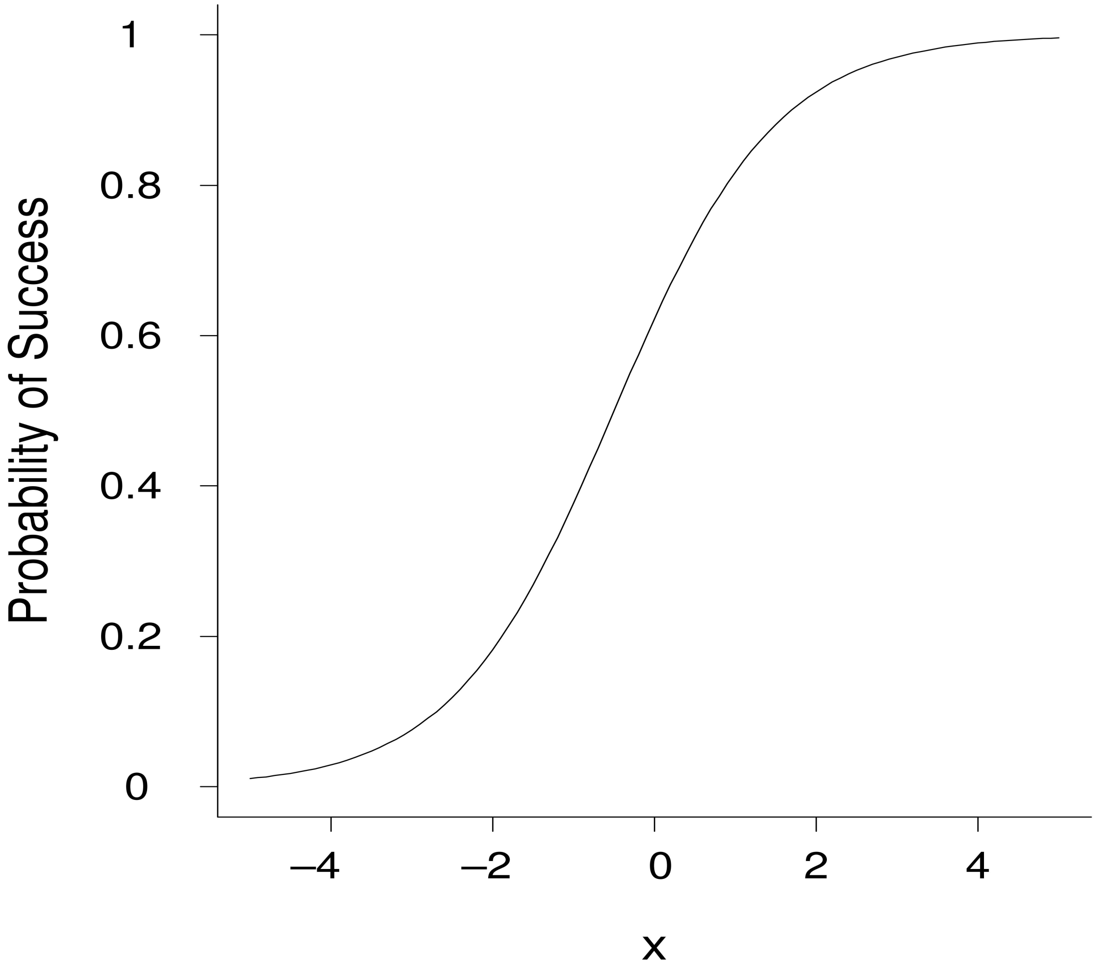


Figure 33: Plot of logistic response function.

Under the assumption that the binary responses are *Bernoulli* random variables, we can use ML estimation to obtain estimates of the logistic regression parameters.

Finally, recall the relationship between “odds” and “probabilities”.

$$\text{Odds} = \frac{\pi}{1 - \pi};$$

$$\pi = \frac{\text{Odds}}{1 + \text{Odds}}.$$

Given the logistic regression model

$$\ln\left[\pi/\left(1-\pi\right)\right] = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

the population intercept, β_1, has interpretation as the log odds of success when all of the covariates take on the value zero.

The population slope, say β_k, has interpretation in terms of the change in log odds of success for a single-unit change in X_k given that all of the other covariates remain constant.

When one of the covariates is dichotomous, say X_2, then β_2 has a special interpretation:

$\exp\left(\beta_2\right)$ is the *odds ratio* or ratio of odds of success for the two possible levels of X_2 (given that all of the other covariates remain constant).

Keep in mind that as:

π increases

$\Rightarrow$ odds of success increases

$\Rightarrow$ log odds of success increases

Similarly, as:

π decreases

$\Rightarrow$ odds of success decreases

$\Rightarrow$ log odds of success decreases

Example: Development of bronchopulmonary dysplasia (BPD) in a sample of 223 low birth weight infants.

Binary Response: $Y = 1$ if BPD is present, $Y = 0$ otherwise.

Covariate: Birth weight of infant in grams.

Consider the following logistic regression model

$$\ln\left[\pi/\left(1-\pi\right)\right] = \beta_1 + \beta_2 \text{Weight}$$

where $\pi = E(Y) = \Pr(Y = 1) = \Pr(\text{BPD})$

For the 223 infants in the sample, the estimated logistic regression (obtained using ML) is

$$\ln\left[\widehat{\pi}/\left(1-\widehat{\pi}\right)\right]=4.0343-0.0042 \text{ Weight}$$

The ML estimate of β_2 implies that, for every 1 gram increase in birth weight, the log odds of BPD decreases by 0.0042.

For example, the odds of BPD for an infant weighing 1200 grams is

$$\exp\left(4.0343-1200*.0042\right)=\exp\left(-1.0057\right)$$

$$=0.3658$$

Thus the predicted probability of BPD is:

$$0.3658/\left(1+0.3658\right)=0.268$$

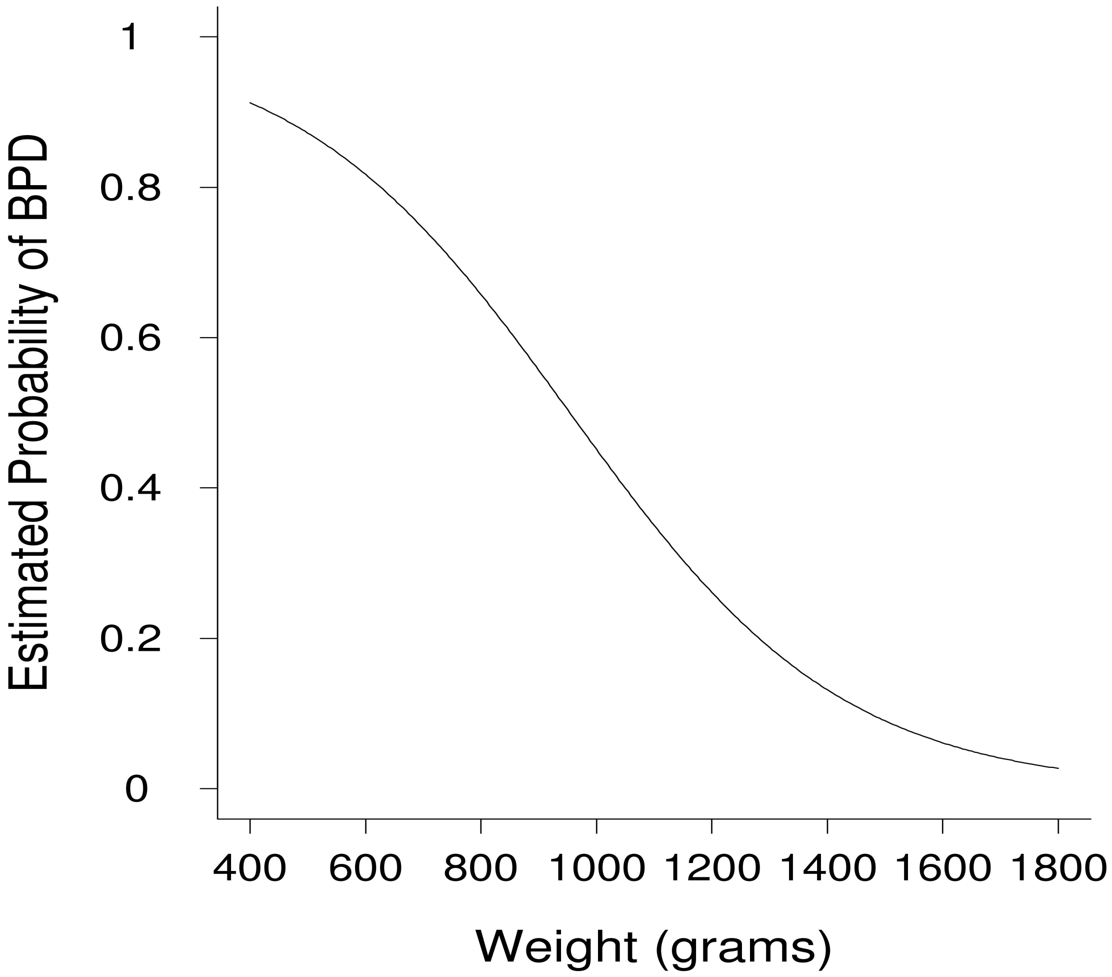


Figure 34: Plot of estimated logistic response function of BPD on birth weight.

Review: Poisson Regression

In Poisson regression, the response variable is a count (e.g. number of cases of a disease in a given period of time).

The Poisson distribution provides the basis of likelihood-based inference.

Often the counts may be expressed as *rates*.

That is, the count or absolute number of events is often not satisfactory because any comparison depends almost entirely on the sizes of the groups (or the "time at risk") that generated the observations.

Like a proportion or probability, a rate provides a basis for direct comparison.

In either case, Poisson regression relates the expected counts or rates to a set of covariates.

The Poisson regression model has two components:

1. The response variable is a count and is assumed to have a Poisson distribution.
   That is, the probability a specific number of events, y, occurs is

   $$\Pr(y \text{ events}) = e^{-\lambda}\lambda^{y}/y!$$

   Note that λ is the expected count or number of events and the expected rate is given by λ/t, where t is a relevant baseline measure (e.g., t might be the number of persons or the number of person-years of observation).

2. $\ln(\lambda/t) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

   Note that since $\ln(\lambda/t) = \ln(\lambda) - \ln(t)$, the Poisson regression model can also be considered as

$$\ln(\lambda) = \ln(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

   where the 'coefficient' associated with $\ln(t)$ is fixed to be 1.

   This adjustment term is known as an "offset".

Therefore, modelling λ (or λ/t) with a log function can be considered equivalent to a linear regression model where the mean of the continuous response has been replaced by the logarithm of the expected count (or rate).

Note that the relationship between λ (or λ/t) and the covariates is non-linear.

We can use ML estimation to obtain estimates of the Poisson regression parameters, under the assumption that the responses are *Poisson* random variables.

Given the Poisson regression model

$$\ln(\lambda/t) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

the population intercept, β_1, has interpretation as the log expected rate when all the covariates take on the value zero.

The population slope, say β_k, has interpretation in terms of the change in log expected rate for a single-unit change in X_k given that all of the other covariates remain constant.

When one of the covariates is dichotomous, say X_2, then β_2 has a special interpretation:

$\exp(\beta_2)$ is the (incidence) rate ratio for the two possible levels of X_2 (given that all of the other covariates remain constant).

Example: Prospective study of coronary heart disease (CHD).

The study observed 3154 men aged 40-50 for an average of 8 years and recorded incidence of cases of CHD.

The risk factors considered include:

Smoking exposure: 0, 10, 20, 30 cigs per day;
Blood Pressure: 0 (< 140), 1 (≥ 140);
Behavior Type: 0 (type B), 1 (type A).

A simple Poisson regression model is:

$$\ln(\lambda/t) = \ln(\textit{rate of CHD}) = \beta_1 + \beta_2 \, \boldsymbol{Smoke}$$

or

$$\ln(\lambda) = \ln(t) + \beta_1 + \beta_2 \, \boldsymbol{Smoke}$$

Person - Years	Smoking	Blood Pressure	Behavior	CHD
5268.2	0	0	0	20
2542.0	10	0	0	16
1140.7	20	0	0	13
614.6	30	0	0	3
4451.1	0	0	1	41
2243.5	10	0	1	24
1153.6	20	0	1	27
925.0	30	0	1	17
1366.8	0	1	0	8
497.0	10	1	0	9
238.1	20	1	0	3
146.3	30	1	0	7
1251.9	0	1	1	29
640.0	10	1	1	21
374.5	20	1	1	7
338.2	30	1	1	12

In this model the ML estimate of β_2 is 0.0318. That is, the rate of CHD increases by a factor of $\exp(0.0318) = 1.032$ for every cigarette smoked.

Alternatively, the rate of CHD in smokers of one pack per day (20 cigs) is estimated to be $(1.032)^{20} = 1.88$ times higher than the rate of CHD in non-smokers.

We can include the additional risk factors in the following model:

$$\ln(\lambda/t) = \beta_1 + \beta_2 \text{ Smoke} + \beta_3 \text{ Type} + \beta_4 \text{BP}$$

Effect	**Estimate**	**Std. Error**
Intercept	-5.420	0.130
Smoke	0.027	0.006
Type	0.753	0.136
BP	0.753	0.129

Now, adjusted rate of CHD (controlling for BP and behavior type) increases by a factor of exp(0.027) = 1.028 for every cigarette smoked.

Adjusted rate of CHD in smokers of one pack per day (20 cigs) is estimated to be $(1.027)^{20} = 1.7$ times higher than rate of CHD in non-smokers.

Finally, note that when a Poisson regression model is applied to data consisting of very small rates (say, $\lambda/t << 0.01$), then the rate is approximately equal to the corresponding probability, p, and

$$\ln(\text{rate}) \approx \ln(p) \approx \ln\left[p/(1-p)\right]$$

Therefore, the parameters for Poisson regression and logistic regression models are approximately equal when the event being studied is rare.

In that case, results from a Poisson and logistic regression will not give discernibly different results.

Overdispersion

Count data (or counts of number of successes) often have variability that far exceeds that predicted by Poisson (or binomial) distribution.

This phenomenon is referred to as *overdispersion*.

Although underdispersion can also arise, it is far less common.

Failure to account for overdispersion has negligible impact of the estimated regression coefficients.

Neglecting overdispersion results in standard errors being underestimated and potentially misleading inferences (e.g., confidence intervals that are too narrow and p-values that are too small).

Example: *Clinical Trial of Antibiotics for Leprosy*

Placebo-controlled clinical trial of 30 patients with leprosy at the Eversley Childs Sanitorium in the Philippines.

Participants were randomized to either of two antibiotics (denoted treatment drug A and B) or to a placebo (denoted treatment drug C).

Baseline data on number of leprosy bacilli at 6 sites of body were recorded.

After several months of treatment, number of bacilli were recorded a second time.

Outcome: Total count of number of leprosy bacilli at 6 sites.

Table 45: Mean count of leprosy bacilli at six sites of the body (and variance) post-treatment.

Treatment Group	Post-Treatment
Drug A (Antibiotic)	5.3
	(21.6)
Drug B (Antibiotic)	6.1
	(37.9)
Drug C (Placebo)	12.3
	(51.1)

Consider outcome (post-treatment) at end of study.

Variability is approximately 4 to 6 times larger than that predicted by Poisson variation.

Adjustments to nominal standard errors to account for overdispersion can be made either by including a scale factor ϕ in specification of the Poisson variance,

$$\mathrm{Var}(Y_i) = \phi\,\mu_i,$$

or by basing standard errors on the so-called “sandwich” estimator of $\mathrm{Cov}(\widehat{\beta})$.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 17

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Introduction to 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/Poisson regression models for count data

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

Later, 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 is a $p \times 1$ vector of 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:

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

where link function $g(\cdot)$ is a known function, e.g., $\log(\mu_i)$.

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

$$\text{Var}\,(Y_i) = \phi\, v(\mu_i),$$

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.

Link Function

The link function applies a transformation to the mean and then links the covariates to the transformed mean,

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

where link function $g(\cdot)$ is known function, e.g., $\log(\mu_i)$.

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

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

Common Examples

Normal distribution:

If we assume that $g(\cdot)$ is the identity function,

$$g\left(\mu\right) = \mu$$

then

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

gives the standard linear regression model, with $\text{Var}\left(Y_i\right) = \phi$.

Note: Variance is unrelated to the mean.

Bernoulli distribution:

For the Bernoulli distribution, $0 < \mu_i < 1$, so we would prefer a link function that transforms the interval $[0, 1]$ on to the entire real line $(-\infty, \infty)$:

$$\begin{aligned} \text{logit} : \ln\left[\mu_i / \left(1 - \mu_i\right)\right] \\ \text{probit} : \Phi^{-1}\left(\mu_i\right) \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

If we assume a logit link function then

$$\log\left[\frac{\mu_i}{(1 - \mu_i)}\right] = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

yields logistic regression model, with $\text{Var}\left(Y_i\right) = \mu_i(1 - \mu_i)$ (Bernoulli variance).

Poisson distribution:

For the Poisson distribution, $\mu_i > 0$, so we would prefer a link function that transforms the interval $(0, \infty)$ on to the entire real line $(-\infty, \infty)$.

If we assume a log link function then

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

yields Poisson or loglinear regression model, with $\mathrm{Var}(Y_i) = \mu_i$ (Poisson variance).

Summary

In generalized linear models:

(i) response assumed to have exponential family distribution, e.g., normal, Bernoulli, binomial, and Poisson distributions.

(ii) transformed mean response is linearly related to the covariates, via an appropriate link function:

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

PROC GENMOD in SAS

Table 46: Illustrative commands for logistic regression using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;
   CLASS  group;
   MODEL  y=group / DIST=BINOMIAL  LINK=LOGIT;
```

PROC GENMOD in SAS

Table 47: Illustrative commands for log-linear regression, with an offset, using PROC GENMOD in SAS.

```
PROC GENMOD;

   CLASS  group;

   MODEL  y=group / DIST=POISSON  LINK=LOG  OFFSET=logtime;
```

Extensions of Generalized Linear Models to Longitudinal Data

When the response variable is categorical (e.g., binary 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.

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)

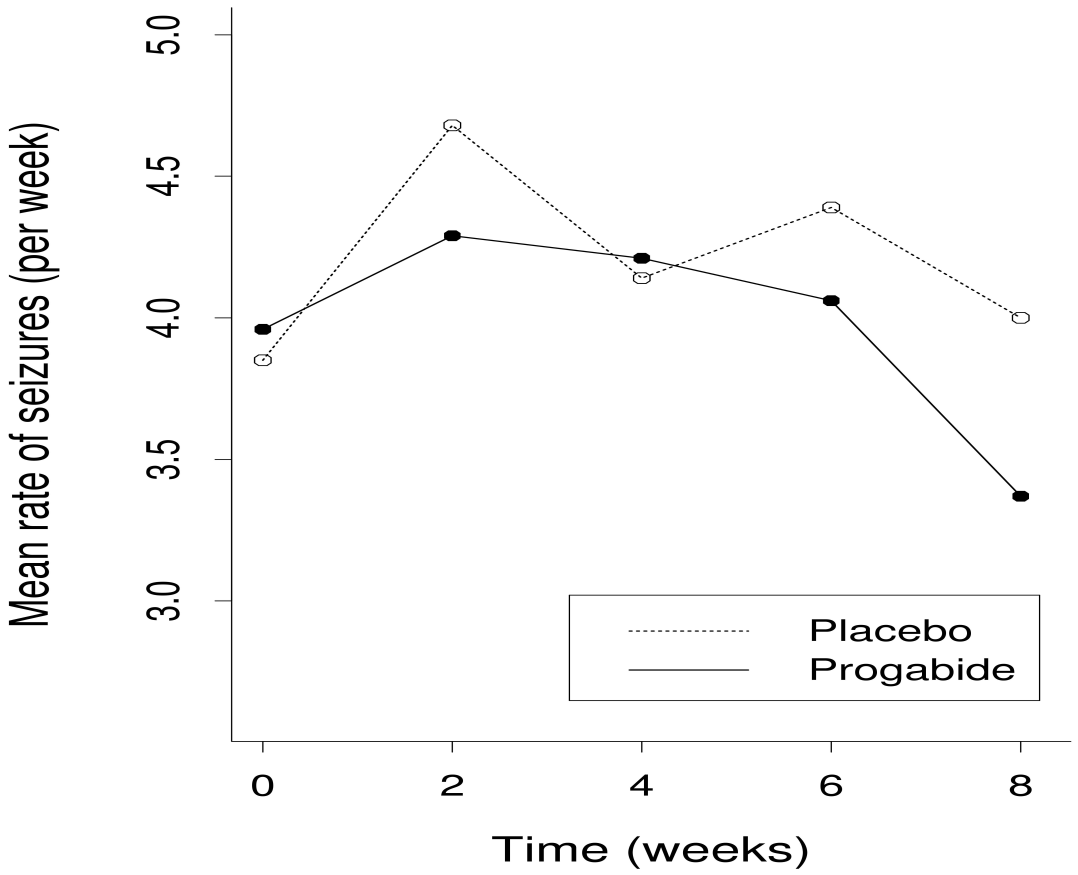
Randomized, placebo-controlled study of treatment of epileptic seizures with progabide.

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


Mean rate of seizures (per week)
5.0
4.5
4.0
3.5
3.0
0
2
4
6
8
Time (weeks)
Placebo
Progabide

Generalized Linear Models for Longitudinal Data

Next, we focus on a number of distinct approaches for analyzing longitudinal responses.

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.

Note: In linear (mixed effects) models for continuous responses, the interpretation of the regression coefficients is independent of the correlation among the responses.

With discrete response data, this is no longer the case.

With non-linear models for discrete data, different approaches for accounting for the correlation leads to models having regression coefficients with distinct interpretations.

We will return to this important issue later (Lecture 20).

In the remainder of this lecture, we will briefly survey three main extensions of generalized linear models.

Suppose that $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$ is a vector of correlated responses from the i^{th} subject.

To analyze such correlated data, we must specify, or at least make assumptions about, the multivariate or joint distribution,

$$f\left(Y_{i1}, Y_{i2}, \ldots, Y_{in}\right)$$

The way in which the multivariate distribution is specified yields three somewhat different analytic approaches:

1. Marginal Models

2. Mixed Effects Models

3. Transitional Models

Marginal Models

One approach is to specify the marginal distribution at each time point:

$$f\left(Y_{ij}\right) \text{ for } j = 1, 2, \ldots, n$$

along with some assumptions about the covariance structure of the observations.

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).

Illustration

Consider the *Oral Treatment of Toenail Infection* study.

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

$$\text{logit}\{Pr(Y_{ij} = 1)\} = \beta_1 + \beta_2\text{Month}_{ij} + \beta_3\text{Trt}_i * \text{Month}_{ij}$$

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

This is an example of a marginal model.

Note, however, that the covariance structure remains to be specified.

Mixed Effects Models

Another possibility is to assume that a subset of the regression parameters in the generalized linear model vary from subject to subject.

Specifically, we could assume that the data for a single subject are independent observations from a distribution belonging to the exponential family, but that the regression coefficients can vary from person to person.

That is, conditional on the random effects, it is assumed that the responses for a single subject are independent observations from a distribution belonging to the exponential family.

Illustration

Consider the *Oral Treatment of Toenail Infection* study.

Suppose, for example, that the probability of onycholysis for participants in the study is described by a logistic model, but that the risk for an individual depends on her latent (perhaps environmentally and genetically determined) “random response level”.

Then we might consider a model where

$$\text{logit}\{Pr(Y_{ij} = 1|b_i)\} = \beta_1 + \beta_2 \text{Month}_{ij} + \beta_3 \text{Trt}_i * \text{Month}_{ij} + b_i$$

Note that such a model also requires specification of the random effects distribution, $F\left(b_i\right)$.

This is an example of a generalized linear mixed effects model.

Transitional (Markov) Models

Finally, another approach is to express the joint distribution as a series of conditional distributions,

$$f\left(Y_{i1}, Y_{i2}, \ldots, Y_{in}\right) = f\left(Y_{i1}\right) f\left(Y_{i2}|Y_{i1}\right) \cdots f\left(Y_{in}|Y_{i1}, \ldots, Y_{i,n-1}\right)$$

This is known as a transitional model (or a model for the transitions) because it represents the probability distribution at each time point as conditional on the past.

This provides a complete representation of the joint distribution.

Illustration

Consider the *Oral Treatment of Toenail Infection* study.

We could write the probability model as

$$f\left(Y_{i1}|X_i\right) f\left(Y_{i2}|Y_{i1}, X_i\right) f\left(Y_{i3}|Y_{i1}, Y_{i2}, X_i\right) \cdots f\left(Y_{i7}|Y_{i1}, Y_{i2}, ..., Y_{i6}, X_i\right)$$

That is, the probability of onycholysis at time 2 is modeled conditional on presence/absence of onycholysis at time 1, and so on.

For example, a "1st-order" logistic model allowing dependence only on previous response, is given by

$$\text{logit}\{Pr(Y_{ij} = 1|Y_{i,j-1})\} = \beta_1 + \beta_2 \text{Month}_{ij} + \beta_3 \text{Trt}_i * \text{Month}_{ij} + \beta_4 Y_{i,j-1}$$

Summary

We have discussed the main features of generalized linear models

We have briefly outlined three main extensions of generalized linear models to longitudinal data:

1. Marginal Models

2. Mixed Effects Models

3. Transitional Models

In the remainder of the course we focus on (i) Marginal Models, and (ii) Mixed Effects Models.

In general, transitional models are somewhat less useful for modelling covariate effects.

Specifically, inferences from a transitional model can be potentially misleading if a treatment or exposure changes risk throughout the follow-up period.

In that case, the conditional risk, given previous history of the outcome, is altered somewhat less strikingly.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 18

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Marginal Models and Generalized Estimating Equations

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).

Features of Marginal Models:

The focus of marginal models is on inferences about population averages.

The marginal expectation, $\mu_{ij} = E\left(Y_{ij}|X_{ij}\right)$, of each response is modelled as a function of covariates.

Specifically, marginal models have the following three part specification:

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

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

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

$$\mathrm{Var}\left(Y_{ij}|X_{ij}\right) = \phi\, v\left(\mu_{ij}\right)$$

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

3. The 'within-subject association' among the responses is a function of the means and of additional parameters, say α, 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:

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

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

In principle, can also specify higher-order moments.

Aside: Measures of Association for Binary Responses

With binary responses correlations are not the best choice for modelling the association because they are constrained by the marginal probabilities.

For example, if $E(Y_1) = Pr(Y_1 = 1) = 0.2$ and $E(Y_2) = Pr(Y_2 = 1) = 0.8$, then $\text{Corr}(Y_1, Y_2) < 0.25$.

The correlations must satisfy certain linear inequalities determined by the marginal probabilities.

These constraints are likely to cause difficulties for parametric modelling of the association.

With binary responses, the odds ratio is a natural measure of association between a pair of responses.

The odds ratio for any pair of binary responses, Y_j and Y_k, is defined as

$$OR(Y_j, Y_k) = \frac{Pr(Y_j = 1, Y_k = 1) Pr(Y_j = 0, Y_k = 0)}{Pr(Y_j = 1, Y_k = 0) Pr(Y_j = 0, Y_k = 1)}.$$

Note that the constraints on the odds ratio are far less restrictive than on the correlation.

$\Longrightarrow$ With binary response can model within-subject association in terms of odds ratios rather than correlations.

Examples of Marginal Models

Example 1. Continuous responses:

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

2. $\mathrm{Var}\,(Y_{ij}|X_{ij}) = \phi_j$
   (i.e., heterogeneous variance, but no dependence of variance on mean)

3. $\mathrm{Corr}\,(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$ $(0 \leq \alpha \leq 1)$
   (i.e., autoregressive correlation)

Example 2. Binary responses:

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

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

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

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

Example 3. Count data:

1. $\text{Log}\,(\mu_{ij}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$.
   (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. $\text{Corr}\,(Y_{ij}, Y_{ik}) = \alpha$
   (i.e., compound symmetry correlation)

Interpretation of Marginal Model Parameters

The regression parameters, β, have ‘population-averaged’ interpretations (where ‘averaging’ is over all individuals within subgroups of the population):

- describe effect of covariates on the average responses
- contrast the means in sub-populations that share common covariate values

$\Longrightarrow$ Marginal models are most useful for population-level inferences.

The regression parameters are directly estimable from the data.

Of note, nature or magnitude of within-subject association (e.g., correlation) does not alter the interpretation of β.

For example, consider the following logistic model,

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

Each element of β measures the change in the log odds of a 'positive' response per unit change in the respective covariate, for sub-populations defined by fixed and known covariate values.

The interpretation of any component of β, say β_k, is in terms of 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,

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

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

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

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

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

$\longrightarrow$ β_k is change in log odds for subgroups of the study population (defined by any fixed values of $X_{ij1}, \ldots, X_{ij(k-1)}, X_{ij(k+1)}, \ldots, X_{ijp}$).

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

Avoid making distributional assumptions about Y_i altogether.

Potential Advantages:

Empirical researcher does not have to be concerned that the distribution of Y_i closely approximates some multivariate distribution.

It circumvents the need to specify models for the three-way, four-way and higher-way associations (higher-order moments) among the responses.

It leads to a method of estimation, known as generalized estimating equations (GEE), that is straightforward to implement.

The GEE approach has become an extremely popular method for analyzing discrete longitudinal data.

It provides a flexible approach for modelling the mean and the pairwise within-subject association structure.

It can handle inherently unbalanced designs and missing data with ease (albeit making strong assumptions about missingness).

GEE approach is computationally straightforward and has been implemented in existing, widely-available statistical software.

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

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

where 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 Y_i.

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

$D_i = \partial\mu_i / \partial\beta$ is the "derivative" matrix (of μ_i with respect to the components of β); $D_i(\beta)$ transforms from the original units of Y_{ij} (and μ_{ij}) to the units of $g(\mu_{ij})$.

Therefore the generalized estimating equations depend on both β and α.

Because the generalized estimating equations depend on both β and α, an iterative two-stage estimation procedure is required:

1. Given current estimates of α and ϕ, an estimate of β is obtained as the solution to the 'generalized estimating equations'

2. Given current estimate of β, estimates of $\boldsymbol{\alpha}$ and ϕ are obtained based on the standardized residuals,

$$r_{ij} = \left(Y_{ij} - \widehat{\mu}_{ij}\right) / v\left(\widehat{\mu}_{ij}\right)^{1/2}$$

For example, ϕ can be estimated by

$$1/\left(Nn - p\right) \sum_{i=1}^{N} \sum_{j=1}^{n} r_{ij}^2$$

The correlation parameters, α, can be estimated in a similar way. For example, unstructured correlations, $\alpha_{jk} = \text{Corr}\left(Y_{ij}, Y_{ik}\right)$, can be estimated by

$$\widehat{\alpha}_{jk} = \left(1/(N-p)\right) \widehat{\phi}^{-1} \sum_{i=1}^{N} r_{ij} r_{ik}$$

Finally, in the two-stage estimation procedure we iterate between steps 1) and 2) until convergence has been achieved.

Properties of GEE estimators

$\widehat{\beta}$, the solution to the generalized estimating equations, has the following properties:

1. $\widehat{\beta}$ is consistent estimator of β

2. In large samples, $\widehat{\beta}$ has a multivariate normal distribution

3. $\text{Cov}(\widehat{\beta}) = B^{-1}MB^{-1}$
   where

$$B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i$$

$$M = \sum_{i=1}^{N} D_i' V_i^{-1} \text{Cov}\left(Y_i\right) V_i^{-1} D_i$$

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

Note: We can use this empirical or so-called 'sandwich' variance estimator even when the covariance has been misspecified.

Summary

The GEE estimators have the following attractive properties:

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 the empirical or so-called 'sandwich' estimator

Case Study 1: *Clinical Trial of Antibiotics for Leprosy*

Placebo-controlled clinical trial of 30 patients with leprosy at the Eversley Childs Sanitorium in the Philippines.

Participants were randomized to either of two antibiotics (denoted treatment drug A and B) or to a placebo (denoted treatment drug C).

Baseline data on number of leprosy bacilli at 6 sites of body were recorded.

After several months of treatment, number of bacilli were recorded a second time.

Outcome: Total count of number of leprosy bacilli at 6 sites.

Table 48: Mean count of leprosy bacilli at six sites of the body (and variance) pre- and post-treatment.

Treatment Group	Baseline	Post-Treatment
Drug A (Antibiotic)	9.3	5.3
	(22.7)	(21.6)
Drug B (Antibiotic)	10.0	6.1
	(27.6)	(37.9)
Drug C (Placebo)	12.9	12.3
	(15.7)	(51.1)

Question: Does treatment with antibiotics (drugs A and B) reduce abundance of leprosy bacilli when compared to placebo (drug C).

We consider the following model for changes in the average count

$$\log E(Y_{ij}) \;=\; \log \mu_{ij} = \beta_1 + \beta_2\,\texttt{time}_{\texttt{ij}} + \beta_3\,\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{1i}} + \beta_4\,\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{2i}},$$

where Y_{ij} is count of bacilli for i^{th} patient in j^{th} period ($j = 1, 2$).

$\texttt{trt}_1$ and $\texttt{trt}_2$ are indicator variables for drugs A and B respectively.

The binary variable, `time`, denotes the baseline and post-treatment follow-up periods, with $\texttt{time} = 0$ for the baseline period (period 1) and $\texttt{time} = 1$ for the post-treatment follow-up period (period 2).

To complete specification of the marginal model, we assume

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

where ϕ can be thought of as an overdispersion factor.

Finally, the within-subject association is accounted for by assuming a common correlation,

$$\text{Corr}(Y_{i1}, Y_{i2}) = \alpha.$$

The log-linear regression parameters, β, can be given interpretations in terms of (log) rate ratios.

Table 49: Parameters of the marginal log-linear regression model for the leprosy bacilli data.

Treatment Group	Period	$\log(\mu_{ij})$
Drug A (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_3$
Drug B (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_4$
Drug C (Placebo)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2$

Table 49 summarizes their interpretation in terms of the log expected counts in the three groups at baseline and during post-treatment follow-up.

For example, e^{β_2} is the rate ratio of leprosy bacilli, comparing the follow-up period to baseline, in the placebo group (drug C).

Similarly, $e^{\beta_2+\beta_3}$ is the corresponding rate ratio in the group randomized to drug A.

Finally, $e^{\beta_2+\beta_4}$ is the corresponding rate ratio in the group randomized to drug B.

Thus, β_3 and β_4 represents the difference between the changes in the log expected rates, comparing drug A and B to the placebo (drug C).

Estimated regression coefficients are displayed in Table 50 (with SEs based on "sandwich" estimator).

A test of H$_0$: $\beta_3 = \beta_4 = 0$, produces a (multivariate) Wald statistic, $W^2 = 6.99$, with 2 degrees of freedom ($p < 0.05$).

Note: Magnitudes of effects are similar and indicate that treatment with antibiotics reduces leprosy bacilli.

A test of H$_0$: $\beta_3 = \beta_4$, produces a Wald statistic, $W^2 = 0.08$, with 1 degree of freedom ($p > 0.7$).

Table 50: Parameter estimates and standard errors from marginal log-linear regression model for the leprosy bacilli data.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
$\texttt{time}_{\texttt{ij}}$	−0.0138	0.1573	−0.09
$\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{1i}}$	−0.5406	0.2186	−2.47
$\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{2i}}$	−0.4791	0.2279	−2.10

Estimated scale or dispersion parameter: $\widehat{\phi} = 3.45$.
Estimated working correlation: $\widehat{\alpha} = 0.797$.

To obtain a common estimate of the log rate ratio, comparing both antibiotics (drugs A and B) to placebo, we can fit the reduced model

$$\log E(Y_{ij}) \;=\; \log \mu_{ij} = \beta_1 + \beta_2\,\texttt{time}_{\texttt{ij}} + \beta_3\,\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{i}},$$

where the variable `trt` is an indicator variable for antibiotics, with $\texttt{trt} = 1$ if a patient was randomized to either drug A or B and $\texttt{trt} = 0$ otherwise.

We retain the same assumptions about the variance and correlation as before.

The estimated regression coefficients are displayed in Table 51

Table 51: Parameter estimates and standard errors from marginal log-linear regression model for the leprosy bacilli data.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
$\texttt{time}_{\texttt{ij}}$	−0.0108	0.1572	−0.07
$\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{i}}$	−0.5141	0.1966	−2.62

Estimated scale or dispersion parameter: $\widehat{\phi} = 3.41$.
Estimated working correlation: $\widehat{\alpha} = 0.780$.

The common estimate of the log rate ratio is -0.5141.

Rate ratio is 0.60 (or $e^{-0.5141}$), with 95% confidence interval, 0.41 to 0.88, indicating that treatment with antibiotics significantly reduces the average number of bacilli when compared to placebo.

For placebo group, there is a non-significant reduction in the average number of bacilli of approximately 1% (or $[1 - e^{-0.0108}] \times 100\%$).

In the antibiotics group there is a significant reduction of approximately 40% (or $[1 - e^{-0.0108-0.5141}] \times 100\%$).

Estimated pairwise correlation of 0.8 is relatively large.

Estimated scale parameter of 3.4 indicates substantial overdispersion.

Case Study 2: *Oral Treatment of Toenail Infection*

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

$$\text{logit}E(Y_{ij}) = \beta_1 + \beta_2\text{Month}_{ij} + \beta_3\text{Trt}_i * \text{Month}_{ij}$$

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

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

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

Table 52: 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 infection has decreases by a factor of 0.19 [exp(-0.14*12)] in treatment group A.

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

4. Odds ratio comparing 12 month decreases in risk of infection 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.

Summary of Key Points

The focus of marginal models is on inferences about population averages.

The regression parameters, β, have 'population-averaged' interpretations (where 'averaging' is over all individuals within subgroups of the population):

- describe effect of covariates on marginal expectations or average responses
- contrast means in sub-populations that share common covariate values

$\Longrightarrow$ Marginal models are most useful for population-level inferences.

Marginal models should not be used to make inferences about individuals ("ecological fallacy").

GEE using PROC GENMOD in SAS

PROC GENMOD in SAS is primarily a procedure for fitting generalized linear models to a single response.

However, PROC GENMOD has incorporated an option for implementing GEE approach using a REPEATED statement (similar to PROC MIXED).

PROC GENMOD, as with almost all software for longitudinal analyses, requires each repeated measurement in a longitudinal data set to be a separate "record".

If the data set is in a *multivariate* mode (or "wide format"), it must be transformed to a *univariate* mode (or "long format") prior to analysis.

Table 53: Illustrative commands for a marginal logistic regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;

   CLASS id group;

   MODEL y=group time group*time / DIST=BINOMIAL LINK=LOGIT;

   REPEATED SUBJECT=id / WITHINSUBJECT=time TYPE=UN;
```

Table 54: Illustrative commands for a marginal logistic regression, with within-subject associations specified in terms of log odds ratios, using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;

   CLASS id group;

   MODEL y=group time group*time / DIST=BINOMIAL LINK=LOGIT;

   REPEATED SUBJECT=id / WITHINSUBJECT=time LOGOR=FULLCLUST;
```

Table 55: Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS.

```
PROC GENMOD;
   CLASS  id  group;
   MODEL  y=group  time  group*time / DIST=POISSON  LINK=LOG;
   REPEATED  SUBJECT=id / WITHINSUBJECT=time  TYPE=UN;
```

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 19

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Generalized Linear Mixed Models

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 (Lectures 11-13).

We also highlight their greater degree of conceptual and analytic complexity relative to marginal models.

Generalized Linear Mixed Models

Postulate unobserved latent variables (random effects) shared by the repeated measures on the same subject.

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

That is, a subset of the regression coefficients (e.g., intercepts and slopes) are assumed to vary across individuals according to some distribution.

Then, conditional on the random effects, it is assumed that the responses for a single individual are independent observations from a distribution belonging to the exponential family.

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,

$$g(E[Y_{ij}|b_i]) = X'_{ij}\beta + Z'_{ij}b_i$$

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 $b_i \sim N\left(0, G\right)$.

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:

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

$Var(Y_{ij}|b_i) = E[Y_{ij}|b_i](1 - E[Y_{ij}|b_i])$ (Bernoulli variance),

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

Example 2:

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

$$\log(E[Y_{ij}|b_i]) = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij}$$

$Var(Y_{ij}|b_i) = E[Y_{ij}|b_i]$ (Poisson variance),

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

Note: G is the covariance matrix for b_{1i} and b_{2i}.

Recall that marginal models consider the consequences of dependence among the repeated measures on the same subject, via a “working” covariance.

In contrast, GLMMs provide a potential explanation for the sources of dependence among the repeated measures on the same subject, via the introduction of random effects.

However, the introduction of random effects also has important implications for the interpretation of the regression parameters in GLMMs.

Interpretation of Fixed Effects

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

Main focus is on the individual and the influence of covariates on a *typical* $(b_i = 0)$ individual's responses.

Regression parameters, $\boldsymbol{\beta}$, measure the change in expected value of response while holding constant other covariates and the random effects.

For example, consider the following logistic model,

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

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

Each element of $\boldsymbol{\beta}$ measures the change in the log odds of a 'positive' response per unit change in the respective covariate, for an individual with propensity to respond positively, b_i.

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

Note: This is not always directly observable from the data.

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

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

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

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

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

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

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

This *subject-specific* interpretation of β_k is more appealing when X_{ijk} is a *time-varying* covariate.

That is, when it is possible to hold b_i (and remaining covariates) fixed and also change the value of the covariate, X_{ijk}.

Recall: Time-varying covariate is one whose value can change over time, e.g., time since baseline, smoking status, and environmental exposures.

When X_{ijk} is *time-invariant* the interpretation of β_k is less transparent.

With a time-invariant covariate (e.g., gender), changing the value of the covariate requires also a change in the index i of X_{ijk}, say $X_{i'jk}$.

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

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

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

Similarly, when $X_{i'jk}$ now takes on some value $x+1$,

$$\log\left[\frac{\Pr(Y_{i'j}=1|b_{i'},X_{i'j1},\ldots,X_{i'jk}=x+1,\ldots,X_{i'jp})}{\Pr(Y_{i'j}=0|b_{i'},X_{i'j1},\ldots,X_{i'jk}=x+1,\ldots,X_{i'jp})}\right] \quad =$$

$$b_{i'} + \beta_1 X_{i'j1} + \cdots + \beta_k (x+1) + \cdots + \beta_p X_{i'jp}.$$

Even when we consider two subjects with identical covariates except for the k^{th}, the difference in log odds is

$$\beta_k + (b_i - b_{i'}).$$

That is, β_k has become confounded with $b_i - b_{i'}$.

This dilemma can only be resolved by assuming same value for the unobserved random effects, $b_i = b_{i'}$; however, this contrast is not directly observable.

Estimation

The joint probability density function is given by:

$$f(Y_i|X_i, b_i)f(b_i)$$

Estimation using maximum likelihood (ML) involves two steps:

First, 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 the distribution of the unobserved random effects, b_i.

However, simple analytic solutions are rarely available.

In general, computations are difficult.

- maximization of the likelihood is iterative
- likelihood evaluation requires many integrations

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

Numerical integration techniques, known as Gaussian quadrature, simply approximate the integral as a weighted sum,

$$L(\beta, \phi, G) \approx \prod_{i=1}^{N} \sum_{k=1}^{K} f(Y_i | b_i = v_k) w_k,$$

where the known quadrature points (the weights, w_k, and the evaluation points, v_k) are chosen to provide an accurate numerical approximation.

The number of quadrature points determines the degree of accuracy of the approximation involved in replacing the integral by a weighted sum.

In the second step, given ML estimates of $\boldsymbol{\beta}$, ϕ and G, the random effects can be predicted as follows,

$$\hat{b}_i = E(b_i|Y_i; \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{G})$$

(Posterior mean)

Note that $E(b_i|Y_i; \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{G})$ involves integrating (or averaging) over the distribution of the unobserved random effects, b_i.

However, simple analytic solutions are rarely available and numerical or Monte Carlo integration techniques are also required.

Case Study 1

Oral Treatment of Toenail Infection

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

$$\text{logit}\,(E[Y_{ij}|b_i]) = \beta_1 + \beta_2\text{Month}_{ij} + \beta_3\text{Trt}_i * \text{Month}_{ij} + b_i$$

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

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

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

Table 56: 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	5.28

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 infection 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 infection 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 $e^{12*0.142}$).

5. Estimated variance of the random intercepts, $\widehat{\sigma}_b^2 = 16.03$ is relatively large.

For example, the estimated variance implies that 95% of patients have a baseline risk of infection between

$$\frac{exp(-1.697 \pm 1.96 \times \sqrt{16.034})}{1 + exp(-1.697 \pm 1.96 \times \sqrt{16.034})}$$

(or between 0 and 0.997).

This suggests substantial heterogeneity of risk.

Case Study 2

Clinical trial of anti-epileptic drug progabide

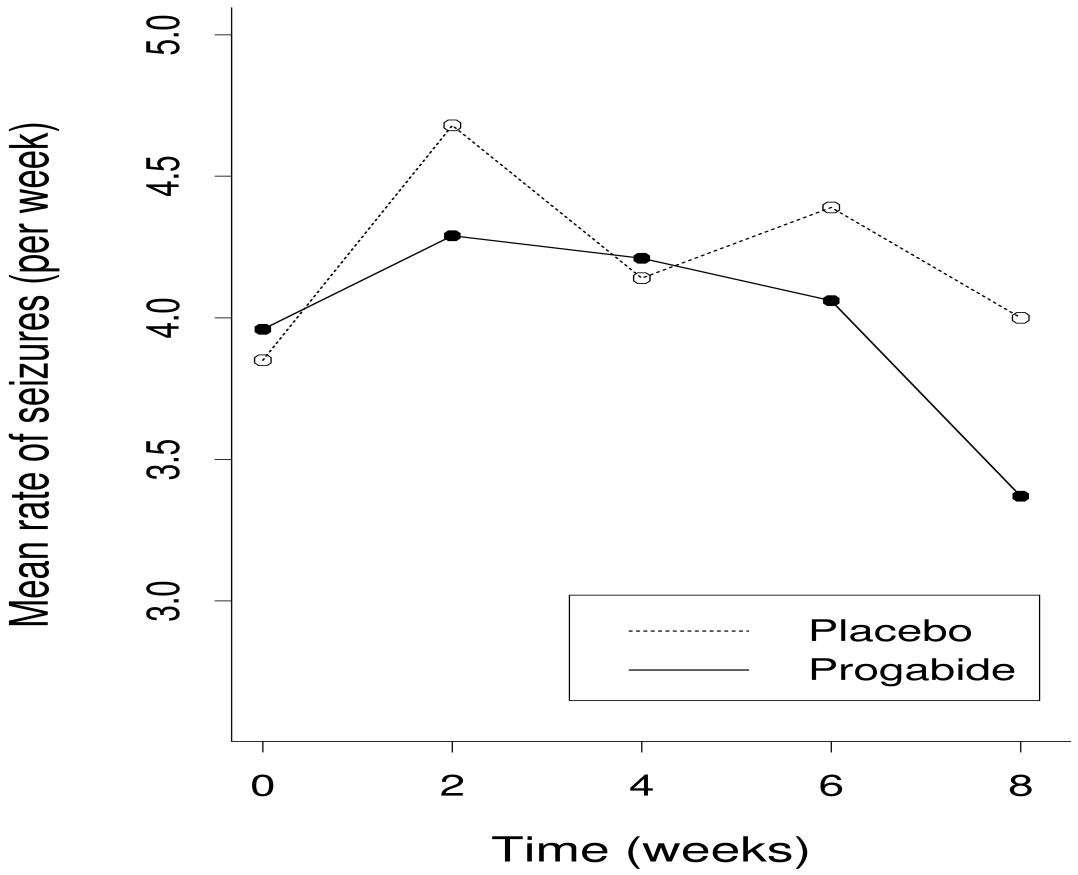
Randomized, placebo-controlled study of treatment of epileptic seizures with progabide.

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

Response variable: Count of number of seizures

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

Interested in the effect of treatment with progabide on changes in an individual's rate of seizures?


Mean rate of seizures (per week)
5.0
4.5
4.0
3.5
3.0
Placebo
Progabide
0
2
4
6
8
Time (weeks)

Assume conditional rate of seizures follows the mixed effects loglinear model,

$$\log(E[Y_{ij}|b_i]) = \log(t_{ij}) + \beta_1 + b_{1i} + \beta_2 \text{time}_{ij} + b_{2i} \text{time}_{ij} + \\ \beta_3 \text{trt}_i + \beta_4 \text{trt}_i * \text{time}_{ij}$$

where t_{ij} = length of period; $\text{time}_{ij} = 1$ if periods 1-4, 0 if baseline; $\text{trt}_i =$ 1 if progabide, 0 if placebo.

(b_{1i}, b_{2i}) are assumed to have a bivariate normal distribution with zero mean and covariance G.

Also, we assume that

$$\text{Var}(Y_{ij}|b_i) = E[Y_{ij}|b_i].$$

Table 57: Subject-specific log expected seizure rates in the two groups at baseline and during post-baseline follow-up.

Treatment Group	Period	$\log\left(\frac{E(Y_{ij} \mid b_i)}{\mathtt{T_{ij}}}\right)$
Placebo	Baseline	$\beta_1 + b_{1i}$
	Follow-up	$(\beta_1 + b_{1i}) + (\beta_2 + b_{2i})$
Progabide	Baseline	$(\beta_1 + b_{1i}) + \beta_3$
	Follow-up	$(\beta_1 + b_{1i}) + (\beta_2 + b_{2i}) + \beta_3 + \beta_4$

Parameter estimates and standard errors from mixed effects log-linear regression model for the seizure data.

Parameter	Estimate	SE	Z
Intercept	1.0707	0.1406	7.62
$\texttt{time}_{\texttt{ij}}$	−0.0004	0.1097	−0.00
$\texttt{trt}_{\texttt{i}}$	0.0513	0.1931	0.27
$\texttt{trt}_{\texttt{i}} \times \texttt{time}_{\texttt{ij}}$	−0.3065	0.1513	−2.03
$\text{Var}(b_{1i})$	0.5010	0.1010	4.96
$\text{Var}(b_{2i})$	0.2334	0.0608	3.84
$\text{Cov}(b_{1i}, b_{2i})$	0.0541	0.0559	0.97

ML based on 50-point adaptive Gaussian quadrature.

Results of the analysis suggest:

1. A patient treated with placebo has the same expected seizure rate before and after randomization [$exp(-0.0004) \approx 1$].

2. A patient treated with progabide has expected seizure rate reduced after treatment by approximately 26% [$1 - exp(-0.0004 - 0.3065) \approx 0.26$].

3. Estimated variance of the random intercepts and slopes is relatively large

4. Heterogeneity should not be ignored

Summary of Key Points

GLMMs extend the conceptual approach represented by the linear mixed effects model.

GLMMs assume natural heterogeneity across individuals in a subset of the regression coefficients.

The focus of GLMMs is on inferences about individuals.

The regression parameters, β, have ‘subject-specific’ interpretations in terms of changes in the transformed mean response for a specific individual.

GLMM using PROC NLMIXED in SAS

PROC NLMIXED in SAS is a very general and versatile procedure for fitting non-linear mixed effects models.

Here we focus on the use of PROC NLMIXED to fit GLMMs to longitudinal data.

PROC NLMIXED, as with almost all software for longitudinal analyses, requires each repeated measurement in a longitudinal data set to be a separate "record".

If the data set is in a *multivariate* mode (or "wide format"), it must be transformed to a *univariate* mode (or "long format") prior to analysis.

PROC NLMIXED directly maximizes an approximate integrated likelihood via numerical quadrature.

Caution: Likelihood approximation may not be accurate if too few quadrature points are used.

PROC NLMIXED has an option for the number of quadrature points used during evaluation of integrals, e.g. QPOINTS=50 specifies that 50 quadrature points be used for each random effect.

Table 58: Illustrative commands for a mixed effects logistic regression, with randomly varying intercepts, using PROC NLMIXED in SAS.

```
PROC NLMIXED QPOINTS=50;

   PARMS beta1=-3.0 beta2=-0.2 beta3=0.5 beta4=0.1 g11=0 to 5 by 0.5;

   eta = beta1 + beta2*time + beta3*group + beta4*group*time + b1;

   mu = exp(eta)/(1+exp(eta));

   MODEL y ~ BINARY(mu);

   RANDOM b1 ~ NORMAL(0, g11)  SUBJECT=id;

   PREDICT mu  OUT=predmean;
```

Table 59: Illustrative commands for mixed log-linear regression, with randomly varying intercepts & slopes, using PROC NLMIXED in SAS.

```
PROC NLMIXED  QPOINTS=50;

   PARMS  beta1=1.0 beta2=0.0 beta3=0.0 beta4=-0.5 g11=0 to 2 by 0.5
        g22=0 to 2 by 0.5 g12=-1 to 1 by 0.25;

   eta = beta1 + beta2*time + beta3*group + beta4*group*time + b1 + b2*time;

   mu = exp(eta);

   MODEL  y ~ POISSON(mu);

   RANDOM  b1 b2 ~ NORMAL([0,0], [g11, g12, g22])   SUBJECT=id;

   PREDICT  beta2+b2   OUT=slopes;
```

PARMS statement: lists names of all parameters (fixed effects and the covariance parameters for the random effects).

PARMS statement is also used to specify initial values (or a grid of initial values) for the parameters.

Caution: Parameters not listed on PARMS statement are assigned initial value of 1; this can be a poor choice and may lead to convergence problems.

Program statements: used to define linear predictor (the fixed and random effects) and to relate mean response to the linear predictor.

PROC NLMIXED allows multiple program statements.

MODEL statement: specifies response variable and conditional distribution of response given the random effects.

PROC NLMIXED includes options for the following exponential family distributions:

NORMAL($\texttt{m}, \texttt{v}$): specifies a normal distribution with mean `m` and variance `v`.
BINARY($\texttt{p}$): specifies a Bernoulli distribution with probability of success `p`.
BINOMIAL($\texttt{n}, \texttt{p}$): specifies a binomial distribution with `n` trials and probability of success `p`.
POISSON($\texttt{m}$): specifies a Poisson distribution with mean `m`.

RANDOM effects $\sim$ distribution SUBJECT=variable;
Random statement defines the random effects and a variable that determines the clustering of observations within an individual via SUBJECT option.

Note: Data should be sorted by the SUBJECT variable since PROC NLMIXED assumes a new realization of the random effects occurs whenever the SUBJECT=variable changes.

All random effects are assumed to have a normal distribution, normal($\mathtt{m}, \mathtt{v}$), with mean (vector) `m` and variance (covariance matrix) `v`.

For a single random effect the syntax is:

$$\text{RANDOM b1} \sim \text{NORMAL}(0, \text{g11})\ \ \text{SUBJECT=id;}$$

For two random effects the corresponding syntax requires the use of brackets for the mean vector and covariance matrix:

$$\text{RANDOM b1 b2} \sim \text{NORMAL(} [0, 0],\ [\text{g11}, \text{g12}, \text{g22}]\text{)}\ \ \text{SUBJECT=id;}$$

Only the non-redundant, lower triangle of the covariance matrix is included in the parameters of the multivariate normal distribution for the random effects.

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 20

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Contrasting Marginal and Mixed Effects Models for Longitudinal Data

So far, we have discussed two main extensions of generalized linear models:

1. Marginal Models
2. Generalized Linear Mixed Models

There are important distinctions between these two broad classes of models that go beyond simple differences in approaches for accounting for the within-subject association.

These two classes of models have somewhat different targets of inference and address subtly different questions regarding longitudinal change in the response.

A **marginal model** for the mean response is given by

$$g(\mu_{ij}) = g[E(Y_{ij}|X_{ij})] = X'_{ij}\beta = \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp},$$

where $g(\cdot)$ is an appropriate non-linear link function (e.g., logit or log).

In marginal models, β's have interpretation in terms of changes in the transformed mean response in the study population, and their relation to covariates.

The population means can be expressed in terms of the inverse link function, say $h(\cdot) = g^{-1}(\cdot)$,

$$h[g(\mu_{ij})] = \mu_{ij} = E(Y_{ij}|X_{ij}) = h(\beta_1 X_{ij1} + \cdots + \beta_p X_{ijp}).$$

Next, consider the **generalized linear mixed model**

$$g[E(Y_{ij}|X_{ij}, b_i)] = X_{ij}'\beta^* + Z_{ij}'b_i,$$

where the random effects b_i have a distribution with mean zero and covariance matrix G.

The regression coefficients β^* have subject-specific interpretations in terms of changes in the transformed mean response for a specific individual.

β^* do not describe changes in the transformed mean response in the study population.

In GLMMs there is an implied model for the marginal means.

This can be obtained by averaging over distribution of the random effects,

$$\begin{aligned}
\mu_{ij} &= E(Y_{ij}|X_{ij}) \\
&= E[E(Y_{ij}|X_{ij}, b_i)] \\
&= E[h(X'_{ij}\beta^* + Z'_{ij}b_i)] \\
&= \int_{-\infty}^{\infty} h(X'_{ij}\beta^* + Z'_{ij}b_i) f(b_i) db_i.
\end{aligned}$$

However, this expression for $E(Y_{ij}|X_{ij})$ does not, in general, have a closed-form expression and, moreover,

$$E(Y_{ij}|X_{ij}) \neq h(X'_{ij}\beta)$$

for any β, e.g., logistic mixed effects model $\neq$ marginal logistic model.

That is, marginalized model doesn't satisfy generalized linear model.

Simple Numerical Illustration

Consider hypothetical data on true propensity for disease, $\Pr(Y_{ij} = 1|b_i)$, for three individuals measured at baseline (pre) and following treatment with a new drug intended to reduce the risk of disease (post).

The three individuals are discernibly different in terms of their underlying propensity for disease at baseline.

This heterogeneity can be expressed in terms of random effects, b_i.

Individuals A, B, and C have "high", "medium" and "low" underlying risk for disease.

Assume target population is comprised of an equal number of individuals that fall into these three distinct risk groups.

Hypothetical data on the true propensity for disease, at baseline and post-baseline, for three individuals with heterogeneous propensities for disease.

Individual	Pre	Post	Difference	Log(OR)
A	0.80	0.67	-0.13	-0.68
B	0.50	0.33	-0.17	-0.71
C	0.20	0.11	-0.09	-0.70
Pop. Average	0.50	0.37	-0.13	

Final row of table contains the population averages (obtained as equally-weighted means).

For a linear function of the propensity for disease (i.e., the difference), the "difference of the averages" is equal to the "average of the differences".

Taking the average of the subject-specific effects (as a single number summary of the subject-specific effects),

$$\frac{-0.13 - 0.17 - 0.09}{3} = -0.13.$$

Alternatively, can compare the average propensity for disease at baseline (0.5) and post-baseline (0.37).

$$(0.37 - 0.50) = -0.13.$$

The latter can be thought of as a contrast of population averages.

With a non-linear function of the propensity for disease, a “non-linear contrast of the averages” is not equal to the “average of the non-linear contrasts”.

Consider the log odds ratios:

Taking the average of the subject-specific effects (as a single number summary of the subject-specific effects),

$$\frac{-0.68 - 0.71 - 0.70}{3} = -0.697.$$

Alternatively, compare the log odds of disease in the population at baseline, $\log(0.5/0.5) = 0$ and post-baseline, $\log(0.37/0.63) = -0.532$.

This comparison yields a measure of effect, -0.532, which is approximately 25% smaller than the summary of the subject-specific effect, -0.697.

Individual	Pre	Post	Difference	Log(OR)
A	0.80	0.67	-0.13	-0.68
B	0.50	0.33	-0.17	-0.71
C	0.20	0.11	-0.09	-0.70
Pop. Average	0.50	0.37	-0.13	

In marginal models, the regression parameters describe the margins of the table.

In GLMMs, the fixed effects describe the interior of the table.

Next we consider a graphical illustration that highlights the differences between these two approaches.

Graphical Illustration

Suppose Y_i is a vector of binary responses and it is of interest to describe changes in the log odds of success over time.

A logistic regression model, with randomly varying intercepts, is given by

$$\text{logit}[E(Y_{ij}|b_i)] = \beta_1^* + \beta_2^* t_{ij} + b_i$$

where $t_{ij} = 0$ at baseline and $t_{ij} = 1$ post-baseline.

The b_i are assumed to have a normal distribution with zero mean and variance $\sigma_b^2 = \text{Var}(b_i)$.

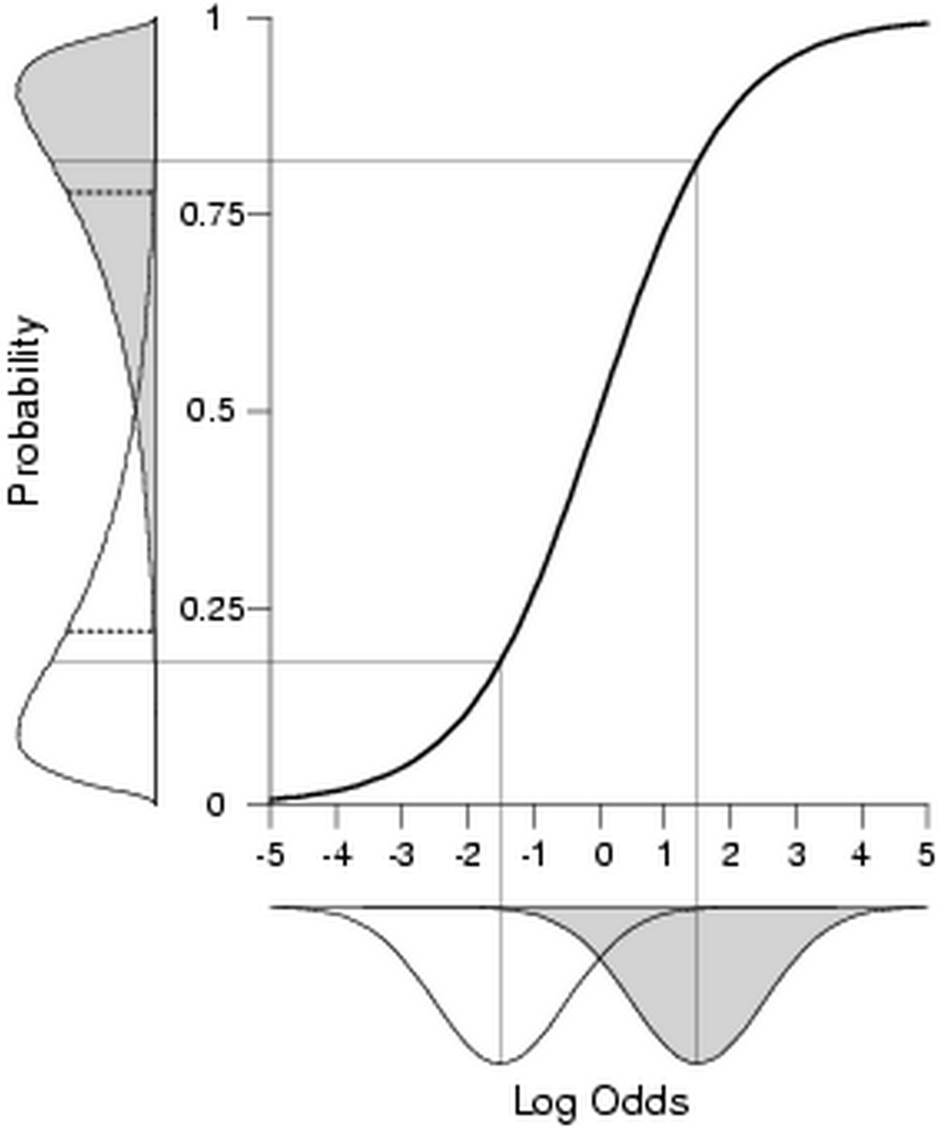
Let $\beta_1^* = 1.5$, $\beta_2^* = -3.0$, and $\text{Var}(b_i) = 1.0$.

At baseline, log odds has a normal distribution with mean = median = 1.5 (see shaded densities).

Note, however, that subject-specific probabilities of disease have a negatively skewed distribution with median, but not mean, of 0.82.

The mean of the subject-specific probabilities is 0.78.

Thus, probability of disease for a “typical” individual from the population (0.82) is not the same as the prevalence of disease in the same population (0.78).

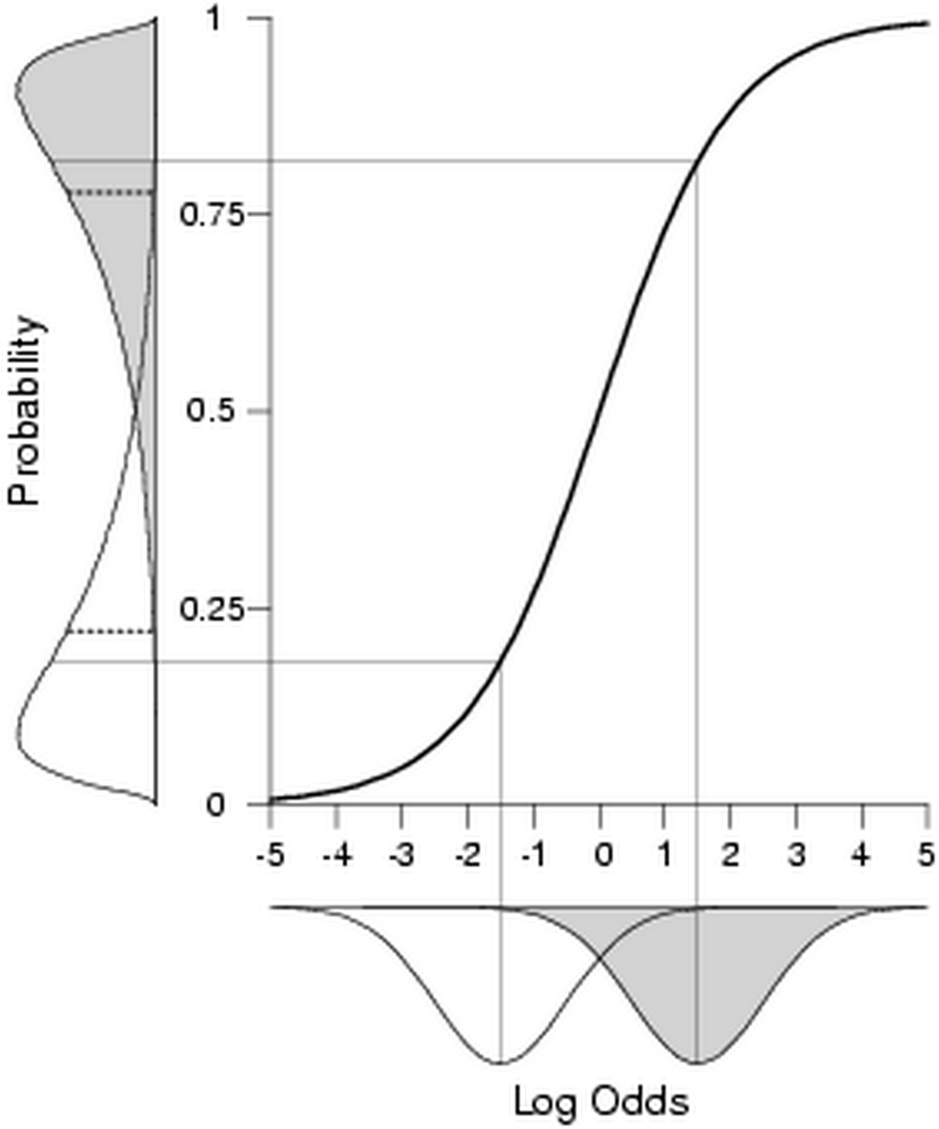

1
0.75
0.5
0.25
0
Probability
-5
-4
-3
-2
-1
0
1
2
3
4
5
Log Odds

Similarly, the log odds of disease post-baseline has a normal distribution with mean = median = -1.5 (see unshaded densities).

However, subject-specific post-baseline probabilities of disease have a positively skewed distribution with median, but not mean, of 0.18.

The mean of the subject-specific probabilities is 0.22.

Thus, probability of disease post-baseline for a "typical" individual from the population (0.18) is not the same as the prevalence of disease in the same population (0.22).


1
0.75
0.5
0.25
0
Probability
-5
-4
-3
-2
-1
0
1
2
3
4
5
Log Odds

The effect of treatment on the log odds of disease for a typical individual from the population, $\beta_2^* = -3.0$, is not the same as the contrast of population log odds.

The latter is what is estimated in a marginal model, say

$$\text{logit}\{E(Y_{ij})\} = \beta_1 + \beta_2\, t_{ij},$$

and can be obtained by comparing the log odds of disease in the population at baseline, $\log(0.78/0.22) = 1.255$, with the log odds of disease in the population post-baseline, $\log(0.22/0.78) = -1.255$.

This yields a population-averaged measure of effect, $\beta_2 = -2.51$, which is approximately 15% smaller than β_2^*, the subject-specific effect of treatment.

Case Study

Cross-Over Trial of Cerebrovascular Deficiency

- Two-period cross-over trial comparing effects of active drug to placebo on cerebrovascular deficiency

- 67 patients randomly allocated to two treatment sequences

- 34 patients receiving Placebo → Active

- 33 patients receiving Active → Placebo

- Each patient has a bivariate binary response vector, $Y_i = (Y_{i1}, Y_{i2})$ denoting whether an electrocardiogram was normal (0) or abnormal (1).

Data from a two-period cross-over trial comparing the effects of active drug to placebo on cerebrovascular deficiency. The response indicates whether an electrocardiogram was normal (0) or abnormal (1).

	Response (Period 1, Period 2)			
Sequence	(1,1)	(1,0)	(0,1)	(0,0)
Sequence 1 (P → A)	6	0	6	22
Sequence 2 (A → P)	9	4	2	18

P: Placebo; A: Active drug.

First, consider marginal logistic model

$$\text{logit}(\mu_{ij}) = \text{logit}[\Pr(Y_{ij} = 1)] = \beta_1 + \beta_2\text{Treatment} + \beta_3\text{Period}$$

where Treatment (0 = Placebo, 1 = Active drug) and Period (0 = Period 1, 1 = Period 2).

The within subject association between the two responses was modelled in terms of a common log odds ratio, α,

$$\log \frac{\Pr(Y_{i1} = 1, Y_{i2} = 1)\Pr(Y_{i1} = 0, Y_{i2} = 0)}{\Pr(Y_{i1} = 1, Y_{i2} = 0)\Pr(Y_{i1} = 0, Y_{i2} = 1)} = \alpha.$$

Parameter estimates and standard errors from marginal logistic regression model for the cerebrovascular deficiency data.

Parameter	Estimate	SE	Z
Intercept	-1.2433	0.2999	-4.15
Treatment	0.5689	0.2335	2.44
Period	0.2951	0.2319	1.27
log OR (α)	3.5617	0.8148	4.37

The results indicate that treatment with the active drug is harmful, increasing the rates of abnormal electrocardiograms.

The odds of an abnormal electrocardiogram is 1.77 (or $e^{0.57}$) times higher when treated with active drug versus placebo.

The estimate of the within-subject association is $\widehat{\alpha} = 3.56$, indicating that there is very strong positive association (OR$= 35.2$).

Next, consider logistic regression model with a random patient effect,

$$\text{logit}[E(Y_{ij}|b_i)] = \beta_1^* + \beta_2^*\text{Treatment} + \beta_3^*\text{Period} + b_i$$

where the random effect b_i is assumed to have a normal distribution with zero mean and variance, $\sigma_b^2 = \text{Var}(b_i)$.

Parameter estimates and standard errors from mixed effects logistic regression model for the cerebrovascular deficiency data.

Parameter	Estimate	SE	Z
Intercept	-4.0817	1.6711	-2.44
Treatment	1.8631	0.9269	2.01
Period	1.0376	0.8189	1.27
$\sigma_b^2 = \text{Var}(b_i)$	24.4365	18.8500	1.30

ML based on 100-point adaptive Gaussian quadrature.

The results also indicate that treatment with the active drug is harmful, increasing the patient-specific risk of an abnormal electrocardiogram.

In particular, a patient's odds of an abnormal electrocardiogram is 6.4 (or $e^{1.86}$) times higher when treated with active drug than when treated with the placebo.

The estimate of the variance of b_i, $\widehat{\sigma}_b^2 = 24.4$, indicates that there is very substantial between-patient variability in their propensity for abnormal electrocardiograms.

Comparison of the two estimated effects of treatment, $e^{\widehat{\beta}_2} = 1.8$ and $e^{\widehat{\beta}_2^*} = 6.4$, from the marginal and mixed effects logistic regression models highlights the distinction between these two analytic approaches.

$\widehat{\beta}_2$ from marginal model describes how the average rates (expressed in terms of odds) of abnormal ECGs could be increased in the study population if patients are treated with the active drug.

$\widehat{\beta}_2^*$ from the mixed effects model describes how the odds of an abnormal ECG increases for any patient treated with the active drug.

Thus, a population-level analysis understates the individual risk, and vice versa.

In summary, the answer to the question “what are the side effects of the active drug” will depend on whether scientific interest is in its impact on the study population or on an individual drawn at random from that population.

With marginal models the main focus is on inferences about the study population.

With generalized linear mixed models the main focus is on inferences about individuals.

Aside

Does the very large estimate of variance, $\widehat{\sigma}_b^2 = 24.4$, accurately reflect between-patient variability in the risk of abnormal electrocardiogram?

In this example, a large proportion of subjects (82%) had same response, (0,0) or (1,1), at both occasions.

This feature can only be captured by a normal distribution for the log odds with large variance.

When number of repeated binary responses is small, and there is a large proportion of subjects with positive (negative) responses at all occasions, the normal assumption for b_i is questionable.

Concluding Remarks

Unlike linear models, where the concepts of regression analysis can be applied quite robustly, longitudinal analysis of categorical data raises many subtle issues.

Different models for categorical outcomes can give discernibly different results.

The choice and meaning of longitudinal models for categorical outcomes require somewhat greater care.

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”

BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 21

INSTRUCTOR: GARRETT FITZMAURICE

Laboratory for Psychiatric Biostatistics

McLean Hospital

Department of Biostatistics

Harvard School of Public Health

Multilevel Models

Until today, this course has focused on the analysis of longitudinal data.

Mixed models can also be used to analyze multilevel data.

Hierarchical or multilevel data arise when there is a *clustered/grouped* structure to the data.

Data of this kind frequently arise in the social, behavioral, and health sciences since individuals can be grouped in so many different ways.

For example, in studies of health services and outcomes, assessments of quality of care are often obtained from patients who are *nested* within different clinics.

Such data can be regarded as hierarchical/multilevel, with patients referred to as the level 1 units and clinics the level 2 units.

In this example there are two levels in the data hierarchy and, by convention, the lowest level of the hierarchy is referred to as level 1.

The term “level”, as used in this context, signifies the position of a unit of observation within a hierarchy.

Clustering in multilevel data can be due to a naturally occurring hierarchy in the target population or a consequence of study design (or sometimes both).

Naturally Occurring Data Hierarchies

Studies of nuclear families: observations on the mother, father, and children (level 1 units) nested within families (level 2 units).

Studies of health services/outcomes: observations on patients (level 1 units) nested within clinics (level 2 units).

Studies of education: observations on children (level 1 units) nested within classrooms (level 2 units).

Note: Naturally occurring hierarchical data structures can have more than two levels, e.g., children (level 1 units) nested within classrooms (level 2 units), nested within schools (level 3 units).

Clustering as Consequence of Study Design

Longitudinal Studies: the clusters are composed of the repeated measurements obtained from a single individual at different occasions.

In longitudinal studies the level 1 units are the repeated occasions of measurement and the level 2 units are the subjects.

Cluster-Randomized Clinical Trials: Groups (level 2 units) of individuals (level 1 units), rather than the individuals themselves, are randomly assigned to different treatments or interventions.

Complex Sample Surveys: Many national surveys use multi-stage sampling, e.g., NHANES.

For example, in 1st stage, “primary sampling units” (PSUs) are defined based on counties in the United States. A first-stage random sample of PSUs are selected. In 2nd stage, within each selected PSU, a random sample of census blocks are selected. In 3rd stage, within selected census blocks, a random sample of households are selected.

Resulting data can be regarded as hierarchical, with households being the level 1 units, area segments the level 2 units, and counties the level 3 units.

Finally, clustering can be due to both study design and naturally occurring hierarchies in the target population.

Example: Clinical trials are often conducted in many different centers to ensure sufficient numbers of patients and/or to assess the effectiveness of the treatment in different settings.

Observations from a multi-center longitudinal clinical trial can be regarded as hierarchical data with 3 levels, with repeated measurement occasions (level 1 units) nested within subjects (level 2 units) nested within clinics (level 3 units).

Distinctive Feature of Multilevel Data

Distinctive feature of multilevel data is that they are *clustered*.

A consequence of this clustering is that measurement on units within a cluster are more similar than measurements on units in different clusters.

For example, two children selected at random from the same family are expected to respond more similarly than two children randomly selected from different families.

The clustering can be expressed in terms of correlation among the measurements on units within the same cluster.

Statistical models for hierarchical data must account for the intra-cluster correlation at each level; failure to do so can result in misleading inferences.

Multilevel Linear Models

The dominant approach to analysis of multilevel data employs a type of linear mixed effects model known as the hierarchical linear model.

The correlation induced by clustering is described by random effects at each level of the hierarchy.

Note: In a multilevel model, the response is obtained at the first level, but covariates can be measured at any level.

For example, if we are studying BMI, we can measure individual diets, family attitudes about food and purchasing habits, and community attributes such as the density of fast-food restaurants.

Combining covariates measured at different levels of the hierarchy within a single regression model is central to hierarchical modelling.

We begin by introducing the ideas with the two-level model.

Later we move to the three-level model to illustrate the general approach.

Two-Level Linear Models

Notation:

Let i index level 1 units and j index level 2 units (by convention, the subscripts are ordered from the lowest to the highest level).

We assume n_2 level 2 units in the sample.

Each of these clusters ($j = 1, 2, \cdots, n_2$) is composed of n_{1j} level 1 units.

For example, in a two-level study of physician practices, we would study n_2 practices, with n_{1j} patients in the j^{th} practice.

Let Y_{ij} denote the response for patient i in the j^{th} practice.

Associated with each Y_{ij} is a $1 \times p$ (row) vector of covariates, X_{ij}

Consider the following model for the mean:

$$E(Y_{ij}) = X_{ij}\beta$$

For example, in a multi-center clinical trial comparing two treatments, we might assume that:

$$E(Y_{ij}) = \beta_1 + \beta_2 \mathrm{Trt}_{ij}$$

where Trt_{ij} is an indicator variable for treatment group (or Trt_j if treatment is constant within practice).

The two-level hierarchical linear model assumes that the correlation within practices can be described by a random effect.

Thus, we assume that

$$Y_{ij} = X_{ij}\beta + b_j + \epsilon_{ij}$$

Or, more generally,

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_j + \epsilon_{ij}$$

with more than 1 random effect.

Features of the Two-Level Linear Model

1. Model defines two sources of variation. Magnitudes of within- and between-cluster variation determine degree of clustering/correlation.
2. For a given level 2 unit, random effects are assumed constant across level 1 units.
3. Conditional expectation of Y_{ij}, given identity of the level 2 group, is

$$X_{ij}\beta + Z_{ij}b_j$$

4. Level 1 observations are assumed to be conditionally independent given the random effects.

The two-level model is identical to the linear mixed model with intraclass correlation structure for repeated measurements (albeit with reversal of subscripting!).

Three-Level Linear Models

Next, consider a three-level *longitudinal* clinical trial in which

(1) physician practices are randomized to treatment,

(2) patients are nested within practices, and

(3) patients are measured at baseline and at three occasions after treatment.

Level 1 is occasions, level 2 is patients, and level 3 is practice.

Let Y_{ijk} denote response at the i^{th} observation of the j^{th} patient in the k^{th} practice.

Covariates can be measured at any of three levels. However, we now introduce random effects to represent clustering at both levels 2 and 3.

The general three-level linear model is written as follows:

$$Y_{ijk} = X_{ijk}\beta + Z_{ijk}^{(3)} b_k^{(3)} + Z_{ijk}^{(2)} b_{jk}^{(2)} + \epsilon_{ijk}$$

Example: Three-Level Model for the Multi-Level Longitudinal Clinical Trial

Let t_{ijk} denote the time from baseline at which Y_{ijk} is obtained.

Also, let Trt_{ij} denote the treatment given to the j^{th} patient at the i^{th} occasion.
The treatment may be constant over occasions for a given patient (Trt_j).

A hierarchical three-level model for the response is given by

$$Y_{ijk} = \beta_1 + \beta_2 t_{ijk} + \beta_3(\text{Trt}_j \times t_{ijk}) + b_k^{(3)} + b_{jk}^{(2)} + \epsilon_{ijk}$$

This model assumes a common intercept and separate linear trends over time in the two treatment groups.

If

$$\mathrm{Var}(b_k^{(3)}) = G^{(3)}, \mathrm{Var}(b_{jk}^{(2)}) = G^{(2)}, \text{ and } \mathrm{Var}(\epsilon_{ijk}) = \sigma^2,$$

and all random effects are assumed to be independent, then

$$\mathrm{Var}(Y_{ijk}) = G^{(2)} + G^{(3)} + \sigma^2$$

and the covariance between two observations from the same patient is

$$G^{(2)} + G^{(3)}$$

Thus, the observations for a given patient have an intraclass correlation structure, with

$$\mathrm{Corr}(Y_{ijk}, Y_{ijl}) = \frac{G^{(2)} + G^{(3)}}{G^{(2)} + G^{(3)} + \sigma^2}.$$

Because this is a linear mixed model,

$$E(Y_{ijk}) = \beta_1 + \beta_2 t_{ijk} + \beta_3(\mathrm{Trt}_{ij} \times t_{ijk})$$

Estimation

For the three-level linear model, the standard distributional assumptions are that:

$$b_k^{(3)} \sim N(0, G^{(3)}), b_{jk}^{(2)} \sim N(0, G^{(2)}), \text{ and } \epsilon_{ijk} \sim N(0, \sigma^2)$$

Given these assumptions, estimation of the model parameters is relatively straightforward. The GLS estimate of β is given by

$$\beta = \Big\{ \sum_{k=1}^{n_3} (X_k' V_k^{-1} X_k) \Big\}^{-1} \sum_{k=1}^{n_3} (X_k' V_k^{-1} Y_k)$$

where Y_k is a column vector of length $\sum_{j=1}^{n_{2k}} n_{1jk}$, the number of observations in the k^{th} cluster. X_k is the corresponding matrix of covariates, and V_k is the covariance matrix of Y_k.

Estimation (Continued)

As before, we use REML (or ML) to obtain estimates of $G^{(3)}$, $G^{(2)}$, and σ^2.

Once these estimates are obtained, we can estimate the covariance matrices, V_k, and substitute those estimates into the expression for the GLS estimator.

This estimation procedure is available in PROC MIXED in SAS.

It is also available in MLwiN and HLM, two stand-alone programs developed for multilevel modeling.

Case Study 1: Developmental Toxicity Study of Ethylene Glycol

Developmental toxicity studies of laboratory animals play a crucial role in the testing and regulation of chemicals.

Exposure to developmental toxicants typically causes a variety of adverse effects, such as fetal malformations and reduced fetal weight at term.

In a typical developmental toxicity experiment, laboratory animals are assigned to increasing doses of a chemical or test substance.

Consider an analysis of data from a development toxicity study of ethylene glycol (EG).

Ethylene glycol is used as an antifreeze, as a solvent in the paint and plastics industries, and in the formulation of various types of inks.

In a study of laboratory mice conducted through the National Toxicology Program (NTP), EG was administered at doses of 0, 750, 1500, or 3000 mg/kg/day to 94 pregnant mice (dams) beginning just after implantation.

Following sacrifice, fetal weight and evidence of malformations were recorded for each live fetus.

In our analysis, we focus on the effects of dose on fetal weight.

Summary statistics (ignoring clustering in the data) for fetal weight for the 94 litters (composed of a total of 1028 live fetuses) are presented in Table 60.

Fetal weight decreases monotonically with increasing dose, with the average weight ranging from 0.97 (gm) in the control group to 0.70 (gm) in the group administered the highest dose.

The decrease in fetal weight is not linear in increasing dose, but is approximately linear in increasing $\sqrt{\text{dose}}$.

Table 60: Descriptive statistics on fetal weight.

Dose				Weight (gm)	
(mg/kg)	$\sqrt{\text{Dose}/750}$	Dams	Fetuses	Mean	St. Deviation[†]
0	0	25	297	0.972	0.098
750	1	24	276	0.877	0.104
1500	1.4	22	229	0.764	0.107
3000	2	23	226	0.704	0.124

†Calculated ignoring clustering.

Because the observations are clustered within dam, the analysis must take account of clustering.

If it does not, the apparent sample size for comparisons between doses will be exaggerated.

To fit a two-level model that is linear in sqrt(dose),

$$Y_{ij} = \beta_1 + \beta_2 \sqrt{\text{dose}/750} + b_j + \epsilon_{ij},$$

we can use the following commands:

```
DATA toxicity;
  INFILE ’c:\bio226\datasets\ethyleneglycol.txt’;
  INPUT id dose weight mal;
  newdose=sqrt(dose/750);
  RUN;

PROC MIXED DATA=toxicity;
  CLASS id;
  MODEL weight = newdose / SOLUTION CHISQ;
  RANDOM INTERCEPT / SUBJECT=id G;
RUN;
```

Results

Variable	Estimate	SE	Z
Fixed Effects			
Intercept	0.98	0.02	61.3
Newdose	-0.13	0.01	-10.9
Random Effects			
Level 2 Variance ($\sigma_2^2 \times 100$)	0.73	0.12	6.1
Level 1 Variance ($\sigma_1^2 \times 100$)	0.56	0.03	21.6

The estimate of σ_2^2 indicates significant clustering of weights within litter. The estimated within-litter correlation is

$$\begin{aligned}
\widehat{\rho} &= \widehat{\sigma}_2^2/(\widehat{\sigma}_2^2 + \widehat{\sigma}_1^2) \\
&= 0.73/(0.73 + 0.56) \\
&= 0.57
\end{aligned}$$

The estimated decrease in weight, comparing the highest dose to 0 dose, is 0.27 (0.22, 0.33).

The model-based and empirical (sandwich) standard errors are very similar (not shown), indicating that the random effects structure is adequate.

It is also easy to test for linearity on the square root scale, though we have data at only four doses.

Case Study 2: The Television, School, and Family Smoking Prevention and Cessation Program

A randomized study with a 2 by 2 factorial design:

Factor 1: A school-based social-resistance curriculum (CC)
Factor 2: A television-based prevention program (TV)

We report results for 1,600 seventh graders from 135 classes in 28 schools in Los Angeles

The response variable, the tobacco and health knowledge scale (THKS), was administered before and after the intervention.

We consider a linear model for post-intervention THKS, with baseline THKS as a covariate.

Descriptive Statistics

			Pre-THKS		Post-THKS	
CC	TV	n	Mean	Std Dev	Mean	Std Dev
No	No	421	2.15	1.18	2.36	1.30
No	Yes	416	2.09	1.29	2.54	1.44
Yes	No	380	2.05	1.29	2.97	1.40
Yes	Yes	383	1.98	1.29	2.82	1.31

The mean value of Pre-THKS does not differ significantly among treatment groups.

Three-Level Model

Model the adjusted change in THKS scores as function of main effects of CC and TV and the CC × TV interaction:

$$Y_{ijk} = \beta_1 + \beta_2\text{Pre-THKS} + \beta_3\text{CC} + \beta_4\text{TV} + \beta_5\text{CC} \times \text{TV} + b_k^{(3)} + b_{jk}^{(2)} + \epsilon_{ijk}.$$

In a slightly modified notation, assume

$$\begin{array}{rcl} \epsilon_{ijk} & \sim & N(0, \sigma_1^2) \\ b_{jk}^{(2)} & \sim & N(0, \sigma_2^2) \\ b_k^{(3)} & \sim & N(0, \sigma_3^2) \end{array}$$

This is the standard hierarchical (or multilevel) linear model with school and classroom effects modelled by incorporating random effects at levels 3 and 2, respectively (level 1 units are the children).

PROC MIXED in SAS

```
DATA tvandcc;
  INFILE 'c:\bio226\datasets\tv.txt';
  INPUT sid cid cc tv baseline THKS;
RUN;

PROC MIXED DATA=tvandcc COVTEST;
  CLASS sid cid;
  MODEL thks = baseline cc tv cc*tv / S;
  RANDOM INTERCEPT / SUBJECT=sid  G ;
  RANDOM INTERCEPT / SUBJECT=cid  G ;
RUN;
```

Table 61: Fixed effects estimates for the THKS scores.

Parameter	Estimate	SE	Z
Intercept	1.702	0.1254	13.57
Pre-Intervention THKS	0.305	0.0259	11.79
CC	0.641	0.1609	3.99
TV	0.182	0.1572	1.16
CC × TV	−0.331	0.2245	−1.47

Table 62: Random effects estimates for the THKS scores.

Parameter	Estimate	SE	Z
Level 3 Variance:			
σ_3^2	0.039	0.0253	1.52
Level 2 Variance:			
σ_2^2	0.065	0.0286	2.26
Level 1 Variance:			
σ_1^2	1.602	0.0591	27.10

Consider REML estimates of the three sources of variability.

Comparing their relative magnitudes, there is variability at both classroom and school levels, with almost twice as much variability among classrooms within a school as among schools themselves.

Correlation among THKS scores for classmates (or children within same classroom within same school) is approximately 0.061 (or $\frac{0.039+0.065}{0.039+0.06+1.602}$).

Correlation among THKS scores for children from different classrooms within same school is approximately 0.023 (or $\frac{0.039}{0.039+0.06+1.602}$).

Next, consider REML estimates of fixed effects for the interventions.

When compared to their SEs, indicate that neither mass-media intervention (TV) nor its interaction with social-resistance classroom curriculum (CC) have an impact on adjusted changes in THKS scores from baseline.

There is a significant effect of the social-resistance classroom curriculum, with children assigned to the social-resistance curriculum showing increased knowledge about tobacco and health.

The estimate of the main effect of CC, in the model that excludes the CC × TV interaction, is 0.47 (SE $= 0.113$, $p < 0.0001$).

The intra-cluster correlations at both the school and classroom levels are relatively small.

It is very tempting to regard this as an indication that the clustering in these data is inconsequential.

However, such a conclusion would be erroneous.

Although intra-cluster correlations are relatively small, they have an impact on inferences concerning the effects of the intervention conditions.

To illustrate this, consider analysis that ignores clustering in the data:

$$Y_{ijk} = \beta_1 + \beta_2 \text{Pre-THKS} + \beta_3 \text{CC} + \beta_4 \text{TV} + \beta_5 \text{CC} \times \text{TV} + e_{ijk},$$

The results of fitting this model to the THKS scores are presented in Table 63 and the estimates of the fixed effects are similar to those reported in Table 61.

However, SEs (assuming no clustering) are misleadingly small for intervention effects and lead to substantively different conclusions about effects of intervention conditions.

This highlights an important lesson: the impact of clustering depends on both the magnitude of the intra-cluster correlation and the cluster size.

For the data from the TVSFP, the cluster sizes vary from 1–13 classrooms within a school and from 2–28 students within a classroom.

With relatively large cluster sizes, even very modest intra-cluster correlation can have a discernible impact on inferences.

Table 63: Fixed effects estimates from analysis that ignores clustering in the THKS scores.

Parameter	Estimate	SE	Z
Intercept	1.661	0.0844	19.69
Pre-Intervention THKS	0.325	0.0258	12.58
CC	0.641	0.0921	6.95
TV	0.199	0.0900	2.21
CC × TV	−0.322	0.1302	−2.47

Generalizations

The multilevel model can be generalized to an arbitrary number of levels.

Generalized linear mixed effects models (GLMMs) have also been developed for the analysis of binary outcomes and counts in the multilevel setting (see FLW, Chapter 17).

Cautionary Remarks

Multilevel modeling can be difficult:

- A covariate can operate at different levels
- It is not always clear how to combine covariates within a single model
- Though hierarchical linear models with random effects are appealing, the extension to generalized linear models raises difficult problems of interpretation.
- As discussed earlier, marginal models and mixed-effects models can give quite different results in the non-linear setting

Summary

Despite certain complexities, multilevel models are now widely used.

In both designed experiments and studies of effects of family/community factors on health, multilevel models provide a usually effective approach to data analysis that accounts for correlations induced by clustering.

Multilevel models are, in one sense, no different than longitudinal models.

Unlike logistic regression and survival analysis, where concept of regression analysis can be applied quite robustly and with few choices, longitudinal and multilevel analysis require more careful thought about the choice and meaning of models.

This is both their challenge and their reward.